Backward Induction Reasoning beyond Backward Induction

BSE Working Paper 1315| February 2022

Emiliano Catonini, Antonio Penta

bse.eu/research

Backward Induction Reasoning beyond Backward Induction*

Emiliano Catonini† NYU-Shanghai

Antonio Penta‡ ICREA, UPF, BSE and TSE

February 1, 2022

Abstract

Backward Induction is a fundamental concept in game theory. As an algorithm, it can only be used to analyze a very narrow class of games, but its logic is also invoked, albeit informally, in several solution concepts for games with imperfect or incomplete information (Subgame Perfect Equilibrium, Sequential Equilibrium, etc.). Yet, the very meaning of 'backward induction reasoning' is not clear in these settings, and we lack a way to apply this simple and compelling idea to more general games. We remedy this by introducing a solution concept for games with imperfect and incomplete information, Backwards Rationalizability, that captures precisely the implications of backward induction reasoning. We show that Backwards Rationalizability satisfies several properties that are normally ascribed to backward induction reasoning, such as: (i) an incomplete-information extension of subgame consistency (*continuation-game consistency*); (ii) the possibility, in finite horizon games, of being computed via a tractable *backwards procedure*; (iii) the view of unexpected moves as *mistakes*; (iv) a characterization of the robust predictions of a 'perfect equilibrium' notion that introduces the backward induction logic and nothing more into equilibrium analysis.

We also discuss a few applications, including a new version of *peer-confirming equilibrium* (Lipnowski and Sadler (2019)) that, thanks to the backward induction logic distilled by Backwards Rationalizability, restores in dynamic games the natural comparative statics the original concept only displays in static settings.

Keywords: backward induction, backwards procedure, backwards rationalizability, incomplete information, interim perfect equilibrium, rationalizability, robustness

JEL codes: C72, C73, D82.

1 Introduction

Backward induction is one of the most fundamental notions of game theory. Strictly speaking, the backward induction algorithm is only defined for games with perfect and complete informa-

*Earlier versions of some of the results in this paper circulated under the title "Backward Induction Reasoning in Incomplete Information Games", by Penta (2012*a*). The present paper is a substantially revised and extended version of that earlier work. This paper benefited from the comments of several seminar and conference audiences. Among the many valuable inputs, we are especially indebted to the generosity of Larbi Alaoui, Pierpaolo Battigalli, George Mailath, Andres Perea, and Bill Sandholm. We also thank Andrea Salvanti for the RA support. The BSE acknowledges the financial support of the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S).

†NYU-Shanghai. E-mail: emiliano.catonini@nyu.edu

‡ICREA, Universitat Pompeu Fabra, BSE and TSE. E-mail: antonio.penta@upf.edu

tion and without 'relevant ties', but its logic has a much broader scope in the discipline. For instance, subgame perfect equilibrium is commonly viewed as the natural extension of backward induction to games with imperfect information. But there is a sense in which also solution concepts for incomplete information games, such as sequential equilibrium (Kreps and Wilson (1982)) or trembling-hand perfect equilibrium (Selten (1975)), are often thought of as having a backward induction flavor. Yet, it is not even clear what "backward induction" means in games with incomplete information, which are typically not solved "backwards", nor to what extent its logic can be separated from equilibrium assumptions. More broadly: *What do we mean by "backward induction reasoning"?* Despite the central position in game theory, there is no comprehensive, formal answer to this question.

The conceptual significance of providing such an answer is obvious, but its relevance is also practical: the many solution concepts that have been developed with a 'backward induction' flavor typically conflate its logic with other kinds of ideas, which often lack the cogency or the tractability of 'plain' backward induction. Identifying a solution concept for general games that distills precisely its logic, *and nothing else*, is thus important to recover the virtues of backward induction in contexts where standard equilibrium concepts lack in tractability, or are hard to justify, or fail to deliver solid economic insights. We discuss a few such cases in Section 6.

In pursuit of an answer to our main question, a good starting point is to inspect the solution concepts that are normally associated with the idea of "backward induction reasoning". Consider first Subgame Perfect Equilibrium (SPE). An influential argument in support of SPE is provided by Harsanyi and Selten (1988)'s notion of subgame consistency:

> "It is natural to require that a solution function for extensive games is subgame consistent in the sense that the behavior prescribed on a subgame is nothing else than the solution to the subgame" (ibid., p.90)

Subgame consistency warrants SPE the recursive structure of backward induction, i.e. the possibility of determining the solution concept's predictions for a subgame by looking at it 'in isolation'. Hence the possibility (in games with finite horizon) to solve for the SPE starting from the terminal nodes and proceeding backwards. This is extremely convenient, and certainly one of the main reasons for the prominence of SPE in applied work.

Several solution concepts extend the idea of SPE to games with incomplete information, often via the introduction of trembles (cf. Selten (1975), Kreps and Wilson (1982), etc.). In these solution concepts, trembles are a shortcut to formalize another idea that is typically associated with the logic of backward induction: that off-equilibrium moves are *mistakes*, unintended deviations.[1] The idea that unexpected moves are mistakes, which disrupt the implementation of one's plan of action, also provides conceptual motivation for the idea that the predictions for the continuation of the game shall only depend on the continuation game itself. In fact, we view these two complementary ideas as the building block of backward induction reasoning. Yet, while the incomplete information counterparts of SPE are typically considered to share its

[1]The view of deviations as 'mistakes' contrasts with the logic of forward induction, which requires instead that unexpected moves be rationalized (if possible) as purposeful deviations (e.g., Pearce (1984), Battigalli (1996)).

backward induction flavor, they do lack its recursive structure. Under Sequential Equilibrium, for instance, the set of predictions from an information set onwards cannot be computed by just looking at the continuation of the game, and neither can the game be solved "backwards". It is thus unclear in what sense, or to what extent, these concepts really are about backward induction reasoning, or what this even means in an incomplete information setting.

The objective of this paper is to identify a solution concept for general games, with possibly imperfect and incomplete information, that captures precisely the logic of backward induction reasoning, and nothing more. In particular, we look for a comprehensive answer that reconciles the following desiderata: (i) first, a recursive structure analogous to that of SPE; (ii) second, the ability to solve the game 'backwards'; (iii) third, a clear formalization of the idea of unexpected deviation as mistakes; (iv) fourth, a connection with a 'perfect equilibrium' concept that introduces backward induction logic and nothing more into equilibrium analysis.

To this end, we introduce *Backwards Rationalizability* ($\mathcal{BR}$ for short), a solution concept for belief-free games with incomplete and imperfect information, which consists of an iterated deletion procedure for the extensive form. At each round, a strategy is eliminated if it is *not* a sequential best response to any conjecture that, at each point in the game, is concentrated on opponents' continuation strategies which are consistent with the previous rounds of deletion. These continuation strategies need not be part of strategies that reach the current information set. With this, players entertain the possibility that the opponents committed *mistakes* in the past.[2] Thus, if an unexpected move of an opponent is interpreted as a mistake, it need not mean anything about her type, hence the inferences a player can draw about others' types, after observing an unexpected move, are unrestricted under $\mathcal{BR}$. This is the key reason why, besides satisfying a convenient *order independence* property (Theorem 1), $\mathcal{BR}$ also satisfies a property analogous to subgame consistency, which we call *continuation-game consistency*: the predictions of $\mathcal{BR}$ about the continuation play from any history onwards coincide with the predictions of $\mathcal{BR}$ in the (belief free) game that starts at that history (Theorem 2).

Continuation-game consistency is suggestive of the possibility, in finite horizon games, that the predictions of $\mathcal{BR}$ can also be computed by 'solving the game backwards'. Indeed, as we show (Theorem 3), the predictions of $\mathcal{BR}$ in these games can be computed by a convenient *backwards procedure*, which consists of the iterated application of belief-free rationalizability to the normal form of the continuation games from each information set considered "in isolation", starting from the end of the game and proceeding backwards.

We introduce next an equilibrium concept for dynamic Bayesian games, *interim perfect equilibrium* (IPE). Bayesian games are obtained appending a model of agents' beliefs, i.e. a type space, to the belief-free game. IPE is the weakest equilibrium notion for Bayesian games that is consistent with sequential rationality and with Bayesian updating, and it coincides with SPE in complete information games (see also Watson (2017)).[3] Furthermore, for reasons related to

[2]For complete information games, this epistemic justification of $\mathcal{BR}$ has indeed been formalized in a recent paper by Battigalli and De Vito (2021). This and other epistemic characterizations are discussed in Section 5.2.

[3]IPE is weaker, for instance, than the Perfect Bayesian Equilibrium notion recently introduced by Watson (2017), which also coincides with SPE under complete information.

the seminal result by Brandenburger and Dekel (1987), we show that the set of $\mathcal{BR}$ strategies in the belief-free game coincides with the set of all strategies that are played in some IPE for some type space (Theorem 4). Hence, $\mathcal{BR}$ characterizes the *robust predictions* of IPE, that is the predictions that do not depend on assumptions on players' exogenous beliefs about each other's types, as normally represented in a standard type space.

At a practical level, our results jointly imply that instead of computing the set of IPE by solving a large (possibly infinite, in fact) number of fixed point problems, one can compute the set of all IPE strategies by means of a tractable backwards procedure. This also shows that a property analogous to subgame consistency holds for the set of IPE strategies: the robust predictions of IPE are continuation-game consistent. As we discuss at the end of the paper, the tractability of the algorithm may prove useful in overcoming the difficulties typically faced in applications, both in complete and in incomplete information settings. At a conceptual level, our results reconcile all the main features that are informally associated with backward induction reasoning, including the recursive structure of the solution, the backwards solvability, and the idea of deviations as unintended mistakes. There is thus a precise sense in which IPE is the incomplete information counterpart of SPE that introduces the backward induction logic and nothing more into equilibrium analysis.

Finally, we discuss a few applications and extensions of our concepts. First, we propose a variation of *peer-confirming equilibrium* (Lipnowski and Sadler (2019)), a solution concept that combines equilibrium and non-equilibrium reasoning, whereby players have correct beliefs only regarding their neighbours in an exogenously given network. In static games, as the network becomes richer, the set of peer-confirming equilibria naturally shrinks, but this is not true in dynamic games, due to a tension in the solution concept between backward and forward induction reasoning. To correct this tension, we propose a variation of peer-confirming equilibrium, based on Backwards Rationalizability. We show that the logic of backward induction reasoning distilled by $\mathcal{BR}$ allows for a smoother integration of the equilibrium and non-equilibrium approaches, and restores in dynamic games the natural monotonicity result of static games. Then, we discuss other applications that are part of our published or ongoing work: Penta (2015) application of Backwards Rationalizability to problems of robust dynamic implementation; and Catonini and Penta (2022)'s extension of $\mathcal{BR}$ to solve a long-lasting puzzle in the industrial organization literature, the two-period Hotelling model of horizontal differentiation with linear transport costs (cf. Hotelling (1929), Osborne and Pitchik (1987)).

The rest of the paper is organized as follows. The next subsection discusses the main connections with the related literature. Section 2 introduces the framework of belief-free dynamic games. In Section 3 we define and analyze Backwards Rationalizability and the backwards procedure. Section 4 introduces Bayesian games and IPE. In Section 5 we discuss some properties and foundational aspects of our construction, and their significance with respect to the most closely related literature. Section 6 discussed the applications, and Section 7 concludes.

1.1 Related Literature

Backwards Rationalizability was first introduced by Penta (2010), for games with imperfect and incomplete information, and independently by Perea (2014), for games with complete information. Epistemic characterizations, which we further discuss in Section 5.2, have been provided by Perea (2014) and Battigalli and De Vito (2021) for complete-information games, and by Penta (2012*a*) for incomplete-information games. Applications are discussed in Section 6.

IPE was first introduced by Penta (2010) and applied by Penta (2015) to full implementation problems. As we will discuss, IPE provides a dynamic extension of interim equilibrium (Bergemann and Morris (2005)), and it is weaker than the notions of Perfect Bayesian Equilibrium (PBE) introduced by Fudenberg and Tirole (1991*b*) and by Watson (2017). Unlike other notions of weak PBE (e.g., Mas-Colell et al. (1995)), however, IPE does coincide with subgame-perfect equilibrium in games with complete information.

In terms of solution concepts, we innovate on the existing literature both for games with complete and incomplete information. For games with incomplete information, Backwards Rationalizability is a novel concept, and provides the first formal extension of backward induction reasoning to this class of games. Nonetheless, the backwards procedure we develop in Section 3.3, and the result that it characterizes Backwards Rationalizability (Theorem 3), are novel also within the special case of complete information games.[4] Furthermore, as we discuss in Section 6, the properties of our solution concept have substantial implications for important economic applications, including in settings with complete information. From a conceptual viewpoint, our analysis also sheds new light on some important aspects of backward induction reasoning, for instance on the role of the *belief persistence* hypothesis in this context (cf. Section 5).

Theorem 4, which shows an identity between Backwards Rationalizability and the set of IPE strategies across all type spaces, can be seen as a dynamic counterpart of the results in Battigalli and Siniscalchi (2003*a*) and Bergemann and Morris (2005), that relate (belief-free) Rationalizability with (interim) Bayesian Equilibrium. Those results, in turn, are incomplete information extensions of the characterization in Brandenburger and Dekel (1987), which brought to light the connection between (*a posteriori*) subjective correlated equilibrium (Aumann (1974)) and (correlated) Rationalizability (Pearce (1984), Bernheim (1984)). The robustness approach pursued in all these papers refers to the set of predictions across all possible type spaces, and it differs from a more recent approach which instead maintains a common prior type space to represent the minimal information of players, and seeks to characterize the set of equilibrium distributions if players have access to extra information (cf. Bergemann and Morris (2016), Bergemann and Morris (2013); Bergemann et al. (2015), Bergemann et al. (2017) – this approach has been brough to the data by Magnolfi and Roncoroni (2020). Dynamic counterparts of the latter approach have been provided by Doval and Ely (2020) and Makris and Renou (2018), who seek to bound or characterize the set of equilibrium distributions over a large class of extensive forms which are consistent with some minimal information about the game.

Backwards Rationalizability is also related to other non-equilibrium concepts for extensive

[4] In complete information games, Perea (2014) defines a "backwards dominance" procedure that, with the appropriate elimination order, can be used to obtain a superset of $\mathcal{BR}$.

form games, such as Extensive Form Rationalizability (Pearce (1984), Battigalli (1996)), and Interim Sequential Rationalizability (Penta (2012*b*)). As it will be explained, Backwards Rationalizability is weaker than the former in terms of outcomes (although not necessarily nested in terms of strategies), and it is stronger than the latter, which is based on Common Belief in Rationality only at the beginning of the game (an idea first due to Ben-Porath (1993), for complete information games). Zuazo-Garin (2017) studied sufficient conditions for the backward induction outcome under uncertainty over the extensive form. A more systematic analysis of the impact of higher order uncertainty about the observability of actions, which may or not induce backward induction outcomes, is provided by Penta and Zuazo-Garin (2021).[5]

2 Belief-Free Dynamic Games

We focus on finite multistage games with observable actions.[6] For each player $i \in N = \{1, ..., n\}$, A_i is the set of actions available to i at some point of the game. Let h^0 denote the initial history. At each non-terminal history h, all players i simultaneously choose an action from the non-empty set $A_i(h) \subseteq A_i$ (player i is actually inactive if $|A_i(h)| = 1$), so histories are sequences of action profiles. Let $\mathcal{H}$ denote the set of (publicly observed) non-terminal histories, and $\mathcal{Z}$ the set of terminal histories. The tree of all histories is endowed with the precedence relation $\prec$ (i.e., given two histories h, h', write $h \prec h'$ when h is a prefix of h').

