Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1894

Backward induction reasoning beyond backward induction

Emiliano Cantonini and Antonio Penta

September 2024

Backward Induction Reasoning beyond Backward Induction*

Emiliano Catonini† Antonio Penta‡

NYU-Shanghai ICREA, UPF, BSE and TSE

September 18, 2024

Abstract

Backward Induction is only defined for games with perfect information, but its logic is also invoked in many equilibrium concepts for games with imperfect or incomplete information. Yet, the meaning of 'backward induction reasoning' is unclear in these settings, and we lack a way to apply its simple logic to general games. We remedy this by introducing a solution concept, Backwards Rationalizability, that satisfies several properties normally ascribed to backward induction reasoning, foremost the possibility of being computed via a tractable backwards procedure. We also show that Backwards Rationalizability characterizes the robust predictions of a 'perfect equilibrium' notion that introduces the backward induction logic and nothing more into equilibrium analysis. We

*Earlier versions of some of our results circulated under the title "Backward Induction Reasoning in Incomplete Information Games", by Penta (2012). 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 Larbi Alaoui, Pierpaolo Battigalli, George Mailath, Andres Perea, Joel Watson and Bill Sandholm. We thank Andrea Salvanti for the RA work. The BSE acknowledges the financial support of the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S). Antonio Penta acknowledges the financial support of the European Research Council, ERC Starting Grant #759424.

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

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

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

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

JEL codes: C72, C73, D82.

1 Introduction

Backward Induction (BI) is one of the fundamental notions of game theory. Strictly speaking, the BI algorithm is only defined for games with perfect and complete information 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 BI to games with imperfect information. or with payoff ties. 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 typically cannot be 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.

In pursuit of an answer, it seems reasonable to inspect the main solution concepts that are traditionally associated with the idea of BI reasoning, starting from 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).

This property 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'. This, in turn, ensures 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 applications.

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 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 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) a recursive structure analogous to that of SPE; (ii) the ability to solve the

[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)).

game 'backwards'; (iii) a clear, 'non-equilibrium' formalization of the idea of unexpected deviations as mistakes; (iv) a connection with a 'perfect equilibrium' concept that introduces backward induction 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 may entertain the possibility that the opponents committed *mistakes* in the past. Note that, 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 that, in finite horizon games, 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*. This procedure consists of the iterated application of (static) belief-free rationalizability, to the normal form of the continuation games obtained 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 by appending a type space to the belief-free game. IPE requires that beliefs are updated from one information set to the next via Bayes' rule, whenever possible, both on and off the path. After

unexpected moves, however, a player's beliefs about the opponents' types are completely unrestricted.[2] We show that the set of $\mathcal{BR}$ strategies in the belief-free game coincides with the set of all IPE strategies, for some type space (Theorem 4). Hence, $\mathcal{BR}$ characterizes the *robust predictions* of IPE, i.e. those which do not depend on assumptions on players' exogenous beliefs about each other's types.

At a practical level, these 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, this is also useful to overcome several difficulties that are typically faced in applications, both in complete and in incomplete information settings (see also Catonini and Penta (2022) Penta (2015)).

Overall, these 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 '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, and that the logic of backward induction in general games is distilled precisely by $\mathcal{BR}$.

Our analysis also uncovers that, somewhat in contrast with the established wisdom, absent equilibrium assumptions, BI-reasoning entails an *agnostic* attitude as to whether unexpected moves are interpreted as mistakes or deliberate choices. As we discuss in Section 5.4, the opposite idea of *belief persistence*, which is natural in an equilibrium context, should not be associated with mere BI-reasoning. In particular, absent equilibrium assumptions, belief persistence does not necessarily follow from making an explicit distinction between plans and moves of the opponents, as in the

[2]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. IPE is weaker, for instance, than Perfect Bayesian Equilibrium recently introduced by Watson (2017). It can be shown that IPE is also consistent with a 'trembling-hand' view of unexpected moves, in which no restrictions are imposed on the possible correlations between the trembles and the other elements of uncertainty.

epistemic justification of $\mathcal{BR}$ provided by Battigalli and De Vito (2021) for games with complete information. We discuss this and the alternative justification of Perea (2014) in Section 5.1.

Finally, we consider 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 neighbors 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. We introduce a variation of peer-confirming equilibrium that is based on Backwards Rationalizability, and we show (cf. Theorem 5) that the plain logic of backward induction reasoning it distills, allows to recover, in dynamic settings, the same natural comparative statics that Lipnowski and Sadler (2019)'s original concept exhibits in static games. Then, we discuss other applications that are part of our published or ongoing work. Namely, Penta (2015)'s 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 discusses the applications, and Section 7 concludes.