Players' preferences over terminal nodes are parameterized by

$$\theta = (\theta_0, ..., \theta_n) \in \Theta = \Theta_0 \times ... \times \Theta_n;$$

Θ_0 is the set of *states of nature* and each Θ_i is the set of i's *payoff types*, all assumed finite. Player i privately observes θ_i at the beginning of the game; nobody observes θ_0. Each player i has payoff function $u_i : \mathcal{Z} \times \Theta \to \mathbb{R}$.

A *belief-free dynamic game* is a tuple

$$\Gamma = \langle N, \mathcal{H}, \mathcal{Z}, \Theta_0, (\Theta_i, u_i)_{i \in N} \rangle .$$

Note that this is not a Bayesian game, as Γ does not include a *type space*, i.e. a model of players' interactive beliefs about Θ and each others' beliefs. Type spaces and Bayesian games are introduced in Section 4.

A strategy is a function $s_i : \mathcal{H} \to A_i$ such that, for each $h \in \mathcal{H}$, $s_i(h) \in A_i(h)$. Let S_i denote the set of i's strategies. Any strategy profile $s \in S = \times_{i \in N} S_i$ induces a terminal history $\boldsymbol{z}(s) \in \mathcal{Z}$. The notation $\boldsymbol{z}(s|h)$ refers to the terminal history induced by strategy profile s,

[5] More broadly, Backwards Rationalizability is related to other versions of rationalizability for incomplete information games, such as: belief-free rationalizability (Battigalli and Siniscalchi (2003*a*), Bergemann and Morris (2005, 2009)); interim independent rationalizability (Ely and Peski (2006)); interim correlated rationalizability (Dekel et al. (2007), Fudenberg et al. (2006), also studied by Weinstein and Yildiz (2007, 2011), Oury and Tercieux (2012), and Penta (2013)); Δ-Rationalizability (Battigalli and Siniscalchi (2003*a*), also studied by Battigalli and Siniscalchi (2003*b*, 2007) and Ollár and Penta (2017, 2021)). For a unified perspective, see Battigalli et al. (2011).

[6] See Fudenberg and Tirole (1991*a*), chapters 3.2 and 8.2. At the expense of heavier notation, the analysis can be easily adapted to all finite dynamic games with perfect recall.

starting from history h. Strategic-form payoff functions can be defined for continuations from any given public history: for each $h \in \mathcal{H}$ and each $(s, \theta) \in S \times \Theta$, let $U_i(s, \theta; h) = u_i(\boldsymbol{z}(s|h), \theta)$ (For the initial history h^0, we will write $U_i(s, \theta)$ instead of $U_i(s, \theta; h^0)$. For each history h and player i, we let $S_i(h)$ denote the set of strategies of i that are compatible with h. Thus, upon reaching history h, player i learns that the behavior of the opponents is consistent with $S_{-i}(h) = \times_{j \neq i} S_j(h)$. (Note that $S_{-i}(h) = S_{-i}(h')$ when h and h' differ only by i's moves.) Finally, for each $h \in \mathcal{H}$, let S_i^h denote the set of strategies in the continuation game starting from h, and for each $s_i \in S_i$, let $s_i|h$ denote the continuation of s_i from history h.

3 Backwards Rationalizability

Backwards Rationalizability is a non-equilibrium solution concept for (dynamic) belief-free games. Similar to baseline Rationalizability (e.g., Pearce (1984)), also Backwards Rationalizability will be defined by an iterative deletion procedure, in which players form conjectures about others' information and behavior, and only entertain strategies that are optimal with respect to those conjectures. We thus need to first introduce a model of players' conjectures, as well as a notion of rationality, both of which will of course reflect the dynamic nature of the environment, and the possibility of incomplete information.

Conjectures: At every history, player i holds a conjecture about the state of nature and the opponents' payoff-types and behavior. These conjectures are represented by a Conditional Probability System, i.e., an array of conditional beliefs, one for each history, which are derived by updating whenever possible.[7] Let $\Theta_{-i} = \times_{j \neq i} \Theta_j$ and $S_{-i} = \times_{j \neq i} S_j$.

Definition 1. *A Conditional Probability System (CPS) over $\Theta_0 \times \Theta_{-i} \times S_{-i}$ is an array of conditional distributions $\mu^i = (\mu^i(\cdot|h))_{h \in \mathcal{H}}$ such that:*

C.1 *For every $h \in \mathcal{H}$, $\mu^i(\Theta_0 \times \Theta_{-i} \times S_{-i}(h)|h) = 1$;*

C.2 *For every h, h' with $h \prec h'$, for every $E \subseteq \Theta_0 \times \Theta_{-i} \times S_{-i}(h')$,*

$$\mu^i(E|h) = \mu^i(E|h') \cdot \mu^i(\Theta_0 \times \Theta_{-i} \times S_{-i}(h')|h). \tag{1}$$

The set of player i's CPSs is denoted by $\Delta_i^{\mathcal{H}}$.

Condition C.1 states that a player is always certain of what she knows at h; condition C.2 states that her beliefs are consistent with the chain rule of probability.

Sequential Rationality: Strategy s_i is sequentially rational for type θ_i given a CPS μ^i if, at each history $h \in \mathcal{H}$, it prescribes optimal behavior in the continuation game given $\mu^i(\cdot|h)$ – what

[7] The original notion of Conditional Probability System, due to Rényi (1955), requires beliefs to satisfy the chain rule, i.e. equation 1, whenever $S_{-i}(h') \subseteq S_{-i}(h)$, even when h and h' are not ordered. Battigalli et al. (2021) show that requiring the chain rule to hold only between ordered histories is meaningful and equivalent to the full-blown chain rule for various solution concepts, including Backwards Rationalizability.

we call *continuation best reply to* $\mu^i(\cdot|h)$. Formally: for any $\theta_i \in \Theta_i$, $s_i \in S_i$, $\mu^i \in \Delta_i^{\mathcal{H}}$, and $h \in \mathcal{H}$, let

$$\bar{U}_i\left(s_i; \mu^i, h, \theta_i\right) = \sum_{(\theta_0, \theta_{-i}, s_{-i}) \in \Theta_0 \times \Theta_{-i} \times S_{-i}(h)} U_i(s_i, s_{-i}, \theta_0, \theta_i, \theta_{-i}; h) \mu^i(\theta_0, \theta_{-i}, s_{-i}|h).$$

Definition 2. *Strategy s_i is sequentially rational for payoff-type θ_i given $\mu^i \in \Delta_i^{\mathcal{H}}$ if for each $h \in \mathcal{H}$ and $s_i' \in S_i$,*

$$\bar{U}_i\left(s_i; \mu^i, h, \theta_i\right) \geq \bar{U}_i\left(s_i'; \mu^i, h, \theta_i\right).$$

The set of sequentially rational strategies for θ_i given μ^i is denoted by $r_i(\mu^i, \theta_i)$. If $s_i \in r_i(\mu^i, \theta_i)$, we also say that μ^i justifies *s_i for θ_i.*

We can now introduce our main solution concept, Backwards Rationalizability ($\mathcal{BR}$):

Definition 3. *For each $i \in N$ and $\theta_i \in \Theta_i$, let $\mathcal{BR}_i^0(\theta_i) = S_i$. Recursively, for $k > 0$, let $\mathcal{BR}_{-i}^{k-1} := \{(\theta_j, s_j)_{j \neq i} \in \Theta_{-i} \times S_{-i} : \forall j \neq i, s_j \in \mathcal{BR}_j^{k-1}(\theta_j)\}$, and let $s_i \in \mathcal{BR}_i^k(\theta_i)$ if there exists $\mu^i \in \Delta_i^{\mathcal{H}}$ such that:*

1. *$s_i \in r_i(\mu^i; \theta_i)$;*
2. *for each $h \in \mathcal{H}$ and $(\theta_{-i}, s_{-i}) \in \Theta_{-i} \times S_{-i}$, if $\mu^i(\Theta_0 \times \{(\theta_{-i}, s_{-i})\}\,|h) > 0$, then there exists $s_{-i}' \in S_{-i}$ such that $s_{-i}'|h = s_{-i}|h$ and $(\theta_{-i}, s_{-i}') \in \mathcal{BR}_{-i}^{k-1}$.*

The set of Backwards Rationalizable *strategies for type θ_i is $\mathcal{BR}_i(\theta_i) = \cap_{k>0} \mathcal{BR}_i^k(\theta_i)$, and we let $\mathcal{BR}_i := \{(\theta_i, s_i) \in \Theta_i \times S_i : s_i \in \mathcal{BR}_i(\theta_i)\}$ and $\mathcal{BR} := \times_{i \in N} \mathcal{BR}_i$.*

In words, $\mathcal{BR}$ is an iterated deletion procedure. At each round, strategy s_i survives for type θ_i if it is justified by a CPS concentrated on opponents' *continuation* strategies that are consistent with the previous round of deletion. Players' conjectures about $\Theta_0 \times \Theta_{-i}$, however, are unrestricted. This property, which we call *unrestricted inference*, will play a crucial role for the interpretation of $\mathcal{BR}$ as backward induction reasoning, as we will show in Section 3.3.

Next, we illustrate $\mathcal{BR}$ with an example.

Example 1. *Ann ($i = a$) and Bob ($i = b$) are privately informed of the size $\theta_i = \{1, 2\}$ of their indivisible endowment. Ann can choose between a barter economy and a production economy. In the barter economy, players can commit to exchanging their endowments or not. Committing to exchange costs $\varepsilon \in (0, 1/2)$, and the exchange goes through only if both players commit. Setting up the production process costs $\gamma \in (\varepsilon, 1/2)$, the total production is $3\left(\theta_a + \theta_b\right)/2$, and it is equally shared between players. The figure displays Ann's payoffs (Bob's payoffs are symmetric).*

Barter (B):

$a \backslash b$	E	N
E	$\theta_b - \varepsilon$	$\theta_a - \varepsilon$
N	θ_a	θ_a

Production (P): $\frac{1}{2} \cdot \frac{3}{2}\left(\theta_a + \theta_b\right) - \gamma$

Backwards rationalizability works as follows.

At the first round, strategies B.E and P.E are not sequentially rational for type $\theta_a = 2$ of Ann, because choosing E at history (B) is not a continuation best reply to any belief. For Ann of type $\theta_a = 1$, instead, strategy B.N is not sequentially rational, because it is not a best reply to any belief at the beginning of the game: it yields a sure payoff of 1, whereas strategies P.E and P.N yield a payoff of at least $3/2 - \gamma > 1$. So we have

$$\begin{aligned} \mathcal{BR}_a^1(\theta_a = 1) &= \{B.E, P.E, P.N\} \\ \mathcal{BR}_a^1(\theta_a = 2) &= \{B.N, P.N\}. \end{aligned}$$

For Bob, at history (B), strategy E is dominated by N for type $\theta_b = 2$, but not for type $\theta_b = 1$. So we have

$$\begin{aligned} \mathcal{BR}_b^1(\theta_b = 1) &= \{E, N\} \\ \mathcal{BR}_b^1(\theta_b = 2) &= \{N\}. \end{aligned}$$

At the second round, for type $\theta_a = 1$ of Ann, strategies B.E and P.E are not sequential best replies to any belief $\mu^a \in \Delta_a^{\mathcal{H}}$ such that $\mu^a(\mathcal{BR}_b^1 | h^0) = 1$ (where we recall that $\mathcal{BR}_b^1 = \{(1,E),(1,N),(2,N)\}$), because they are not continuation best replies at history (B): they yield payoff $(1-\varepsilon)$ with probability 1, whereas choosing N yields a sure payoff of 1. So we have

$$\begin{aligned} \mathcal{BR}_a^2(\theta_a = 1) &= \{P.N\} \\ \mathcal{BR}_a^2(\theta_a = 2) &= \{P.N, B.N\}. \end{aligned}$$

Analogously, for Bob of type $\theta_b = 1$ at history (B) strategy E is not a best reply to any belief over $\mathcal{BR}_a^1|(B) = \{(1,E),(1,N),(2.N)\}$. So we have $\mathcal{BR}_b^2(\theta_b) = \{N\}$ for both $\theta_b = 1, 2$.

All the step-2 type-strategy pairs survive the third step of $\mathcal{BR}$. For both types of Bob and for type $\theta_a = 1$ of Ann, we are left with just one strategy, so it cannot be eliminated. In particular, for each type of Bob, choosing N is optimal for every belief over $\mathcal{BR}_a^2|(B) = \{(1,N),(2.N)\}$. For Ann of type $\theta_a = 2$, strategy B.N is a sequential best reply to every belief $\mu^a \in \Delta_a^{\mathcal{H}}$ such that $\mu^a((1,N)|h^0) = 1$, while strategy P.N is a sequential best reply to every belief $\mu^a \in \Delta_a^{\mathcal{H}}$ such that $\mu^a((2,N)|h^0) = 1$. In conclusion, we have that $\mathcal{BR}_a(\theta_a = 1) = \{P.N, B.N\}$ and $\mathcal{BR}_a(\theta_a = 2) = \{P.N\}$ for Ann, and $\mathcal{BR}_b(\theta_b) = \{N\}$ for both types $\theta_b = 1, 2$ of Bob. ▲.

3.1 Algorithmic properties

Since the game is finite, $\mathcal{BR}$ ends in finitely many steps. Hence:

Remark 1. *There exists $K \in \mathbb{N}$ such that $\mathcal{BR}^K = \mathcal{BR}$.*

It is also easy to check that $\mathcal{BR}$ satisfies the following standard fixed-point property:[8]

Remark 2. *For each $i \in N$ and $(\theta_i, s_i) \in \Theta_i \times S_i$, we have $(\theta_i, s_i) \in \mathcal{BR}_i$ if and only if $s_i \in r_i(\mu^i; \theta_i)$ for some $\mu^i \in \Delta_i^{\mathcal{H}}$ that satisfies the following property: for each $h \in \mathcal{H}$ and*

[8] Finiteness makes the property obvious, but it also holds in nicely-behaved infinite games, as standard.

(θ_{-i}, s_{-i}) *with* $\mu^i(\Theta_0 \times \{(\theta_{-i}, s_{-i})\}\,|h) > 0$*, there exists* $s'_{-i} \in S_{-i}$ *such that* $s'_{-i}|h = s_{-i}|h$ *and* $(\theta_{-i}, s'_{-i}) \in \mathcal{BR}_{-i}$.

$\mathcal{BR}$ is also robust to changes in the order of elimination of type-strategy pairs, which is helpful in practice. *Order independence* is also strongly related to the conceptually more relevant possibility of "solving the game backwards", as we will see in Section 3.3.

To formalize order independence, we rewrite $\mathcal{BR}$ as a reduction procedure, which is possible because it has *one-step memory*: computing $\mathcal{BR}^k$ only requires $\mathcal{BR}^{k-1}$, not $\mathcal{BR}^{k-2}, ..., \mathcal{BR}^0$. Fix $\hat{\Omega} = \times_{i \in N} \hat{\Omega}_i$, where every $\hat{\Omega}_i$ contains at least one element (θ_i, s_i) for each $\theta_i \in \Theta_i$. For each $i \in N$, let $\rho_i^{\mathcal{BR}}(\hat{\Omega})$ be the set of all $(\theta_i, s_i) \in \hat{\Omega}_i$ such that $s_i \in r_i(\mu^i; \theta_i)$ for some $\mu^i \in \Delta_i^{\mathcal{H}}$ that satisfies the following property: for each $h \in \mathcal{H}$ and $(\theta_{-i}, s_{-i}) \in \Theta_{-i} \times S_{-i}$, if $\mu^i(\Theta_0 \times \{(\theta_{-i}, s_{-i})\}\,|h) > 0$, then there exists $s'_{-i} \in S_{-i}$ such that $s'_{-i}|h = s_{-i}|h$ and $(\theta_{-i}, s'_{-i}) \in \hat{\Omega}_{-i}$. Finally, define the reduction operator $\rho^{\mathcal{BR}}(\hat{\Omega}) = \times_{i \in N} \rho_i^{\mathcal{BR}}(\hat{\Omega})$.

Definition 4. *An elimination order for* $\rho^{\mathcal{BR}}$ *is a chain* $\Omega = \hat{\Omega}^0 \supseteq \hat{\Omega}^1 \supseteq .... \supseteq \hat{\Omega}^M$ *such that:*

1. *for each* $m = 1, ..., M$, $\hat{\Omega}^m \supseteq \rho^{\mathcal{BR}}(\hat{\Omega}^{m-1})$;
2. $\rho^{\mathcal{BR}}(\hat{\Omega}^M) = \hat{\Omega}^M$.

$\mathcal{BR}$ is the maximal elimination order for $\rho^{\mathcal{BR}}$, that is, for each $k = 1, ..., \mathcal{BR}^k = \rho^{\mathcal{BR}}(\mathcal{BR}^{k-1})$. Any alternative elimination order $(\hat{\Omega}^m)_{m=0}^M$ is "slower" than $\mathcal{BR}$, in that at some step m, not every type-strategy pair that can be eliminated is actually eliminated: $\hat{\Omega}^m \supset \rho^{\mathcal{BR}}(\hat{\Omega}^{m-1})$.

To see that all elimination orders are equivalent, note that $\rho^{\mathcal{BR}}(\hat{\Omega}) \subseteq \rho^{\mathcal{BR}}(\hat{\Omega}')$ whenever $\hat{\Omega} \subset \hat{\Omega}'$. This monotonicity implies that "forgetting" to eliminate some type-strategy pair cannot result in a set $\hat{\Omega}^m$ that does not contain $\mathcal{BR}$. Morever, as long as $\hat{\Omega}^m$ is actually larger than $\mathcal{BR}$, it cannot have the fixed-point property highlighted in Remark 2: any set with this property would clearly survive all steps of $\mathcal{BR}$. Therefore, since an order of elimination can stop only when a fixed point is reached (see point 2 in Definition 4), the final output will coincide with $\mathcal{BR}$. This proves the following result (the formal proof is omitted):

Theorem 1. $\mathcal{BR}$ *is order-independent: for every order of elimination* $(\hat{\Omega}^m)_{m=0}^M$ *for* $\rho^{\mathcal{BR}}$*, we have* $\hat{\Omega}^M = \mathcal{BR}$.

As we mentioned, this is a technical property that is especially convenient when one needs to solve for an iterated deletion procedure. But, more importantly, this property is also useful for the main results of the next subsections, which provide two of the desiderata of backward induction reasoning that we discussed in the introduction; namely, continuation-game consistency and the backwards solution.

3.2 Continuation-game Consistency

In the Introduction, we identified the following distinctive feature of backward induction reasoning: the set of predictions for the whole game, when restricted to a part of the game, should coincide with the set of predictions for that part of the game analyzed in isolation. Does $\mathcal{BR}$

replicate this feature? The answer is affirmative. As customary, we take a "part of the game" to be everything that follows a certain history. In incomplete information games, however, a history does not define a "subgame", because it does not suffice as a starting point: the "initial conditions" shall also include the beliefs about payoff-types at that history. Instead of "subgame consistency", we thus look for *continuation-game consistency*: the predictions of $\mathcal{BR}$ from a history onwards shall coincide with the predictions of $\mathcal{BR}$ for a hypothetical game that starts at that history, hence under all possible initial belief about payoff types. The key intuition for why $\mathcal{BR}$ satisfies continuation-game consistency is the unrestricted inference property: According to $\mathcal{BR}$, every belief about payoff-types is possible at any history, because reaching the history is either the only rationalizable behavior of the opponents, or it can surprise our player, who is then free to infer whatever she wants about the opponents' payoff-types. To formalize, let $\mathcal{BR}^h$ denote $\mathcal{BR}$ for the game with root h. Also, for each player i, let

$$\mathcal{BR}_i|h = \{(\theta_i, s_i^h) \in \Theta_i \times S_i^h : \exists s_i \in S_i \text{ s.t. } (\theta_i, s_i) \in \mathcal{BR}_i \text{ and } s_i|h = s_i^h\},$$

and let $\mathcal{BR}|h = \times_{i \in N} \mathcal{BR}_i|h$.

Theorem 2. *For each $h \in \mathcal{H}$, for each $k \geq 0$, $\mathcal{BR}^k|h = \mathcal{BR}^{h,k}$.*

In words, after unexpected moves, players who reason according to $\mathcal{BR}$ can focus on the continuation of the game to predict the opponents' future behavior. This is what we call *continuation-game consistency*. In terms of epistemic conditions for $\mathcal{BR}$, this requires players to have common belief in their *future* rationality only (cf. Section 5.2).

The proof, which is provided in the Appendix, is based on a simple inductive argument. At each step k, every viable belief for $\mathcal{BR}_i^k$ can be replicated for $\mathcal{BR}_i^{h,k}$, by just taking its 'projection' in the continuation game that starts at h. Conversely, every viable belief for $\mathcal{BR}_i^{h,k}$ can be replicated for $\mathcal{BR}_i^k$, by attaching it to a CPS where h or an earlier history from which h is always reached comes as a surprise for i. In both directions, one obtains the same continuation best replies, at h and onwards.

Continuation-game consistency also suggests that, when reasoning about the overall game, players can anticipate the $\mathcal{BR}$-solution of the continuation game that follows some future history, and hence solve the game backwards, starting from preterminal histories. The next result will formalize this intuition.

3.3 The Backwards Procedure

Continuation-game consistency and order independence provided important clues for the possibility of computing the predictions of $\mathcal{BR}$ by "solving the game backwards". To verify this, we first need to clarify what it means to solve backwards a game with imperfect and incomplete information. In absence of any assumption of equilibrium, we solve the game backwards with a recursive use of belief-free rationalizability on the normal form of the continuation games that follow any given history, starting from preterminal histories and proceeding backwards. The normal form of each continuation game with root h will be first reduced to the type-strategy pairs that are consistent with the solution of the smaller continuation games.

We recall briefly the definition of belief-free rationalizability, which is useful to define also for strategic forms that consist of subsets of type-strategy pairs of the original game. So, fix a strategic form $G = \langle N, \Theta_0, (\widetilde{\Omega}_i, U_i)_{i\in N}\rangle$, where each $\widetilde{\Omega}_i \subseteq \Theta_i \times S_i$ is a subset of type-strategy pairs of player i, and $U_i : \Theta \times S \to \mathbb{R}$. For every i, let $R_i^0 := \widetilde{\Omega}_i$, and recursively, for every $k > 0$, $R_i^k := \{(\theta_i, s_i) \in \widetilde{\Omega}_i : \exists \nu \in \Delta(\Theta_0 \times R_{-i}^{k-1}) \text{ s.t. } s_i \in \hat{r}_i(\nu; \theta)\}$, where $\hat{r}_i(\nu; \theta_i)$ denotes the set of strategies that are a best response for type θ_i to a conjecture $\nu \in \Delta(\Theta_0 \times \Theta_{-i} \times S_{-i})$.[9] Then, for each $i \in N$, the set of *belief-free rationalizable type-strategy pairs* is $R_i := \cap_{k>0} R_i$.

Our *Backwards Procedure*, $\mathcal{BP}$, is formally defined as follows:[10]

Definition 5. *For each preterminal history h, define $\mathcal{BP}^h = \times_{i\in N}\mathcal{BP}_i^h$ as the output of belief-free rationalizability on $\times_{j\in N}(\Theta_j \times S_j^h)$.*

Moving backwards, fix now a history h and suppose that $\mathcal{BP}^{h'}$ was defined for every immediate successor h' of h. For each player i, let $\mathcal{BP}_i^{h,0}$ denote the set of pairs $(\theta_i, s_i^h) \in \Theta_i \times S_i^h$ such that $\left(\theta_i, s_i^h | h'\right) \in \mathcal{BP}_i^{h'}$ for every immediate successor h' of h. We define $\mathcal{BP}^h$ as the output of belief-free rationalizability on the strategic form that consists of the type-strategy pairs $(\mathcal{BP}_j^{h,0})_{j\in N}$.

The next example illustrates the procedure.

Example 2. *Consider again the game of Example 1. The backwards procedure works as follows:*

First we start from applying belief-free rationalizability to the continuation game with root $h = (B)$, the preterminal history. The game is symmetric: for each player $i = a, b$, if $\theta_i = 2$, E is dominated by N, so we have

$$\mathcal{BP}_i^{h,1} = \{(1, N), (2, N)\}.$$

Next, for every belief over $\mathcal{BP}_j^{h,1}$ $(j \neq i)$ E yields payoff $1 - \varepsilon$ with probability 1, whereas N yields a sure payoff of 1, so we have

$$\mathcal{BP}_i^{h,2} = \{(1, N), (2, N)\} = \mathcal{BP}_i^h.$$

Proceeding backwards, we move to the beginning of the game, $h = h^0$, and we initialize belief-free rationalizability with the set of type-strategy pairs whose continuations are belief-free rationalizable following B. That is:

$$\begin{aligned} \mathcal{BP}_a^0 &= \{(1, B.N), (1, P.N), (2, B.N), (2, P.N)\}, \\ \mathcal{BP}_b^0 &= \{(1, N), (2, N)\}. \end{aligned}$$

For Ann of type $\theta_a = 1$, strategy $B.N$ is not a best reply to any belief, because it yields a sure payoff of 1, whereas strategies $P.E$ and $P.N$ yield a payoff of at least $3/2 - \gamma > 1$. For ann of type $\theta_a = 2$, strategy $B.N$ is a best reply to a belief that assigns probability 1 to $(1, N)$, and strategy $P.N$ is a best reply to a belief that assigns probability 1 to $(2, N)$. No type-strategy pair can be eliminated for Bob, as we are already left with just one strategy for each type, therefore

[9] Formally, $\hat{r}_i(\nu; \theta_i) := \{argmax_{s_i \in S_i} \sum U_i(s_i, s_{-i}, \theta_i, \theta_{-i}, \theta_0) \cdot \nu(\theta_0, \theta_{-i}, s_{-i})\}$.

[10] A definition of $\mathcal{BP}$ that also includes the steps of belief-free rationalizability is provided in the Appendix.

no further type-strategy pair can be eliminated for Ann at the second step. In conclusion,

$$\begin{aligned}\mathcal{BP}_a &= \{(1, P.N), (2, B.N), (2, P.N)\}, \\ \mathcal{BP}_b &= \{(1, N), (2, N)\}.\end{aligned}$$

Note that $\mathcal{BP}_a = \mathcal{BR}_a$ *and* $\mathcal{BP}_b = \mathcal{BR}_b$. ▲

As noted, in this example $\mathcal{BP}$ yields exactly the predictions of $\mathcal{BR}$ in terms of behavior of every single player, from every history onwards. The next result shows that in fact this is a general property. The only possible difference between the two concepts is that, unlike $\mathcal{BP}$, in some games $\mathcal{BR}$ may exclude certain combinations of behavior of player i from h onwards and from a later history h' onwards, when i's behavior alone precludes reaching h' from h. Hence, the two concepts do not necessarily coincide in terms of full strategies, but this difference does not affect the actual predictions on players' behavior in any continuation game. To formalize this, we introduce the notion of realization-equivalence of continuation strategies: given a continuation strategy s_i^h, the realization-equivalent class $\left[s_i^h\right]$ is the set of all strategies $\widetilde{s}_i^h \in S_i^h$ that, for every $s_{-i} \in S_{-i}^h$, yield the same terminal history as s_i^h. We also write $\left[(\theta_i, s_i^h)\right]$ for $\{\theta_i\} \times \left[s_i^h\right]$, and given a subset $\widetilde{\Omega}_i \subseteq \Theta_i \times S_i^h$, we let $[\widetilde{\Omega}_i] = \cup_{\omega_i \in \widetilde{\Omega}_i} [\omega_i]$.

Theorem 3. *For each* $h \in \mathcal{H}$*, for each* $i \in N$, $[\mathcal{BR}_i|h] = \left[\mathcal{BP}_i^h\right]$.

In words, for every continuation game, the strategies that survive the backwards procedure for a type are realization-equivalent to the backwards rationalizable ones. Thus, while $\mathcal{BP}$ may include more strategies than $\mathcal{BR}$, the extra strategies would only differ for the behavior they entail at histories h' which are prevented from being reached by the strategies themselves – hence, they are realization-equivalent to strategies in $\mathcal{BR}$. Furthermore, if one conditions on h' the entire sets $\mathcal{BR}_i$ and $\mathcal{BP}_i$, the resulting sets of continuation strategies are still realization-equivalent from h' onwards. Hence, effectively, the possible behavior of each player in each continuation game is exactly the same under the two solution concepts. [11]

The proof of Theorem 3, which is provided in the Appendix, combines continuation-game consistency and order independence. Continuation-game consistency allows to focus on $\mathcal{BR}^h$ in place of $\mathcal{BR}|h$ for the comparison with $\mathcal{BP}^h$. By order independence, we can focus on a slow elimination order for $\mathcal{BR}^h$, where the strategies that are not continuation best replies at histories that follow h for a type are iteratively eliminated first. This order of elimination yields type-strategy pairs whose continuations after each h' that immediately follows h coincide with $\mathcal{BR}^{h'}$. Assuming $\left[\mathcal{BR}^{h'}\right] = [\mathcal{BP}_i^{h'}]$ by induction from the bottom of the game, we have thus mirrored (in terms of realization-equivalent classes) the initialization of $\mathcal{BP}^h$. Then, the equivalence between $\mathcal{BP}^h$ and $\mathcal{BR}|h$ is preserved step by step as we carry out $\mathcal{BP}^h$ on one side, and we move on to the elimination of strategies that are not continuation best replies at h on the other side.

[11] This point is further explained in Section 5.3.