2 Belief-Free Games

We focus on finite multistage games with observable actions.[3] 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'). Each player i has payoff function $u_i : \mathcal{Z} \times \Theta \to \mathbb{R}$, where $\Theta = \Theta_0 \times ... \times \Theta_n$, with typical element $\theta := (\theta_0, ..., \theta_n)$. *Payoff types* θ_i are private information of each i; the *state of nature* θ_0 is unobserved. A *belief-free game* is thus a tuple $\Gamma = \langle N, \mathcal{H}, \mathcal{Z}, \Theta_0, (\Theta_i, u_i)_{i \in N} \rangle$.[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 terminal history induced by strategy profile s, starting from history h, is denoted $z(s|h)$.[5] Strategic-form payoffs are defined from any 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 $U_i(s, \theta)$). For each h and i, we let $S_i(h)$ denote the set of i's strategies that are compatible with h. Thus, upon reaching history h, player i learns that the behavior of the opponents' strategies are in $S_{-i}(h) = \times_{j \neq i} S_j(h)$. Finally, for each $h \in \mathcal{H}$, S_i^h denotes the set of strategies in the continuation game starting from h, and for each $s_i \in S_i$, $s_i|h$ denotes the continuation of s_i from h.

[3] 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.

[4] Tuple Γ is not a Bayesian game because it does not include a *type space*. Type spaces and Bayesian games are introduced in Section 4.

[5] Notice that $z(s|h)$ does not coincide with $z(s|h^0)$ when h is not along the path induced by s.

3 Backwards Rationalizability

Backwards Rationalizability is a non-equilibrium solution concept for belief-free games. Similar to baseline Rationalizability (e.g., Pearce (1984)), it will also be defined by an iterative deletion procedure, in which players form conjectures about others' information and behavior, and play sequential best responses.

Conjectures are modelled as Conditional Probability System: an array of conditional beliefs, one for each history, derived from Bayesian updating whenever possible. Formally, let $\Theta_{-i} = \times_{j\neq i}\Theta_j$ and $S_{-i} = \times_{j\neq i}S_j$. A Conditional Probability System (CPS) over $\Theta_0 \times \Theta_{-i} \times S_{-i}$ is an array of conditional distributions $\mu^i = (\mu^i(\cdot|h))_{h\in\mathcal{H}}$:

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 $\Delta_i^{\mathcal{H}}$.

Strategy s_i is *sequentially rational* for θ_i, given a CPS μ^i, if at every $h \in \mathcal{H}$, it prescribes optimal behavior in the continuation game given $\mu^i(\cdot|h)$. Formally: for each $h \in \mathcal{H}$ and $s_i' \in S_i$, $\bar{U}_i\left(s_i; \mu^i, h, \theta_i\right) \geq \bar{U}_i\left(s_i'; \mu^i, h, \theta_i\right)$, where

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

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 define Backwards Rationalizability:

Definition 1. *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:(i) $s_i \in r_i(\mu^i; \theta_i)$; (ii) 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* $(\theta_{-i}, s'_{-i}) \in \mathcal{BR}^{k-1}_{-i}$*.*

The set of Backwards Rationalizable *strategies for type* θ_i *is* $\mathcal{BR}_i(\theta_i) = \cap_{k>0}\mathcal{BR}^k_i(\theta_i)$*, 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. Note that we do not ask beliefs to be concentrated on the *full* strategies that survive the previous round, i.e., we do not require $\mu^i(\Theta_0 \times \mathcal{BR}^{k-1}_{-i}|h) = 1$. Such stronger requirement, combined with the property of a CPS that the beliefs at history h should be concentrated on $S_{-i}(h)$, would yield extensive-form rationalizability (and, hence, forward induction reasoning). Players' conjectures about $\Theta_0 \times \Theta_{-i}$ are instead 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.

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* $\gamma \in (1/4, 1/2)$*, the total production is* $3\,(\theta_a + \theta_b)\,/2$*, 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): $\dfrac{1}{2} \cdot \dfrac{3}{2}\,(\theta_a + \theta_b) - \gamma$

At the first round, strategies B.E and P.E are not sequentially rational for $\theta_a = 2$*, because choosing E at history* (B) *is not a continuation best reply to any belief. For 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 $\theta_a = 1$, *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

The following standard properties are easy to check for $\mathcal{BR}$:

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

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 (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 $(\theta_{-i}, s'_{-i}) \in \mathcal{BR}_{-i}$.*[6]

$\mathcal{BR}$ is also robust to changes in the order of elimination of type-strategy pairs, which is helpful in practice. To formalize this property, we rewrite $\mathcal{BR}$ as a reduction procedure. Fix $\hat{\Omega} = \times_{i \in N} \hat{\Omega}_i$, where each $\hat{\Omega}_i$ 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 $\rho^{\mathcal{BR}}(\hat{\Omega}) := \times_{i \in N} \rho_i^{\mathcal{BR}}(\hat{\Omega})$.

Definition 2. *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 $\hat{\Omega}^m \supset \rho^{\mathcal{BR}}(\hat{\Omega}^{m-1})$ at some step m. 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 $\mathcal{BR}$. Moreover, 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

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

clearly survive all steps of $\mathcal{BR}$. Since an order of elimination can stop only when a fixed point is reached (see point 2 in Def. 2), the final output will coincide with $\mathcal{BR}$. This proves the following:

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

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

As discussed, a distinctive feature of backward induction reasoning is that the set of predictions for the whole game, when restricted to a part of the game, coincides with the set of predictions for that part of the game analyzed in isolation. In incomplete information games, a history does not define a "subgame". Thus, instead of "subgame consistency", we focus on *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, under all possible initial beliefs about payoff types.

Formally, let $\mathcal{BR}^h$ denote $\mathcal{BR}$ for the game with root h, and for each 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 $\mathcal{BR}|h = \times_{i \in N} \mathcal{BR}_i|h$.

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*. As the next result will show, this

property 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.

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". But we first need to clarify what it means to solve backwards a game with imperfect and incomplete information. Absent any equilibrium assumption, we do this with a recursive use of belief-free rationalizability on the normal form of each continuation game, starting from preterminal histories and proceeding backwards, at each point maintaining that players can only believe in type-strategy strategies pairs that are consistent with the solution of the subsequent 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 $k > 0$, $R_i^k := \{(\theta_i, s_i) \in \widetilde{\Omega}_i : \exists \nu \in \Delta(\Theta_0 \times R_{-i}^{k-1}) \text{ s.t. } s_i \in \hat{r}_i(\nu; \theta_i)\}$, where $\hat{r}_i(\nu; \theta_i)$ denotes the set of strategies that are a best response for type θ_i to a conjecture $\nu \in \Delta(\Theta_0 \times \Theta_{-i} \times S_{-i})$.[7] Then, for each $i \in N$, the set of *belief-free rationalizable type-strategy pairs* is $R_i := \cap_{k>0} R_i^k$.

Our *Backwards Procedure*, $\mathcal{BP}$, is formally defined as follows:[8]

Definition 3. *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.

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

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

We define $\mathcal{BP}^h$ as the output of belief-free rationalizability on the strategic form that consists of the type-strategy pairs $(\mathcal{BP}_j^{h,0})_{j\in N}$.

Example 2. *Consider again the game of Example 1. To apply the backwards procedure, first we apply 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

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

Going backwards, at $h = h^0$, we initialize belief-free rationalizability with 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 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. 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 $s_{-i}^h \in S_{-i}^h$, 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.[9]

4 Interim Perfect Equilibrium

This section explores the connection between $\mathcal{BR}$ and equilibrium predictions. Following the standard approach, we model exogenous belief hierarchies implicitly (Harsanyi (1967)), by means of *type spaces*, i.e. $\mathcal{T} = (T_i, \vartheta_i, \tau_i)_{i \in N}$, where for each i, T_i is a finite set of types, and $\vartheta_i : T_i \to \Theta_i$ is an onto function assigning the payoff-type to each type.[10] 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

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

[10] Because the game is finite, the restriction to finite type spaces is without loss of generality for our purposes.

initial belief of each type t_i about state of nature and opponents' types. We write $\tau_i(\theta_0, t_{-i}|t_i)$ for the probability that type t_i assigns to (θ_0, t_{-i}). Appending a type space $\mathcal{T}$ to the belief-free game Γ, yields the *Bayesian Game* $\Gamma^{\mathcal{T}}$.

In a Bayesian game, we call *interim strategies* the elements of S_i, and *interim mixed strategies* the elements of $\Delta(S_i)$.[11] 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 $s_i \in S_i$.

In a Bayesian game, an equilibrium is described as a strategy profile coupled with systems of beliefs about the state of nature and the opponents' types, which associates each type of player i with an *array* of beliefs about the state of nature and the opponents' types, one for each history. Formally, $p_i : T_i \to (\Delta(\Theta_0 \times T_{-i}))^{\mathcal{H}}$, and 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 $p = (p_i)_{i \in N}$. For consistency with the type space, we require the initial beliefs specified by p_i to coincide with the ones specified by τ_i. For internal consistency, we require p_i to satisfy Bayesian updating under the assessment. For each h^0, let $\pi(h)$ denote the immediate predecessor of h.