4 Interim Perfect Equilibrium

This section explores the connection between $\mathcal{BR}$ and equilibrium predictions. In an incomplete information game, equilibrium means that players have correct beliefs about how the opponents would play given their payoff-relevant information and their belief hierarchy about the payoff-relevant information of everyone. At the beginning of the game, these belief hierarchies cannot be determined endogenously from equilibrium conditions; they must be specified exogenously. To do this, we follow the traditional approach (Harsanyi (1967)) of modeling the belief hierarchies implicitly, by means of *type spaces*. Appending a type space to a game yields a *Bayesian game*.

4.1 Bayesian Games

Because the game is finite, it is without loss of generality for equilibrium predictions to focus on finite type spaces. For each player i, fix a finite set T_i of types. Fix also an onto function $\vartheta_i : T_i \to \Theta_i$ assigning the payoff-type to each type. The *belief map* $\tau_i : T_i \to \Delta(\Theta_0 \times T_{-i})$, where $T_{-i} = \times_{j \neq i} T_j$, specifies the initial belief of each type t_i about state of nature and opponents' types. We call $\mathcal{T} = (T_i, \vartheta_i, \tau_i)_{i \in N}$ a (Θ-based) *type space*.

A Bayesian Game is obtained by appending the type space $\mathcal{T}$ to the belief-free game Γ:

$$\Gamma^{\mathcal{T}} = \langle N, \mathcal{H}, \mathcal{Z}, \Theta_0, (T_i, \vartheta_i, \tau_i, \Theta_i, u_i)_{i \in N} \rangle .$$

We write $\tau_i(\theta_0, t_{-i} | t_i)$ for the probability that type t_i assigns to (θ_0, t_{-i}).

In a Bayesian game, we call *interim strategies* the elements of S_i, and *interim mixed strategies* the elements of $\Delta(S_i)$.[12] We will also use *replacement plans*: given an interim strategy s_i and a history h, let $\varrho_{i,h}(s_i)$ denote the interim strategy $s_i' \in S_i(h)$ such that $s_i'(h') = s_i(h')$ for every $h' \not\prec h$. We call just *strategies* the functions $b_i : T_i \to \Delta(S_i)$ that assign an interim mixed strategy to each epistemic type. We write $b_i(s_i | t_i)$ for the probability that $b_i(t_i)$ assigns to the interim strategy $s_i \in S_i$.

4.2 Interim Perfect Equilibrium

In a dynamic Bayesian game, an equilibrium is described as a strategy profile coupled with systems of beliefs about state of nature and opponents' types. These belief systems must be part of the definition of equilibrium, because the beliefs after a deviation from equilibrium behavior are not pinned down by initial beliefs and equilibrium conditions. Formally, we introduce a map $p_i : T_i \to (\Delta(\Theta_0 \times T_{-i}))^{\mathcal{H}}$ that associates each type of player i with an array of beliefs about state of nature and opponents' types, one for each history. We write $p_i(\theta_0, t_{-i} | h; t_i)$ for the probability that $p_i(t_i)$ assigns to (θ_0, t_{-i}) at history h.

An *assessment* consists of a strategy profile $b = (b_i)_{i \in N}$ and a profile of belief systems $p = (p_i)_{i \in N}$. Given an assessment (b, p), for each $i \in N$ and $t_i \in T_i$, it will be useful to derive a

[12] We use mixed strategies in place of behavior strategies because it is notationally more convenient. Given a type space, there could be more IPE in mixed strategies than in behavior strategies. However, our equivalence between $\mathcal{BR}$ and IPE across types spaces would hold also if we restrict the attention to IPE in behavior strategies, because the IPE we construct for the proof actually uses only pure strategies.

CPS $\hat{\mu}^{t_i}_{(b,p)} = (\hat{\mu}^{t_i}_{(b,p)}(\cdot|h))_{h\in\mathcal{H}}$ over $\Theta_0 \times T_{-i} \times S_{-i}$ through the following recursive procedure.

For each $(\theta_0, (t_j)_{j\neq i}, (s_j)_{j\neq i}) \in \Theta_0 \times T_{-i} \times S_{-i}$, let

$$\hat{\mu}^{t_i}_{(b,p)}(\theta_0, (s_j, t_j)_{j\neq i}|h^0) = p_i(\theta_0, (t_j)_{j\neq i}|h^0; t_i) \cdot \prod\nolimits_{j\neq i} b_j(s_j|t_j). \tag{2}$$

Now fix $h \neq h^0$, let $p(h)$ denote the immediate predecessor of h, and suppose that $\hat{\mu}^{t_i}_{(b,p)}(\cdot|p(h))$ was defined. If $\hat{\mu}^{t_i}_{(b,p)}(\Theta_0 \times T_{-i} \times S_{-i}(h)|p(h)) > 0$, for each $\omega \in \Theta_0 \times T_{-i} \times S_{-i}(h)$, let

$$\hat{\mu}^{t_i}_{(b,p)}(\omega|h) = \frac{\hat{\mu}^{t_i}_{(b,p)}(\omega|\widetilde{h})}{\hat{\mu}^{t_i}_{(b,p)}(\Theta_0 \times T_{-i} \times S_{-i}(h)|\widetilde{h})}, \tag{3}$$

otherwise, for each $(\theta_0, (t_j)_{j\neq i}, (s_j)_{j\neq i}) \in \Theta_0 \times T_{-i} \times S_{-i}$, let

$$\hat{\mu}^{t_i}_{(b,p)}(\theta_0, t_{-i}, s_{-i}|h) = p_i(\theta_0, (t_j)_{j\neq i}|h; t_i) \cdot \prod\nolimits_{j\neq i} b_j(\varrho^{-1}_{j,h}(s_j)|t_j). \tag{4}$$

In the CPS $\hat{\mu}^{t_i}_{(b,p)}$, the beliefs about types and strategies of the opponents are derived from the assessment at the beginning of the game (eq. 2) and after every unexpected move (eq. 4). At the other histories, the beliefs are derived by updating (eq. 3).

For consistency of the assessment with the context described by the type space, one shall require that the initial beliefs specified by p_i (hence by each $\hat{\mu}^{t_i}_{(b,p)}(\cdot|h^0)$) coincide with the ones specified by τ_i. For internal consistency, one shall require that, for each t_i, the beliefs specified by $p_i(t_i)$ at each history are updated in view of b_{-i} whenever possible, as in $\hat{\mu}^{t_i}_{(b,p)}$.[13]

Definition 6. *An assessment (b,p) is* weakly pre-consistent *if, for every $i \in N$, $t_i \in T_i$, and for every $(\theta_0, t_{-i}) \in \Theta_0 \times T_{-i}$:*

1. $p_i(\theta_0, t_{-i}|h^0; t_i) = \tau_i(\theta_0, t_{-i}|t_i)$,
2. $p_i(\theta_0, t_{-i}|h; t_i) = \hat{\mu}^{t_i}_{(b,p)}(\{(\theta_0, t_{-i})\} \times S_{-i}|h)$ *for each* $h \in \mathcal{H}$.

Given the CPS $\hat{\mu}^{t_i}_{(b,p)}$, we can derive a CPS $\mu^{t_i}_{(b,p)}$ over the payoff-relevant uncertainty $\Theta_0 \times \Theta_{-i} \times S_{-i}$ as follows: for each $h \in \mathcal{H}$ and $(\theta_0, \theta_{-i}, s_{-i}) \in \Theta_0 \times \Theta_{-i} \times S_{-i}$, let

$$\mu^{t_i}_{(b,p)}(\theta_0, \theta_{-i}, s_{-i}|h) = \hat{\mu}^{t_i}_{(b,p)}(\{\theta_0\} \times \vartheta^{-1}_{-i}(\theta_{-i}) \times \{s_{-i}\}\,|h), \tag{5}$$

where $\vartheta^{-1}_{-i}((\theta_j)_{j\neq i}) = \times_{j\neq i}\vartheta^{-1}_j(\theta_j)$.

When (b,p) is an equilibrium, the CPSs $\mu^{t_i}_{(b,p)}$ constitute the equilibrium beliefs over the relevant uncertainty. With this, the definition of Interim Perfect Equilibrium is straightforward.

Definition 7. *An assessment (b,p) is an Interim Perfect Equilibrium of $\Gamma^{\mathcal{T}}$ if:*

1. *it is weakly pre-consistent;*
2. *for all $i \in N$, $t_i \in T_i$, and $s_i \in S_i$ with $b_i(s_i|t_i) > 0$, s_i is sequentially rational under $\mu^{t_i}_{(b,p)}$.*

[13] When this is impossible, note that $p_i(\cdot|h; t_i)$ can be arbitrary and $\hat{\mu}^{t_i}_{(b,p)}(\cdot|h)$ is consistent with it by definition, so that at the later histories beliefs are updated from the actual belief at h.

Example 3. *Append to the game of Example 1 a type space* $\mathcal{T} = (T_i, \vartheta_i, \tau_i)_{i=a,b}$ *where, for every* $i = a, b$, $T_i = \{t_i^1, t_i^2\}$ *and* $\vartheta_i(t_i^k) = k$ *for each* $k = 1, 2$. *We study the set of IPE, as the belief maps* $(\tau_i)_{i=a,b}$ *vary.*

Let $(b_i, p_i)_{i=a,b}$ *be a candidate IPE. For type* t_a^2 *of Ann, at history* (B), *action* N *is dominant, therefore we must have* $b_a(t_a^2) \in \Delta(\{B.N, P.N\})$. *Then, for Bob, for each* $k = 1, 2$, *we have*

$$\hat{\mu}_{(b,p)}^{t_b^k}(\{(t_a^1, B.E), (t_a^1, B.N), (t_a^2, B.N)\} \,|h = (B)) = 1.$$

Thus, the payoff of strategy E *is* $1 - \varepsilon$, *whereas the payoff of strategy* N *is* k. *Hence, we must have* $b_b(t_b^k) = N$. *Then, for each* $k = 1, 2$, *we have*

$$\hat{\mu}_{(b,p)}^{t_a^k}(\{(t_b^1, N), (t_b^2, N)\} \,|h^0) = 1.$$

It follows that, for type t_a^1 *of Ann, the only sequential best reply is* $P.N$, *thus* $b_a(t_a^1) = P.N$. *There only remains to determine* $b_a(t_a^2)$. *This depends on* $\tau_a(t_a^2)$. *Note indeed that, by weak preconsistency,*

$$p_a(t_b^k | h^0; t_a^2) = \tau_a(t_b^k | t_a^2), \quad k = 1, 2$$

and by construction of $\hat{\mu}_{(b,p)}^{t_a^2}$,

$$\hat{\mu}_{(b,p)}^{t_a^2}(\{t_b^k\} \times S_b | h^0) = p_a(t_b^k | h^0; t_a^2), \quad k = 1, 2.$$

Therefore, strategy $P.N$ *is optimal for* t_a^2 *if*

$$\frac{3}{4}\left(2 + (2\tau_a(t_b^2|t_a^2) + \tau_a(t_b^1|t_a^2))\right) - \gamma \geq 2,$$

while strategy $B.N$ *is optimal with the opposite weak inequality. Thus, we get*

$$\begin{cases} b_a(t_a^2) = P.N & \text{if } \tau_a(t_b^2|t_a^2) > \frac{4}{3}\gamma - \frac{1}{3} \\ b_a(t_a^2) = B.N & \text{if } \tau_a(t_b^2|t_a^2) < \frac{4}{3}\gamma - \frac{1}{3} \\ b_a(t_a^2) \in \Delta(\{P.N, B.N\}) & \text{if } \tau_a(t_b^2|t_a^2) = \frac{4}{3}\gamma - \frac{1}{3} \end{cases}.$$

This completes the characterization of the IPE of the game, as a function of players' beliefs. ▲

IPE can be seen as a dynamic counterpart of interim equilibrium, as defined in Bergemann and Morris (2005), obtained by imposing two natural conditions: (i) weak pre-consistency of the belief system, and (ii) sequential rationality. Notice that weak preconsistency imposes no restrictions on the beliefs held at histories that receive zero probability at the preceding node. Hence, even if agents' initial beliefs admit a common prior, IPE is weaker than the notions of Perfect Bayesian Equilibrium (PBE) introduced by Fudenberg and Tirole (1991*b*) and by Watson (2017). However, unlike other notions of weak PBE (see, e.g., Mas-Colell et al. (1995)), IPE requires players' beliefs to be consistent with Bayesian updating also off-the-equilibrium path. Hence, in complete information games, IPE does coincide with subgame-perfect equilibrium.

4.3 Characterization of the set of IPE

The set of IPE of a Bayesian game crucially depends on the exogenous hierarchies of beliefs about payoff-relevant types. By contrast, $\mathcal{BR}$ is a belief-free solution concept, where no exogenous structure on beliefs is imposed. Moreover, IPE is an equilibrium concept, thus players are assumed to hold correct beliefs about how the opponents will behave, conditional on their possible types. By contrast, $\mathcal{BR}$ is a rationalizability procedure, where beliefs are purely subjective and need not be correct. Despite these important differences, is it possible to draw a connection between the two solution concepts?

There are two clues in our previous analysis that suggest a connection. First, the $\mathcal{BR}$ set has a fixed point property, by Remark 2. This means that all backward rationalizable type-strategy pairs can be justified under the correct belief that, from every point on, the opponents will follow backwards rationalizable continuation strategies. Second, IPE is based on type spaces which in principle may contain types with different beliefs about the types of others. Therefore, different types can have different beliefs about the payoff-types and the strategies of the opponents, and hence optimally play differently (even when they have the same payoff-type). There remains the important difference that the type structure fixes a subset of hierarchies of beliefs about payoff-types, while such beliefs are free in $\mathcal{BR}$. But this difference can be overcome by looking at the set of *all* possible IPE across *all* type structures.

The conclusion is that there is indeed a very strong connection between the two solution concepts: $\mathcal{BR}$ yields exactly the set of IPE strategies across all type structures.

Theorem 4. *Fix a belief-free game Γ. For each $i \in N$, $(\theta_i, s_i) \in \mathcal{BR}_i$ if and only if there exists a type space $\mathcal{T}$, an IPE (b^*, p^*) of $\Gamma^{\mathcal{T}}$, and a type $t_i \in T_i$ s.t. $\vartheta_i(t_i) = \theta_i$ and $b_i^*(s_i|t_i) > 0$.*

Thus, $\mathcal{BR}$ characterizes the set of predictions on players' strategies that are consistent with IPE, but which do not depend on exogenous restrictions on the type space. In that sense, $\mathcal{BR}$ characterizes the *robust predictions* of IPE, across all the exogenous hierarchies of beliefs.

The proof of Theorem 4 is in the Appendix. The 'if' direction is straightforward: given an IPE, all the type-interim strategy pairs that can be derived from the IPE strategies iteratively survive $\mathcal{BR}$ because they "justify each other". The 'only if' direction is proven through the construction of just one type space that comprises all the backwards rationalizable type-strategy pairs. In particular, for each payoff-type and each associated interim strategy, we construct a type with initial belief over the opponents' types that mirrors the belief we obtain for the pair from the fixed-point property of $\mathcal{BR}$ (see Remark 2).

5 Discussion

In this section we discuss some important or subtle aspects of our concepts and results.

5.1 Complete Information, Redundant Types and IPE

In games with complete and perfect information and no relevant ties, Backwards Rationalizability coincides with the backward induction solution, hence with SPE. The next example

(borrowed from Perea (2014)) shows that if the game has complete but imperfect information, the set of strategies played in the SPE of the game may be a strict subset of $\mathcal{BR}$:

Example 4. *Consider the game in the following figure:*

$\mathcal{P}\ell.\ 1$

$Out \swarrow$ $\qquad$ $\searrow In$

$4, 0$

$1\backslash 2$	f	g	h
c	$2, 3$	$5, 1$	$2, 0$
d	$3, 1$	$2, 3$	$2, 0$
e	$1, 4$	$1, 3$	$6, 0$

In this game, $\mathcal{BR}_1 = \{Out.c, Out.d, In.c\}$ *and* $\mathcal{BR}_2 = \{f, g\}$. *The game, however, has only one SPE, in which player* 1 *chooses Out: in the proper subgame, the only Nash equilibrium entails the mixed (continuation) strategies* $\frac{1}{2}c + \frac{1}{2}d$ *and* $\frac{3}{4}f + \frac{1}{4}g$, *yielding a continuation payoff of* $\frac{11}{4}$ *for player* 1. *Hence, player* 1 *chooses Out at the first node.* ▲

In games with complete information, IPE coincides with SPE, but $\mathcal{BR}$ in general is weaker than SPE. At first glance, this may appear in contradiction with Theorem 4, which says that $\mathcal{BR}$ characterizes the set of strategies played in IPE across models of beliefs. The reason is that IPE is a solution concept for Bayesian games (i.e., for pairs $\langle \Gamma, \mathcal{T} \rangle$), and even if the environment has no payoff uncertainty (i.e., if Θ is a singleton), the complete information model in which T_i is a singleton for every i is not the only possible one: models with *redundant types* may exist, for which the IPE strategies differ from the SPE strategies that are played in the complete information model. The source of the discrepancy is analogous to the one between Nash equilibrium and subjective correlated equilibrium (Aumann (1974); see also Brandenburger and Dekel (1987)), with the type space playing the role of the correlating device.[14]We illustrate the point by constructing a type space and an IPE in which strategy $In.c$ is played by some type of player 1. Consider a type space $\hat{\mathcal{T}}$ such that $\hat{T}_1 = \{t_1^{Out.c}, t_1^{Out.d}, t_1^{In.c}\}$ and $\hat{T}_2 = \{t_2^f, t_2^g\}$, with the following beliefs:[15]

$$\begin{aligned} \tau_1(t_2^f | t_1) &= \begin{cases} 1 & \text{if } t_1 = t_1^{Out.d} \\ \frac{1}{2} & \text{if } t_1 = t_1^{Out.c} \\ 0 & \text{if } t_1 = t_1^{In.c} \end{cases}, \\ \tau_2(t_1^{Out.d} | t_2^g) &= 1, \text{ and } \tau_2(t_1^{Out.c} | t_2^f) = 1. \end{aligned}$$

The equilibrium strategy profile b is such that, for each player i and type $t_i^{s_i}$, $b_i\left(t_i^{s_i}\right) = s_i$. The belief systems agree with the beliefs of the type space at the initial history. At history (In), the belief of player 1 remains the same by updating, whereas the belief of player 2 must be revised,

[14]For more on the effects that redundant types may have on expanding the set of predictions for solution concepts that incorporate conditional independence restrictions on agent's conjectures (such as Bayes-Nash equilibrium, Interim Independent Rationalizability, etc.), see Ely and Peski (2006)

[15]It is easy to see that such a difference is not merely due to the possibility of zero-probability types. Also the relaxation of the common prior assumption is not crucial for this particular point.

but we can maintain the same belief each type had at the beginning of the game. Then, it is easy to verify that (b, p) is an IPE.

On the other hand, if Θ is a singleton and the game has *perfect information* (no stage with simultaneous moves), then $\mathcal{BR}$ does coincide with the set of SPE strategies. Hence, in environments with no payoff uncertainty and with perfect information, only SPE strategies are played as part of an IPE for any model of beliefs.

5.2 Epistemic characterizations

An earlier version of this paper (Penta (2012*a*)) showed that $\mathcal{BR}$ characterizes the behavioral implications of Rationality and Common Belief in Future Rationality (RCBFR). The same result was also independently provided by Perea (2014), but for complete information games. Common belief in future rationality means that, at every point in the game, there is common belief that everybody will follow an optimal continuation plan, without necessarily assuming that past moves were part of the same plan.

In complete information games with observable actions, Battigalli and De Vito (2021) construct an epistemic model where plans and actual play are formally distinguished. (Their epistemic model can also be extended to incomplete-information games.) With this, they show that $\mathcal{BR}$ characterizes the behavioral implications of the following epistemic hypotheses. First, players formulate an optimal plan for the entire game and execute it correctly. At every history, there is common belief that everybody has an optimal plan for the entire game, not just for the future. However, there is only common belief that everybody will correctly execute her plan in the future. Therefore, after observing an unexpected move, a player is free to believe that the move was carried out by mistake. On the other hand, we emphasize that these epistemic conditions do not rule out the possibility that the unexpected move *was* part of an optimal plan, just different than previously believed. We clarify this aspect in the next subsection.

5.3 On the $\mathcal{BR}$-$\mathcal{BP}$ comparison

To better understand the observations made in Section 3.3, on the possible difference between $\mathcal{BR}$ and $\mathcal{BP}$ in terms of full strategies, note that compared to $\mathcal{BP}$, $\mathcal{BR}$ further eliminates any strategy that is a continuation best reply to some belief at h, but not a continuation best reply to the *same* belief at some later history h' that is precluded by the strategy itself. Such combinations of continuation best replies to different beliefs are instead allowed by $\mathcal{BP}$, which uses normal form best replies in place of sequential rationality. In other words, some strategies in $\mathcal{BP}$ may not be dynamically consistent plans under the beliefs allowed by $\mathcal{BR}$, but these inconsistencies only occur after a deviation from the own plan and do not introduce any non-backwards rationalizable continuation plan.[16] This is the reason why, as we explained in Section 3.3, the possible behavior of each player in each continuation game is exactly the same under

[16] In a previous version of this paper, $\mathcal{BR}$ allowed players to change their beliefs after a deviation from their own plan, by using CPSs over strategy profiles instead of just opponents' strategies. In this way, the equality $\mathcal{BR} = \mathcal{BP}$ was established. Thus, modifying the CPS in the definition of $\mathcal{BR}$ in this way would weaken the solution concept without affecting its predictions in terms of outcomes, conditional on every history, and would establish the full equivalence with $\mathcal{BP}$.

the two solution concepts. We provide next an example of a game without payoff uncertainty in which the difference between $\mathcal{BR}$ and $\mathcal{BP}$ in terms of eliminations of strategies emerges (and makes $\mathcal{BP}$ much easier to compute than $\mathcal{BR}$).

Example 5. *Consider the following complete information game:*

$\mathcal{P}\ell.\ 2$

$Out \swarrow$ $\qquad$ $\searrow In$

$0, 4$ $\qquad$ $\mathcal{P}\ell.\ 1$

$\ell \swarrow$ $\qquad$ $\searrow r$

$3, 5$

$1\backslash 2$	w	m	e
n	$2,3$	$2,2$	$2,0$
c	$1,2$	$8,0$	$8,1$
s	$0,0$	$9,2$	$9,3$

Strategies $r.n$ and $\ell.s$ are not sequentially rational for player 1: the first is dominated by ℓ at history (In), the second is not a continuation best reply at (In, r) to any belief that makes ℓ optimal. All the strategies of player 2 are sequentially rational, therefore at the second step of $\mathcal{BR}$ no strategy of player 1 can be eliminated. Thus, consider the possible beliefs of player 2. There are two cases.

Case 1: *player 2 is initially certain that player 1 will play (a strategy that prescribes) ℓ. Then, player 2 will choose In, and upon observing r, player 2 must revise her belief. Every action of player 1 at history (In, r) is prescribed by some sequentially rational strategy, therefore player 2 can form any belief about the continuation play. As a consequence, all the strategies of player 2 that prescribe In survive the second step of $\mathcal{BR}$.*

Case 2: *player 2 gives positive initial probability to $\{r.c, r.s\}$. If this probability is large enough, player 2 will plan to choose Out, and if she unintentionally plays In and observes r, she must update her initial belief. Hence, she still cannot give positive probability to $r.n$. As a consequence, strategy $Out.m$ does not survive the second step of $\mathcal{BR}$ (while strategies $Out.w$ and $Out.e$ do). This elimination is however immaterial for the beliefs of player 1 at the third step of $\mathcal{BR}$, therefore all the remaining strategies are backwards rationalizable.*

Move now to $\mathcal{BP}$. In the simultaneous-moves subgame with root (In, r), every action is rationalizable. Consider thus the (non-reduced) subgame with root (In). Strategies $r.c$ and $r.s$ are normal-form best replies to sufficiently optimistic beliefs about the action of Player 2 at (In, r). Strategy $r.n$ is instead eliminated, because it yields a sure payoff of 2, while the strategies $\ell.n, \ell.c, \ell.s$ yield 3. The latter strategies all survive $\mathcal{BP}$, because they are normal-form best replies to a sufficiently pessimistic belief about the action of Player 2 at (In, r), including strategy $\ell.s$, which is not backwards rationalizable. Note however that there do exist backwards rationalizable strategies of Player 1 that prescribe s, namely $r.s$. Finally, move to the root of the game and consider the reduced strategic-form obtained after removing $r.n$. Every strategy of Player 2 is a strategic-form best reply to some belief: the strategies that prescribe Out are best replies to beliefs concentrated on the strategies of Player 1 that prescribe r, whereas the stategies that prescribe In are best replies to beliefs concentrated on the strategies of Player 1

that prescribe ℓ. Hence, all the strategies of Player 2 survive $\mathcal{BP}$, including strategy Out.m which is not backwards rationalizable, but again, there do exist backwards rationalizable strategies of Player 2 that prescribe Out or m. ▲

5.4 Belief persistence

Subgame perfect equilibrium embodies another idea which is commonly associated with backward induction reasoning: the idea of "belief persistence". Belief persistence means that players never change their belief about the strategies of the opponents, no matter how many deviations from the expected strategies they have observed. A possible way to interpret belief persistence is the following: Upon observing an unexpected move, players are *fully convinced* that the move was carried out by mistake, as a deviation from the optimal plan (they don't merely entertain such a possibility). This attitude, we argue, is a consequence of equilibrium play and not of backward induction reasoning per se: absent equilibrium restrictions, after being surprised a player may as well question her previous belief regarding the opponents' types and strategies, and form new beliefs by focusing on the continuation game. Indeed, some backward rationalizable strategies can only be justified without imposing the strong form of belief persistence.

Example 6. *Consider the game of Example 5. Recall that strategies $\ell.s$ and $r.n$ are not sequentially rational for player 1. Introduce now belief persistence: if player 2 is initially certain of ℓ but then observes r, she remains convinced that player 1 planned to choose ℓ but executed r by mistake. At the second step of reasoning, player 2 must give zero probability to $\{\ell.s, r.n\}$ at the beginning of the game. Therefore, if she is initially certain of ℓ, she must give probability 1 to $\{\ell.n, \ell.c\}$. Under belief persistence, this means that at history (In, r) she gives probability 1 to $\{r.n, r.c\}$. Therefore, she will not choose m. If she gives positive initial probability to $\{r.c, r.s\}$, she must give probability 1 to $\{r.c, r.s\}$ at history (In, r), but then again she has no incentive to play m. Hence, not just Out.m, but also In.m would not survive the second step of reasoning under belief persistence. However, In.m is backwards rationalizable. Therefore, imposing belief persistence refines the possible paths.* ▲

Backwards Rationalizability thus captures an *agnostic* attitude as to whether the unexpected moves of the opponents are mistakes or deliberate choices. Extensive-form rationalizability (Pearce (1984), Battigalli (1997)) captures instead the view that unexpected moves are definitely deliberate utility maximizing choices (if possible). By doing so, it refines $\mathcal{BR}$ (cf. Perea (2018), Catonini (2020)).[17] In contrast, *belief persistence* means that unexpected moves are definitely interpreted as mistakes; hence, restricting beliefs to satisfy belief persistence at every step of elimination would also refine $\mathcal{BR}$, albeit differently from EFR.

[17] These papers show that extensive-form rationalizability refines $\mathcal{BR}$ in terms of outcomes. But it also makes sense to define extensive-form rationalizability explicitly as a refinement of $\mathcal{BR}$, that is, initializing the procedure with the backward rationalizable strategies. In this way, both views of unexpected moves are captured, but differently than in $\mathcal{BR}$, with an *epistemic priority ordering* (Catonini (2019)): unexpected moves are interpreted as utility maximizing choices if possible, as mistakes otherwise. The proofs of Perea (2018) and Catonini (2020) indeed show that this procedure is outcome-equivalent to extensive-form rationalizability.

A natural question arises at this point: Given the lack of belief persistence, how is it possible that $\mathcal{BR}$ captures the robust implications of IPE, which – just like any standard equilibrium concept for Bayesian games – is based on the very notion of belief-persistence?

To see this, consider a game with no payoff uncertainty (i.e., Θ is a singleton), and let $\mathcal{T}$ denote a type space. If $\mathcal{T}$ contains only one type for every player, so that $\langle \Gamma, \mathcal{T} \rangle$ is a standard game with complete information, then IPE boils down to SPE. So, in the example above, $In.m$ cannot be played with positive probability in any IPE. However, as discussed in Section 5.1, even in a game without payoff uncertainty there can be many types of an opponent. For instance, similar to the type space $\hat{\mathcal{T}}$ in Section 5.1, one can think of a type for each of the backwards rationalizable strategies: while all such types would share the same (degenerate) belief hierarchies about the (commonly known) payoffs of the game, they would differ in their belief hierarchies about each others' strategies in the game.[18] Then, after observing an unexpected move, a player can change her belief about the *type* of the opponent, and hence also change her belief about his moves in the continuation game, despite keeping fixed the (correct) belief about how each type would play in the game.

This means that, from the viewpoint of an external analyst, *belief-persistence* only has bite insofar as the analyst has information about the precise set of *types* that players have in mind (that is, the 'mental states' that players may use to index others' behavior, with the associated beliefs), also when they revise their beliefs after observing an unexpected move. These are precisely the restrictions that are captured by the type space in a standard Bayesian game. If, however, the analyst does not wish to exogenously restrict such universe of conceivable types, and hence wishes to capture the set of all IPE-predictions across all possible type spaces, then the richness of the resulting type space voids the belief-persistence of IPE of any bite: the set of all such predictions is captured by a solution concept for belief-free games (namely, $\mathcal{BR}$) which does *not* satisfy belief-persistence.

6 Applications and Extensions

In this section we briefly discuss some applications of Backwards Rationalizability to illustrate its relevance and tractability. First, we apply backwards rationalizability to develop a variation of a recent work by Lipnowski and Sadler (2019), who put forward a solution concept that allows for a combination of equilibrium and non-equilibrium reasoning. In this context, we show that Backwards Rationalizability allows for a smoother integration of the two approaches, and for a natural extension of important properties of their solution concept from static to dynamic settings. Then, we discuss other applications that are part of our published or ongoing work.