Definition 4. *An assessment (b, p) is* weakly pre-consistent *if, for all $i \in N$ and for all $t_i \in T_i$:*

1. *$p_i(\theta_0, t_{-i}|h^0; t_i) = \tau_i(\theta_0, t_{-i}|t_i)$ for all $(\theta_0, t_{-i}) \in \Theta_0 \times T_{-i}$;*
2. *for each $h \succ h^0$, $p_i(\cdot|h; t_i)$ is derived with Bayes rule from $p_i(\cdot|\pi(h); t_i)$ given b_{-i} whenever possible, and it is arbitrary otherwise.*

Given an assessment (b, p), for each player i and type t_i, one can construct the CPS $\hat{\mu}^{t_i}_{(b,p)}$ over $\Theta_0 \times T_{-i} \times S_{-i}$ induced by $p_i(t_i)$ and b_{-i}, and in turn, the CPS $\mu^{t_i}_{(b,p)}$ over $\Theta_0 \times \Theta_{-i} \times S_{-i}$ induced by $\hat{\mu}^{t_i}_{(b,p)}$.[12]

[11] We use mixed strategies instead of behavior strategies for notational convenience. Given a type space, there could be more IPE in mixed strategies than in behavior strategies. But the equivalence between $\mathcal{BR}$ and IPE across types spaces also holds for IPE in behavior strategies.

[12] See the proof of Theorem 4 for the formal definitions and Example 3 for a construction.

Definition 5. *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 given $\mu^{t_i}_{(b,p)}$.*

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*

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

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

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

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 $\hat{\mu}^{t_a^2}$,

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

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. ▲

IPE can be seen as a dynamic counterpart of *interim equilibrium* (Bergemann and Morris (2005)), obtained by imposing two natural conditions: (i) weak preconsistency, and (ii) sequential rationality. Weak preconsistency implicitly imposes a standard equilibrium notion of *belief persistence* (cf. Section 5.4), in that it requires the beliefs about strategies to be consistent with the equilibrium strategy profile, b, both on and off the equilibrium path. Beyond this, however, it imposes no restrictions on the beliefs about others' types that players hold 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.

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, $\mathcal{BR}$ – which is a non-equilibrium concept for belief-free dynamic games – 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, where the robustness criterion refers to the predictions across *all* type spaces, whether or not they admit a common prior.

5 Discussion

5.1 Epistemic justifications of Backwards Rationalizability

In games with complete information, Perea (2014) justified $\mathcal{BR}$ with the epistemic conditions of "rationality and common belief of future rationality". Such epistemic justification highlights the forward-looking nature of $\mathcal{BR}$: at the every information set, the behavior of an opponent is predicted based on the continuation of the game, and not on her past moves (as opposed to strong rationalizability, i.e., forward-induction reasoning). Also focusing on games with complete information, Battigalli and De Vito (2021) introduce in their epistemic model a formal separation between a player's plan and her actual moves. In this world, sequential rationality is captured by "optimal planning and consistency", and $\mathcal{BR}$ is justified by the additional conditions of "common full belief in optimal planning and *continuation* consistency," that is, the correct implementation of the plan from every information set onwards. Under these epistemic hypotheses, upon observing an unexpected move by the opponent, a player revises her joint belief about the opponent's plan and implemented strategy, whereas keeping the marginal belief on the plan and revising the one on the strategy would capture the form of belief persistence that we analyze in detail in Section 5.4.

5.2 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 SPE strategies 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 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 those played under the complete information. 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.[13] We illustrate the point by constructing a type space $\hat{\mathcal{T}}$ and an IPE in which $In.c$ is played by some

[13]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)

type of player 1:Let $\hat{T}_1 = \{t_1^{Out.c}, t_1^{Out.d}, t_1^{In.c}\}$ and $\hat{T}_2 = \{t_2^f, t_2^g\}$, with beliefs:[14]

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

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 $\mathcal{T}$ 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, 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.

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

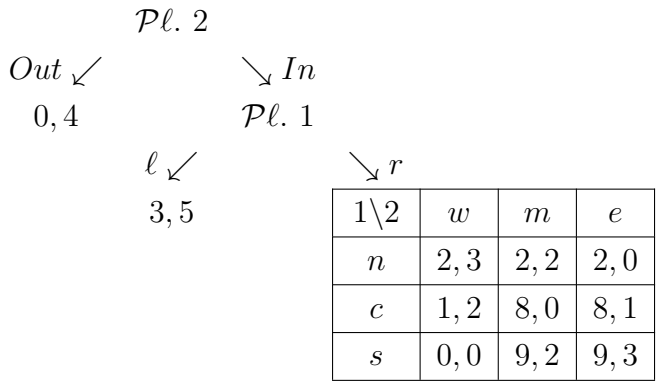
To understand the difference between $\mathcal{BR}$ and $\mathcal{BP}$ in terms of full strategies, note that compared to $\mathcal{BP}$, $\mathcal{BR}$ further eliminates any strategy that is a continuation best reply to some belief at h, but not a continuation best reply to the *same* belief at some 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.[15] This is the reason why the

[14]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.