6.1 Peer-Confirming Equilibrium with Backward Induction Reasoning

In a recent paper, Lipnowski and Sadler (2019) define the notion of *peer-confirming equilibrium* (PCE) for complete information games in which players are organized in a network. In a PCE, players have correct beliefs about the strategies of their neighbours; the beliefs about the other

[18] The proof of Theorem 4 shows how to construct such a type structure.

players are consistent with common belief in rationality and in correctness of beliefs about neighbours' play. In static games, PCE spans from Nash equilibrium, when the network is complete, to rationalizability, when the network is empty, and a nice monotonicity result holds: as the number of the connections in the network increases, the set of PCE shrinks.

Lipnowski and Sadler (2019) also apply their concept to dynamic games. Players are assumed to have correct beliefs about their neighbors also off-path, and these beliefs need not be consistent with forward induction reasoning. Thus, players display belief persistence towards their neighbors, and when the network is complete, PCE coincides with subgame perfect equilibrium. In contrast, when there are opponents who are not in a player's neighbourhood, then this players' beliefs about non-neighbors must be consistent, whenever possible and both on- and off-the-path, with common belief in rationality and correctness of beliefs about neighbours. Therefore, forward induction considerations ensue. As a result, when the network is empty, PCE coincides with *extensive-form rationalizability* (Pearce (1984)), and thus the monotonicity result from the static settings is not preserved: as it is well-known, subgame perfect equilibrium and extensive-form rationalizability yield non-nested predictions.

The reason behind the lack of monotonicity is the tension in PCE between the subgame perfect equilibrium logic that players apply to their neighbors and the forward induction logic that they apply to the other players. Without a way to capture the non-equilibrium implications of backward induction reasoning, this tension in Lipnowski and Sadler (2019) was in a way unavoidable: the key idea of PCE of weakening equilibrium restrictions only for non-neighbors translates into a hybrid of plain subgame perfect equilibrium and extensive form rationalizability, thereby mixing backwards and forward induction logic. Endowed with the tools we developed above, we propose next a modification of peer-confirming equilibrium that is entirely based on *backward induction reasoning*. As in Lipnowski and Sadler (2019), we maintain that players have correct beliefs about their neighbors, as well as the equilibrium view that a player never changes beliefs about their continuation play. Regarding the non-neighbors, instead, we drop belief persistence – which as argued pertains to an equilibrium logic, not to backward induction per se – but we maintain the view that anyone's unexpected moves may be regarded as mistakes, and hence they need not mean anything about their continuation play. As a result, our version of peer-confirming equilibrium (a solution concept we formally denote by $\mathcal{PC}$ below) spans from subgame perfect equilibrium, when the network is complete, to *backwards rationalizability*, when the network is empty. Thus, the monotonicity result is restored: peer-confirming equilibrium with backward induction reasoning (i.e., $\mathcal{PC}$) does become more restrictive as the network becomes richer.

Formally, for each player $i \in N$, let $N^i \subseteq N$ denote i's network neighbourhood, which includes her neighbours and herself. As in Lipnowski and Sadler (2019), we focus on games without payoff uncertainty, therefore we omit everywhere the sets of types.

Definition 8. *Let $\mathcal{PC}^0 = S$. For any $k > 0$, $s^* = (s_i^*)_{i \in N} \in \mathcal{PC}^k$ if and only if, for each $i \in N$, there exists $\mu^i \in \Delta_i^{\mathcal{H}}$ such that: (i) $s_i^* \in r_i(\mu^i)$; and (ii) for each $h \in \mathcal{H}$ and $s_{-i} \in \text{supp}\mu^i(\cdot|h)$, $s_{-i}|h = s'_{-i}|h$ for some $s' = (s'_j)_{j \in N} \in \mathcal{PC}^{k-1}$ such that $(s'_j)_{j \in N^i} = (s_j^*)_{j \in N^i}$. Then, the set of* peer-confirming equilibria with backward induction reasoning *is defined as $\mathcal{PC} := \cap_{k>0} \mathcal{PC}^k$.*

$\mathcal{PC}$ is an iterated elimination procedure for strategy *profiles*, rather than strategies. This is important because the candidate strategy profile restricts the viable beliefs of players: at every history, a player shall assign probability one to continuation strategies that are consistent with a strategy profile where all the neighbours (and herself) play as in the candidate profile. Without this restriction (that is, with the empty network), the focus on profiles becomes immaterial and $\mathcal{PC}$ coincides with plain $\mathcal{BR}$.

Remark 3. *If $N^i = \{i\}$ for every $i \in N$, then $\mathcal{PC} = \mathcal{BR}$.*

With the complete network, given a candidate profile s^*, each player i is forced to believe in $s^*_{-i}|h$ from every history h onwards. Therefore, $\mathcal{PC}$ boils down to the set of pure SPE of the game (which of course can be empty).

Remark 4. *If $N^i = N$ for every $i \in N$, then $\mathcal{PC}$ is the set of pure SPE.*

In the original definition of peer-confirming equilibrium of Lipnowski and Sadler (2019), the requirement on players' beliefs is split into two. The first requirement concerns the neighbours: the beliefs about their continuation play must coincide with the candidate profile. The second requirement concerns the other players: at every history h, the beliefs about their play must be consistent with strategy profiles of step $k-1$ that coincide after h with the candidate profile in i's neighbourhood *and reach* h, if any; otherwise, these beliefs are unrestricted. A richer network restrains the set of viable beliefs at the first step of reasoning, so that fewer profiles survive. However, fewer profiles reach fewer histories, therefore the second-step beliefs with the richer network need not be a subset of those with a poorer network. This is the source of the non-monotonicity in the original notion of peer-confirming equilibrium. By contrast, under backward induction reasoning, a smaller set of possible strategy profiles entails a smaller set of viable beliefs. This observation was key for the order independence of $\mathcal{BR}$, and is key here for the monotonicity of $\mathcal{PC}$ with respect to the network structure.

Theorem 5. *Suppose that $\hat{N}^i \supseteq \bar{N}^i$ for every $i \in N$. Let $\hat{\mathcal{PC}}$ and $\bar{PC}$ denote, respectively, $\mathcal{PC}$ under $(\hat{N}_i)_{i\in N}$ and under $(\bar{N}_i)_{i\in N}$. We have $\hat{\mathcal{PC}} \subseteq \bar{PC}$.*

Proof. By induction. The basis step, $\hat{\mathcal{PC}}^0 \subseteq \bar{PC}^0$, is trivial. Fix now $k > 0$ and suppose that $\hat{\mathcal{PC}}^{k-1} \subseteq \bar{PC}^{k-1}$. Fix $s^* \in \hat{\mathcal{PC}}^k$. We want to show that $s^* \in \bar{PC}^k$. Fix $i \in N$. Fix $\mu^i \in \Delta_i^{\mathcal{H}}$ such that μ_i and s_i^* satisfy requirement (ii) in Definition 8 with $\mathcal{PC}^{k-1} = \hat{\mathcal{PC}}^{k-1}$ and $N^i = \hat{N}^i$. Requirement (ii) is then satisfied also with $\mathcal{PC}^{k-1} = \bar{PC}^{k-1}$ and $N^i = \bar{N}^i$ because $\bar{N}^i \subseteq \hat{N}^i$ and $\bar{PC}^{k-1} \supseteq \hat{\mathcal{PC}}^{k-1}$ by the inductive hypothesis. Hence, s^* satisfies the requirements for $\bar{PC}^k$. ■

6.2 Other Applications

Compared to the earlier literature, Backwards Rationalizability provides the first well-defined notion of backward induction reasoning in incomplete information settings. An early application with incomplete information is provided by Penta (2015), who studies the problem of robust implementation in dynamic settings. In that context, Backwards Rationalizability enables two main achievements. First, it extends the robust implementation approach of Bergemann and

Morris (2009) to dynamic environments, in which agents may obtain information over time. In principle, doing so would require studying whether a mechanisms exists for which, for all possible models of beliefs over the stochastic process that generates players types over periods, and for all the perfect Bayesian equilibria associated with each such model, agents behave so as to induce outcomes consistent with the designer's objective. A direct approach to the question would thus require the solution of a continuum of complicated fixed-points problems, each of difficult solution, since even for a single model of beliefs characterizing the set of PBE can be very challenging. Resorting to Backwards Rationalizability instead makes it possible to pursue a much more tractable approach, that enables a seamless extension of Bergemann and Morris (2009) static analysis to dynamic environments. Second, the analysis in Penta (2015) also sheds light on Bergemann and Morris (2007) results on the advantages of using dynamic mechanisms in static environments. In particular, Penta (2015) results show that the fundamental insight that robustness may be favored by the reduction of strategic uncertainty that backwards induction grants in a dynamic mechanism under complete and perfect information, does not survive the introduction of incomplete information.

Importantly, however, we would like to stress that the advantages of Backwards Rationalizability can be seen even without the extra complexity associated with incomplete information. That is, attaining a precise understanding of backward induction reasoning, independent of other equilibrium restrictions, may prove useful in applications *even* under complete information. A notable case in point – which is important both economically and historically – is provided by the original two-period Hotelling model of horizontal differentiation. Backward induction is a natural way of reasoning in this game: before considering the possible positioning, a firm wants to understand which prices could emerge in the second stage depending on the locations chosen in the first stage. Yet, in the baseline specification with linear transportation cost (Hotelling (1929)), subgame perfect equilibrium fails to provide a tractable and intuitive solution to the location problem. A numerical solution was found by Osborne and Pitchik (1987), whereby the chosen locations induce a complicated mixed pricing equilibrium where firms may engage in a price war, whereas slightly higher differentiation would induce certain prices and overall higher profits. Attempts to recover some tractability have explored alternative cost functions, but have produced insights that often clash with basic economic intuition.[19] As a consequence, the literature on the two-period Hotelling model has pretty much died, ending up considering it as a sort of puzzle, despite the inherent plausibility of the baseline model.

In an ongoing project (Catonini and Penta (2022)), we show that the tools developed in this paper may be fruitfully applied to think about the two-period Hotelling model afresh. That is, as a way to still maintain the fundamental logic of backward induction reasoning – which is inherently compelling in this setting – without the entanglement with other kinds of assumptions that are implicitly in the SPE notion. In particular, we append Backwards Rationalizability with a simple refinement, that formalizes the idea that firms know the *path of*

[19] For instance, d'Aspremont et al. (1979) explored the variation of the model with quadratic transport costs, and showed the existence of an easy-to-compute SPE. Such an equilibrium, however, induces maximal differentiation (the two firms position themselves at the opposite extremes of the Hotelling interval), a result which is considered at odds with factual observation.

play but face strategic uncertainty after a deviation, plus an additional hypothesis of "closedness under rational behavior" (Basu and Weibull (1991)) that is motivated by a requirement of self-enforceability along the path. We show that the resulting concept rules out coordination on specific randomizations, since they are not self-enforceable, and we identify the transportation-efficient location pair as the *only* location pair that is consistent with this solution concept.[20]

Overall, in our view the applications above show that Backwards Rationalizability not only is a tractable and ready-to-use solution concept, but it may also serve as basis to impose extra desiderata (dictated by the specific economic context) over the simple and compelling logic of backward induction, separate from other kinds of assumptions that are entangled with it in existing solution concepts, and which do not always prove tractable or plausible.

7 Conclusion

The idea of backward induction reasoning is informally associated with several solution concepts for dynamic games with incomplete information. Yet, we lack a precise understanding of what backward induction reasoning means in these context, to what extent it can be separated from other kinds of assumptions, and to what extent familiar ideas that are intuitively associated with backward induction reasoning can be reconciled with incomplete information.

This paper covers this gap by introducing a new solution concept, Backwards Rationalizability, that captures precisely the behavioral implications of backward induction reasoning in games with imperfect and incomplete information, without extraneous restrictions on players' beliefs or equilibrium assumptions. Our results show that Backwards Rationalizability satisfies several desiderata that are more or less directly associated with backward induction reasoning. Namely: (i) it satisfies *continuation-game consistency*, which provides a natural incomplete-information extension of the recursive structure of subgame perfect equilibrium; (ii) in finite horizon games, it can be computed with a 'backwards procedure', that starts from the end of the game and proceeds backwards, considering each continuation-game in isolation (very much like one can do, under complete information, for backward induction or subgame perfection); (iii) third, the solution concept directly embodies the view of unexpected moves as possible mistakes; (iv) fourth, it characterizes the set of equilibrium strategies that may be played, across all type spaces that could be defined over the underlying space of payoff uncertainty, for an equilibrium concept that introduces backward induction logic *and nothing more* into equilibrium analysis.

Besides these results, which jointly provide a unified perspective on the meaning of backward induction reasoning in games with incomplete and imperfect information, we stress that as a solution concept for belief-free (dynamic) games Backwards Rationalizability does not impose the equilibrium assumption that players have correct beliefs about others' behavior. This allows to make predictions based on a backward induction logic, even in contexts in which the equilibrium assumptions are difficult to operationalise, or in which they are not necessarily compelling.

[20] An earlier paper by Catonini (2021) also analyzed an example of the Hotelling model with linear costs, but simplified to a discrete space of locations, and adopting a notion of self-enforcing agreements under *forward induction* reasoning. The results in Catonini and Penta (2022) therefore strengthen the earlier ones in that they are based on a much weaker solution concept, which is both easier to apply and closer to more standard notions in the applied literature, and for being based on the original (non discretized) model with a continuum of locations.

Backwards Rationalizability may thus prove fruitful in a number of applications, even in settings with complete information (which of course are a special case of those considered in this paper).

A notable case in point, which we discussed in Section 6, is provided by the two-period Hotelling model of horizontal differentiation: the basic logic of backward induction is inherently compelling in that model, and yet standard subgame perfect equilibrium analysis has proven intractable and has produced results that are at odds with basic economic intuition. In Catonini and Penta (2022) we show that the plain logic of backward induction that we have distilled in this paper, appended with a basic requirement of self-enforceability on path, yields a unique solution in the Hotelling model that is both easy to compute and consistent with sound economic intuition. Another example is the version of *peer-confirming equilibrium* that we developed in Section 6, modifying the original concept of Lipnowski and Sadler (2019) so as to accommodate backward induction reasoning. As we showed, besides making it easier to apply the idea of peer-confirming to dynamic settings, the adoption of Backwards Rationalizability in this context restores the natural comparative statics results that the original concept only features in static games. Since most equilibrium refinements are based on backward induction reasoning, the seamless integration of equilibrium and strategic reasoning achieved by Backwards Rationalizability for peer-confirming equilibrium is likely to carry over to other models of partial coordination that could be further explored.

In our view, the extensions we discussed here and in Section 6 show that distilling the precise implications of backward induction reasoning, separate from everything else (such as the equilibrium hypothesis, belief persistence, and other kinds of assumptions), may prove useful in applications both to increase the tractability of the analysis, and to restore natural economic intuition, in settings with complete and incomplete information alike.

Appendix

A Proofs

We first introduce some additional terminology that will be used in the proofs.

Fix an elimination procedure of type-strategy pairs $((\hat{\Omega}_i^{h,k})_{i\in N})_{k\geq 0}$ for the continuation game with root h. We say that a CPS $\mu^{i,h}$ over $\Theta_0 \times \Theta_{-i} \times S_{-i}^h$ is *viable* for $\hat{\Omega}_i^{h,k}$ when, for every $h' \succeq h$ and (θ_{-i}, s_{-i}^h) such that $\mu^{i,h}(\Theta_0 \times \{(\theta_{-i}, s_{-i}^h)\}|h') > 0$, there is $(\tilde{\theta}_{-i}, \tilde{s}_{-i}^h) \in \hat{\Omega}_{-i}^{h,k-1}$ such that $\tilde{s}_{-i}^h|h' = s_{-i}^h|h'$ and $\tilde{\theta}_{-i} = \theta_{-i}$. We say that $\mu^{i,h}$ "*justifies* $(\theta_i, s_i^h) \in \hat{\Omega}_i^{h,k}$" when $\mu^{i,h}$ is viable for $\hat{\Omega}_i^{h,k}$ and s_i^h is a sequential best reply to $\mu^{i,h}$ for θ_i.

Proof of Theorem 2.

The statement is an identity for $h = h^0$, so suppose $h \neq h^0$.

Trivially, $\mathcal{BR}^0|h = \mathcal{BR}^{h,0}$. Now fix $k > 0$ and suppose by induction that $\mathcal{BR}^{k-1}|h = \mathcal{BR}^{h,k-1}$.

First we show $\mathcal{BR}^k|h \subseteq \mathcal{BR}^{h,k}$. Fix $i \in N$, $(\theta_i, s_i) \in \mathcal{BR}_i^k$, and a CPS μ^i that justifies this. Define the map

$$\varsigma : (\theta_0, \theta_{-i}, s_{-i}) \mapsto (\theta_0, \theta_{-i}, s_{-i}|h)\,.$$

Construct a CPS $\mu^{i,h} = \left(\mu^{i,h}(\cdot|h')\right)_{h' \succeq h}$ over $\Theta_0 \times \Theta_{-i} \times S_{-i}^h$ as follows: for each $h' \succeq h$, let $\mu^{i,h}(\cdot|h')$ be the pushforward of $\mu^i(\cdot|h')$ through ς. By the induction hypothesis, $\mathcal{BR}_{-i}^{k-1}|h' = \mathcal{BR}_{-i}^{h,k-1}|h'$, so the fact that μ^i is viable for $\mathcal{BR}_i^k$ implies that $\mu^{i,h}$ is viable for $\mathcal{BR}_i^{h,k}$. Moreover, $\mu^i(\cdot|h')$ and $\mu^{i,h}(\cdot|h')$ induce the same distribution over types and continuation strategies, therefore the fact that s_i is a continuation best reply to $\mu^i(\cdot|h')$ for θ_i implies that so is $s_i|h$ to $\mu^{i,h}(\cdot|h')$. Thus, $(\theta_i, s_i|h) \in \mathcal{BR}_i^{h,k}$.

Now we show $\mathcal{BR}^k|h \supseteq \mathcal{BR}^{h,k}$. Fix $i \in N$. Let $\overline{h} \preceq h$ be the shortest history such that $s_{-i}|\overline{h} \in S_{-i}^{\overline{h}}(h)$ for all $(\theta_{-i}, s_{-i}) \in \mathcal{BR}_{-i}^{k-1}$. Thus, if $\overline{h} \neq h^0$, there exists $(\overline{\theta}_{-i}, \overline{s}_{-i}) \in \mathcal{BR}_{-i}^{k-1}$ such that $\overline{s}_{-i}|p(\overline{h}) \notin S_{-i}^{p(\overline{h})}(h)$, but since $\overline{s}_{-i}|\overline{h} \in S_{-i}^{\overline{h}}(h)$, we must have $\overline{s}_{-i}|p(\overline{h}) \notin S_{-i}^{p(\overline{h})}(\overline{h})$. Let $\bar{\mu}^i$ be a viable CPS for $\mathcal{BR}_i^k$ such that, if $\overline{h} \neq h^0$, at every history $h' \prec \overline{h}$, player i assigns probability 1 to $(\overline{\theta}_{-i}, \overline{s}_{-i}|h')$,[21] so that $\bar{\mu}^i(\Theta_0 \times \Theta_{-i} \times S_{-i}(\overline{h})|p(\overline{h})) = 0$. Fix a map ς that associates each $(\bar{\theta}_0, \bar{\theta}_{-i}, s_{-i}^h) \in \Theta_0 \times \Theta_{-i} \times S_{-i}^h$ with some[22] $(\bar{\theta}_0, \bar{\theta}_{-i}, s_{-i}) \in \Theta_0 \times \Theta_{-i} \times S_{-i}(h)$ such that (a) $s_{-i}|h = s_{-i}^h$ and (b) if $\left(\bar{\theta}_{-i}, s_{-i}^h\right) \in \mathcal{BR}_{-i}^{h,k-1}$, then $(\bar{\theta}_{-i}, s_{-i}|\overline{h}) \in \mathcal{BR}_{-i}^{k-1}|\overline{h}$ — requirement (b) is compatible with (a) and with $s_{-i} \in S_{-i}(h)$ because, by the induction hypothesis, $s_{-i}^h = \hat{s}_{-i}|h$ for some $(\bar{\theta}_{-i}, \hat{s}_{-i}) \in \mathcal{BR}_{-i}^{k-1}$, and by definition of $\overline{h}$, $\hat{s}_{-i}|\overline{h} \in S_{-i}^{\overline{h}}(h)$, so one can choose any $s_{-i} \in S_{-i}(\overline{h})$ such that $s_{-i}|\overline{h} = \hat{s}_{-i}|h$.

Now fix $(\theta_i, s_i^h) \in \mathcal{BR}_i^{h,k}$ and a CPS $\mu^{i,h}$ that justifies this. Construct an array of conditional beliefs $\mu^i = (\mu^i(\cdot|h))_{h \in \mathcal{H}}$ as follows:

1. for each $h' \succeq h$, let $\mu^i(\cdot|h')$ be the pushforward of $\mu^{i,h}(\cdot|h')$ through ς;
2. for each $h' \succeq \overline{h}$ with $h' \prec h$, let $\mu^i(\cdot|h') = \mu^i(\cdot|h)$;
3. for every other $h' \succ \overline{h}$ with $\mu^i(\Theta_0 \times \Theta_{-i} \times S_{-i}(h')|\overline{h}) > 0$, derive $\mu^i(\cdot|h')$ from $\mu^i(\cdot|\overline{h})$ by conditioning;
4. for every other h', let $\mu^i(\cdot|h') = \bar{\mu}^i(\cdot|h')$.

It is easy to check that μ^i is a CPS. For each $h' \succeq h$ and (θ_{-i}, s_{-i}) with $\mu^i(\Theta_0 \times \{(\theta_{-i}, s_{-i})\}\,|h') > 0$, by 1. and (a) we have $\mu^{i,h}(\Theta_0 \times \{(\theta_{-i}, s_{-i}|h)\}\,|h') > 0$, so by the fact that $\mu^{i,h}$ is viable for $\mathcal{BR}_i^{h,k}$, we get $(\theta_{-i}, s_{-i}|h') \in \mathcal{BR}_{-i}^{h,k-1}|h'$, and hence by the induction hypothesis $(\theta_{-i}, s_{-i}|h') \in \mathcal{BR}_{-i}^{k-1}|h'$. For each $h' \succeq \overline{h}$ with $h' \prec h$ and (θ_{-i}, s_{-i}) with $\mu^i(\Theta_0 \times \{(\theta_{-i}, s_{-i})\}\,|h') > 0$, by 2. we have $\mu^i(\Theta_0 \times \{(\theta_{-i}, s_{-i})\}\,|h) > 0$, so as just argued $(\theta_{-i}, s_{-i}|h) \in \mathcal{BR}_{-i}^{h,k-1}$, and hence by 1. and (b) we get $\left(\theta_{-i}, s_{-i}|\overline{h}\right) \in \mathcal{BR}_{-i}^{k-1}|\overline{h}$, which implies $(\theta_{-i}, s_{-i}|h') \in \mathcal{BR}_{-i}^{k-1}|h'$. Therefore, together with 3. and 4., we can conclude that μ^i is viable for $\mathcal{BR}_i^k$. Finally, for each $h' \succeq h$, by 1. and (a), $\mu^i(\cdot|h')$ and $\mu^{i,h}(\cdot|h')$ induce the same distribution over types and continuation strategies, therefore μ^i justifies $(\theta_i, s_i) \in \mathcal{BR}_i^k$ for some s_i with $s_i|h = s_i^h$. ■

[21] Since $\bar{\mu}^i(\cdot|h')$ is a probability measure over $\Theta_0 \times \Theta_{-i} \times S_{-i}$, not $\Theta_0 \times \Theta_{-i} \times S_{-i}^{h'}$, we refer of course to the belief induced over the continuation strategies.

[22] Note that we keep fixed $(\bar{\theta}_0.\bar{\theta}_{-i})$. Throughout all the proofs, we will always do so in the construction of maps.

For the proof of Theorem 3, we will refer to the following definition of $\mathcal{BP}$, which includes the steps of belief-free rationalizability. For each $h \in \mathcal{H}$, let $\phi(h)$ denote the set of immediate successors of h in $\mathcal{H}$ (if any).

Definition 9. *Fix $h \in \mathcal{H}$ and suppose that, for each $h' \succ h$ (if any) $\mathcal{BP}^{h'}$ has already been defined.*

Step 0: *For each $i \in N$, let*

$$\mathcal{BP}_i^{h,0} = \left\{(\theta_i, s_i^h) \in \Theta_i \times S_i^h : \forall h' \in \phi(h), (\theta_i, s_i^h|h') \in \mathcal{BP}_i^{h'}\right\}.$$

(if h is preterminal, $\phi(h) = \emptyset$, thus $\mathcal{BP}_i^{h,0} = \Theta_i \times S_i^h$).

Step k: *For each $i \in N$ and $(\theta_i, s_i^h) \in \Theta_i \times S_i^h$, let $(\theta_i, s_i^h) \in \mathcal{BP}_i^{h,k}$ if there exists $\nu_i^h \in \Delta(\Theta_0 \times \Theta_{-i} \times S_{-i}^h)$ such that:*

$BP1^h$: $s_i^h \in \hat{r}_i^h(\nu_i^h; \theta_i)$.

$BP2^h$: $\nu_i^h(\Theta_0 \times \mathcal{BP}_{-i}^{h,k-1}) = 1$.

For each $i \in N$, let $\mathcal{BP}_i^h = \cap_{k>0}\mathcal{BP}_i^{h,k}$.

Proof of Theorem 3.

We are going to write $s_i^h \simeq \widetilde{s}_i^h$ when $\left[s_i^h\right] = \left[\widetilde{s}_i^h\right]$ (i.e., s_i^h and $\widetilde{s}_i^h$ belong to the same realization-equivalent class).

By Theorem 2, $\mathcal{BR}|h = \mathcal{BR}^h$, so we can prove $\left[\mathcal{BR}_i^h\right] = \left[\mathcal{BP}_i^h\right]$ for all $i \in N$. The proof is recursive on the length of histories, starting from preterminal histories and moving backwards. So, suppose that the result holds for every history longer than history h.

Define an elimination procedure $((\hat{\Omega}_i^{h,k})_{i\in N})_{k\geq 0}$ as follows. For each $i \in N$, let $\hat{\Omega}_i^{h,0} = \Theta_i \times S_i^h$. For each $k > 0$, let $\left(\theta_i, s_i^h\right) \in \hat{\Omega}_i^{h,k}$ if there exists a viable CPS μ_i^h for $\hat{\Omega}_i^{h,k}$ such that, for each $h' \succ h$, s_i^h is a continuation best reply to $\mu_i^h(\cdot|h')$ for θ_i, even if not at h. Let K be the first step k such that $\hat{\Omega}^{h,k} = \hat{\Omega}^{h,k+1}$. Now define an elimination order of the backwards rationalizability operator in the continuation game with root h, denoted by $\mathcal{B}\hat{\mathcal{R}}^h$, as follows. For each $k = 0, ..., K$, let $\mathcal{B}\hat{\mathcal{R}}^{h,k} = \hat{\Omega}^{h,k}$. For each $k > K$ and $i \in N$, let $\left(\theta_i, s_i^h\right) \in \mathcal{B}\hat{\mathcal{R}}_i^{h,k}$ if there exists a CPS μ_i^h that justifies this. By Theorem 1, $\mathcal{B}\hat{\mathcal{R}}^h = \mathcal{BR}^h$, so we can prove $[\mathcal{B}\hat{\mathcal{R}}_i^h] = \left[\mathcal{BP}_i^h\right]$.

It is easy to see that, for every $i \in N$, $(\theta_i, s_i^h) \in \mathcal{B}\hat{\mathcal{R}}_i^{h,K}$ if and only if $(\theta_i, s_i^h|h') \in \mathcal{BR}_i^{h'}$ for all $h' \in \phi(h)$. (Thus, $\mathcal{B}\hat{\mathcal{R}}^{h,K}|h' = \mathcal{BR}^{h'} = \mathcal{BR}^h|h' = \mathcal{B}\hat{\mathcal{R}}^h|h'$, where the second equality is by Theorem 2 and the last equality by Theorem 1.) By definition, $(\theta_i, s_i^h) \in \mathcal{BP}_i^{h,0}$ if and only if $(\theta_i, s_i^h|h') \in \mathcal{BP}_i^{h'}$ for all $h' \in \phi(h)$. By the recursive hypothesis, $[\mathcal{BR}_i^{h'}] = [\mathcal{BP}_i^{h'}]$. Hence, $[\mathcal{B}\hat{\mathcal{R}}_i^{h,K}] = [\mathcal{BP}_i^{h,0}]$.

Now fix $k > 0$ and assume by way of induction that $[\mathcal{B}\hat{\mathcal{R}}_i^{h,K+k-1}] = [\mathcal{BP}_i^{h,k-1}]$ for all $i \in N$.

Fix $(\theta_i, s_i^h) \in \mathcal{B}\hat{\mathcal{R}}_i^{h,K+k}$ and $\mu^{i,h}$ that justifies this. By the induction hypothesis, we can fix a map ς that associates each $(\bar{\theta}_0, (\bar{\theta}_j, s_j^h)_{j\neq i}) \in \Theta_0 \times \mathcal{B}\hat{\mathcal{R}}_{-i}^{h,K+k-1}$ with some $(\bar{\theta}_0, (\bar{\theta}_j, \widetilde{s}_j^h)_{j\neq i}) \in \Theta_0 \times \mathcal{BP}_{-i}^{h,k-1}$ such that $s_j^h \simeq \widetilde{s}_j^h$ for every $j \neq i$. Let ν_i^h be the pushforward of $\mu^{i,h}(\cdot|h)$ through

ς; it satisfies BP2^h. Since s_i^h is a continuation best reply to $\mu^{i,h}(\cdot|h)$, it satisfies BP1^h with ν_i^h, so $(\theta_i, s_i^h) \in \mathcal{BP}_i^{h,k}$.

Fix $(\theta_i, s_i^h) \in \mathcal{BP}_i^{h,k}$ and ν_i^h that satisfies BP1^h and BP2^h at step k. By the induction hypothesis, there exists $\hat{s}_i^h \simeq s_i^h$ such that $(\theta_i, \hat{s}_i^h) \in \mathcal{B}\hat{\mathcal{R}}_i^{h,K}$. Thus, there exists a CPS $\hat{\mu}^{i,h}$ such that, for each $h' \succ h$, $\hat{s}_i^h$ is a continuation best reply to $\hat{\mu}^{i,h}(\cdot|h')$ for θ_i, and for each (θ_{-i}, s_{-i}^h) with $\hat{\mu}^{i,h}(\Theta_0 \times \{(\theta_{-i}, s_{-i}^h)\}|h') > 0$, $(\theta_{-i}, s_{-i}^h|h') \in \mathcal{B}\hat{\mathcal{R}}_{-i}^{h,K}|h' = \mathcal{B}\hat{\mathcal{R}}_{-i}^{h}|h'$, where the equality is given by the argument in brackets above. By the induction hypothesis, we can fix a map ς that associates each $(\bar{\theta}_0, (\bar{\theta}_j, s_j^h)_{j\neq i}) \in \Theta_0 \times \mathcal{BP}_{-i}^{h,k-1}$ with some $(\bar{\theta}_0, (\bar{\theta}_j, \tilde{s}_j^h)_{j\neq i}) \in \Theta_0 \times \mathcal{B}\hat{\mathcal{R}}_{-i}^{h,K+k-1}$ such that $\tilde{s}_j^h \simeq s_j^h$ for every $j \neq i$. Construct $\mu^{i,h}$ as follows: let $\mu^{i,h}(\cdot|h)$ be the pushforward of ν_i^h through ς, and for each $h' \succ h$, derive $\mu^{i,h}(\cdot|h')$ from $\mu^{i,h}(\cdot|h)$ by conditioning if possible, otherwise let $\mu^{i,h}(\cdot|h') = \hat{\mu}^{i,h}(\cdot|h')$. It is easy to see that $\mu^{i,h}$ justifies $(\theta_i, \hat{s}_i^h) \in \mathcal{B}\hat{\mathcal{R}}_i^{h,K+k}$. ■

Proof of Theorem 4.

"If" part. Every interim strategy s_i that is played with positive probability in the IPE (b, p) by some type t_i is sequentially rational for $\vartheta_i(t_i)$ given the CPS $\mu_{(b,p)}^{t_i}$. At every history h, $\mu_{(b,p)}^{t_i}$ assigns positive probability only to pairs (θ_j, s_j') where $s_j'|h = s_j|h$ for some s_j that is played with positive probability in the IPE by some type $t_j \in \vartheta_j^{-1}(\theta_j)$. Hence, a simple inductive argument shows that all type-interim strategy pairs induced by b survive all steps of $\mathcal{BR}$.

"Only if" part. Construct a type structure as follows. For each $i \in N$, let $T_i = \mathcal{BR}_i$, and for each $t_i = (\theta_i, s_i) \in T_i$, let $\vartheta_i(t_i) = \theta_i$. Now fix $t_i = (\theta_i, s_i) \in T_i$. By the fixed-point property of $\mathcal{BR}$, we can fix μ^{t_i} such that (i) $s_i \in r_i(\mu^{t_i}; \theta_i)$ and (ii) for each $h \in \mathcal{H}$, there is a map $\xi_h^{t_i}$ that associates each $(\bar{\theta}_0, (\bar{\theta}_j)_{j\neq i}, (s_j)_{j\neq i}) \in \text{Supp}\mu^{t_i}(\cdot|h)$ with some $(\bar{\theta}_0, (\bar{\theta}_j, s_j')_{j\neq i}) \in \Theta_0 \times T_{-i}$ such that $s_{-i}'|h = s_{-i}|h$. Let $\tau_i(\cdot|t_i)$ be the pushforward of $\mu^{t_i}(\cdot|h^0)$ through $\xi_h^{t_i}$.

Now construct the desired IPE as follows. For each $i \in N$, define the strategy b_i as $b_i(t_i) = s_i$ for each $t_i = (\theta_i, s_i)$. For each $t_i \in T_i$, define $p_i(\cdot|t_i)$ recursively as follows. First, let $p_i(\cdot|h^0; t_i) = \tau_i(\cdot|t_i)$. So, p satisfies condition 1 of weak preconsistency. From this, derive $\hat{\mu}_{(b,p)}^{t_i}(\cdot|h^0)$ with equation 2. Now fix $h \succ h^0$ and suppose that $\hat{\mu}_{(b,p)}^{t_i}(\cdot|p(h))$ was defined. If $\hat{\mu}_{(b,p)}^{t_i}(\Theta_0 \times T_{-i} \times S_{-i}(h)|p(h)) > 0$, derive $\hat{\mu}_{(b,p)}^{t_i}(\cdot|h)$ with equation 3 and let $p_i(\cdot|h; t_i)$ be its marginal on $\Theta_0 \times T_{-i}$; otherwise, let $p_i(\cdot|h; t_i)$ be the pushforward of $\mu^{t_i}(\cdot|h)$ through $\xi_h^{t_i}$ and derive $\hat{\mu}_{(b,p)}^{t_i}(\cdot|h)$ with equation 4; either way, $p_i(\cdot|h; t_i)$ satisfies condition 2 of pre-consistency. Thus, to prove that (b, p) is an IPE, there only remains to show the optimality of b.

Fix $i \in N$ and $t_i = (\theta_i, s_i)$. Fix $h \in \mathcal{H}$ such that $h = h^0$ or $\hat{\mu}_{(b,p)}^{t_i}(\Theta_0 \times T_{-i} \times S_{-i}(h)|p(h)) = 0$. Then, for each $\omega = (\theta_0, (t_j)_{j\neq i}, (s_j)_{j\neq i})$, we have

$$\hat{\mu}_{(b,p)}^{t_i}(\omega|h) \overset{\text{(Eqs 2,4)}}{=} p_i(\theta_0, t_{-i}|h; t_i) \cdot \textstyle\prod_{j\neq i} b_j(\varrho_{j,h}^{-1}(s_j)|t_j)$$
$$\overset{\text{(def. of } p\text{)}}{=} \mu_i^{t_i}((\xi_h^{t_i})^{-1}(\theta_0, t_{-i})\,|h) \cdot \textstyle\prod_{j\neq i} b_j(\varrho_{j,h}^{-1}(s_j)|t_j)$$
$$\overset{\text{(def. of } b\text{)}}{=} \begin{cases} \mu_i^{t_i}((\xi_h^{t_i})^{-1}(\theta_0, t_{-i})\,|h) & \text{if } s_j = \varrho_{j,h}(b_j(t_j)) \quad (\forall j \neq i), \\ 0 & \text{otherwise} \end{cases}.$$

Now, if $s_j = \varrho_{j,h}(b_j(t_j))$ for every $j \neq i$, for each $\omega' = (\theta_0, (\theta_j)_{j\neq i}, (s_j')_{j\neq i}) \in \left(\xi_h^{t_i}\right)^{-1}(\omega)$, we have $\theta_j = \vartheta_j(t_j)$ and $s_j'|h = s_j|h$ for every $j \neq i$. Hence, $\mu_{(b,p)}^{t_i}(\cdot|h)$ (which is the pushforward of

$\hat{\mu}^{t_i}_{(b,p)}(\cdot|h)$ through $\times_{j\neq i}\vartheta_j$) and $\mu^{t_i}_i(\cdot|h)$ induce the same distribution over types and continuation strategies, therefore s_i is a continuation best reply also to $\mu^{t_i}_{(b,p)}(\cdot|h)$ for θ_i. At every other $h \in \mathcal{H}$, s_i is a continuation best reply as well because $\mu^{t_i}_{(b,p)}$ satisfies the chain rule. ■

References

Aumann, R. J. (1974), 'Subjectivity and correlation in randomized strategies', *Journal of mathematical Economics* **1**(1), 67–96.

Basu, K. and Weibull, J. W. (1991), 'Strategy subsets closed under rational behavior', *Economics Letters* **36**(2), 141–146.

Battigalli, P. (1996), 'Strategic rationality orderings and the best rationalization principle', *Games and Economic Behavior* **13**(2), 178–200.

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

Battigalli, P., Catonini, E. and Manili, J. (2021), Belief change, rationality, and strategic reasoning in sequential games, Working Paper 679, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Battigalli, P. and De Vito, N. (2021), 'Beliefs, plans, and perceived intentions in dynamic games', *Journal of Economic Theory* pp. 1–43.

Battigalli, P. and Siniscalchi, M. (2003*a*), 'Rationalizability and incomplete information', *Advances in Theoretical Economics* **3**(1), article 3.

Battigalli, P. and Siniscalchi, M. (2003*b*), 'Rationalizable bidding in first-price auctions', *Games and Economic Behavior* **45**(1), 38–72.

Battigalli, P. and Siniscalchi, M. (2007), 'Interactive epistemology in games with payoff uncertainty', *Research in Economics* **61**, 165–184.

Battigalli, P., Tillio, A. D., Grillo, E. and Penta, A. (2011), 'Interactive epistemology and solution concepts for games with asymmetric information', *The B.E. Journal of Theoretical Economics* **11**(1).

Ben-Porath, E. (1993), 'Repeated games with finite automata', *Journal of Economic Theory* **59**(1), 17–32.

Bergemann, D., Brooks, B. and Morris, S. (2015), 'The limits of price discrimination', *American Economic Review* **105**(3), 921–57.

Bergemann, D., Brooks, B. and Morris, S. (2017), 'First-price auctions with general information structures: Implications for bidding and revenue', *Econometrica* **85**(1), 107–143.

Bergemann, D. and Morris, S. (2005), 'Robust mechanism design', *Econometrica* pp. 1771–1813.

Bergemann, D. and Morris, S. (2007), 'An ascending auction for interdependent values: Uniqueness and robustness to strategic uncertainty', *American Economic Review* **97**(2), 125–130.

Bergemann, D. and Morris, S. (2009), 'Robust implementation in direct mechanisms', *The Review of Economic Studies* **76**(4), 1175–1204.

Bergemann, D. and Morris, S. (2013), 'Robust predictions in games with incomplete information', *Econometrica* **81**(4), 1251–1308.

Bergemann, D. and Morris, S. (2016), *Bayes correlated equilibrium andthe comparison of information structures in games*, Vol. 11.

Bernheim, B. D. (1984), 'Rationalizable strategic behavior', *Econometrica: Journal of the Econometric Society* pp. 1007–1028.

Brandenburger, A. and Dekel, E. (1987), 'Rationalizability and correlated equilibria', *Econometrica: Journal of the Econometric Society* pp. 1391–1402.

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

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

Catonini, E. (2021), 'Self-enforcing agreements and forward induction reasoning', *The Review of Economic Studies* **2**, 610–642.

Catonini, E. and Penta, A. (2022), 'A simple solution to the hotelling problem', *mimeo* .

d'Aspremont, C., Gabszewicz, J. J. and Thisse, J.-F. (1979), 'On hotelling's "stability in competition"', *Econometrica* **47**(5), 1145–1150.

Dekel, E., Fudenberg, D. and Morris, S. (2007), 'Interim correlated rationalizability', *Theoretical Economics* .

Doval, L. and Ely, J. C. (2020), 'Sequential information design', *Econometrica* **88**(6), 2575–2608.

Ely, J. C. and Peski, M. (2006), 'Hierarchies of belief and interim rationalizability', *Theoretical Economics* **1**(1), 19–65.

Fudenberg, D., Dekel, E. and Morris, S. (2006), 'Topologies on types', *Theoretical Economics* **1**.

Fudenberg, D. and Tirole, J. (1991*a*), *Game theory*, MIT press.

Fudenberg, D. and Tirole, J. (1991*b*), 'Perfect bayesian equilibrium and sequential equilibrium', *journal of Economic Theory* **53**(2), 236–260.

Harsanyi, J. C. (1967), 'Games with incomplete information played by "bayesian" players, i–iii part i. the basic model', *Management science* **14**(3), 159–182.

Harsanyi, J. C. and Selten, R. (1988), 'A general theory of equilibrium selection in games', *MIT Press Books* **1**.

Hotelling, H. (1929), 'Stability in competition', *The Economic Journal* **39**(153), 41–57.

Kreps, D. M. and Wilson, R. (1982), 'Sequential equilibria', *Econometrica: Journal of the Econometric Society* pp. 863–894.

Lipnowski, E. and Sadler, E. (2019), 'Peer-confirming equilibrium', *Econometrica* **87**(2), 567–591.

Magnolfi, L. and Roncoroni, C. (2020), Estimation of Discrete Games with Weak Assumptions on Information, The Warwick Economics Research Paper Series (TWERPS) 1247, University of Warwick, Department of Economics.

Makris, M. and Renou, L. (2018), Information design in multi-stage games, Working Papers 861, Queen Mary University of London, School of Economics and Finance.

Mas-Colell, A., Whinston, M. D., Green, J. R. et al. (1995), *Microeconomic theory*, Vol. 1, Oxford university press New York.

Ollár, M. and Penta, A. (2017), 'Full implementation and belief restrictions', *American Economic Review* **107**(8), 2243–77.

Ollár, M. and Penta, A. (2021), 'A network solution to robust implementation: The case of identical but unknown distributions', *BSE working paper series, w.p. 1248* .

Osborne, M. J. and Pitchik, C. (1987), 'Cartels, profits and excess capacity', *International Economic Review* pp. 413–428.

Oury, M. and Tercieux, O. (2012), 'Continuous implementation', *Econometrica* **80**(4), 1605–1637.

Pearce, D. G. (1984), 'Rationalizable strategic behavior and the problem of perfection', *Econometrica: Journal of the Econometric Society* pp. 1029–1050.

Penta, A. (2010), 'Incomplete information and robustness in strategic environments', *Penn Dissertation* (131).

Penta, A. (2012*a*), 'Backward induction reasoning in games with incomplete information', *mimeo* .

Penta, A. (2012*b*), 'Higher order uncertainty and information: Static and dynamic games', *Econometrica* **80**(2), 631–660.

Penta, A. (2013), 'On the structure of rationalizability for arbitrary spaces of uncertainty', *Theoretical Economics* **8**(2), 405–430.

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

Penta, A. and Zuazo-Garin, P. (2021), 'Rationalizability, observability and common knowledge', *The Review of Economic Studies* **forthcoming**.

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

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.

Rényi, A. (1955), 'On a new axiomatic theory of probability', *Acta Mathematica Academiae Scientiarum Hungarica* **6**, 285–335.

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

Watson, J. (2017), 'A general, practicable definition of perfect bayesian equilibrium', *unpublished draft* .

Weinstein, J. and Yildiz, M. (2007), 'A structure theorem for rationalizability with application to robust predictions of refinements', *Econometrica* **75**(2), 365–400.

Weinstein, J. and Yildiz, M. (2011), 'Sensitivity of equilibrium behavior to higher-order beliefs in nice games', *Games and Economic Behavior* **72**(1), 288–300.

Zuazo-Garin, P. (2017), 'Uncertain information structures and backward induction', *Journal of Mathematical Economics* **71**, 135–151.