[15]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

possible behavior of each player in each continuation game is exactly the same under 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:*



$1\backslash 2$	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*

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 $\mathcal{BP}$.

In and observes r, she must update her initial belief. Since $r.n$ initially received zero probability, she still cannot give positive probability to it. Hence, 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, and that is "belief persistence": the idea that players never change their belief about the strategies of the opponents, no matter how many deviations from the expected strategies they have observed.[16] A possible

[16] Also our definition of IPE captures this idea: the beliefs about the behavior of the opponent are entirely determined by the opponent's equilibrium strategy b_{-i}.

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. 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)). 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.

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 5.2, 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 5.2, 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. (The proof of Theorem 4 shows how to construct such a type structure.) 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 point is related to the one we discussed in Section 5.2, but it is distinct in that it speaks directly to the role that type spaces have in determining the bite of the *belief persistence* assumption in Bayesian games. Specifically, 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 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 player's beliefs about non-neighbors must be consistent, whenever possible and both on- and off-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 backward induction logic that players apply to their neighbors (embedded in subgame perfect equilibrium), 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 developed in this paper, we propose 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 (Theorem 5).

Formally, for each $i \in N$, let $N^i \subseteq N$ denote i's 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 6. *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, in the spirit of equilibrium consistency restrictions, the candidate strategy profile may restrict 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 Lipnowski and Sadler (2019)'s definition of peer-confirming equilibrium, 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 6 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 provides a seamless extension of Bergemann and Morris (2009) static analysis to dynamic environments, in which agents may obtain information over time. Second, it sheds light on Bergemann and Morris (2007)'s results on the advantages of using dynamic mechanisms in static environments. In particular, it shows that the intuition that robustness may be favored by the adoption of dynamic mechanisms, thanks to the reduction of strategic uncertainty granted by backwards induction, does not extend to incomplete information settings, at least if no restrictions on beliefs are imposed.[17]

Importantly, however, the advantages of Backwards Rationalizability can be seen even in settings with *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

[17]For robust implementation with belief restrictions, see Ollár and Penta (2017, 2023, 2022).

chosen in the first stage. Yet, in the baseline specification with linear transportation cost (Hotelling (1929)), SPE fails to provide a tractable and convincing solution.[18] 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 faded, leaving it as some kind of puzzle, despite its inherent simplicity. The tools developed in this paper may be fruitfully applied to think about the two-period Hotelling model afresh. In Catonini and Penta (2022), we look at subsets of backwards rationalizable strategy profiles that are "closed under rational behavior" (Basu and Weibull (1991)) along the induced path. It turns out that the transportation-efficient location pair is the *only* location pair that is consistent with this solution concept, which captures the idea that firms know the path of play, but face strategic uncertainty after a deviation, and try to reduce it with backward induction reasoning.

Overall, these applications show that Backwards Rationalizability not only is a tractable and ready-to-use solution concept, but it may also serve as a basis to impose extra desiderata 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. These extra desiderata may be dictated by the specific economic context, and and super-imposed on Backwards Rationalizability, which may thus serve as a template solution concept.

Appendix

[18] 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.

[19] For instance, d'Aspremont et al. (1979) considered 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 interval), a result which is considered at odds with factual observation.

A Proofs

We 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.

We will repeatedly use two facts, which we report without a formal proof. Fix two histories h, h', a CPS $\mu^{i,h}$ over $\Theta_0 \times \Theta_{-i} \times S_{-i}^h$, and 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 $(\bar{\theta}_0, \bar{\theta}_{-i}, s_{-i}^{h'}) \in \Theta_0 \times \Theta_{-i} \times S_{-i}^{h'}$ such that $s_{-i}^{h'} = s_{-i}^h | h'$ if $h' \succeq h$, $s_{-i}^{h'} | h = s_{-i}^h$ and $s_{-i}^{h'} \in S_{-i}^{h'}(h)$ otherwise — note that the map keeps $(\bar{\theta}_0, \bar{\theta}_{-i})$ fixed, and all the maps we will use will do so. Consider the array of beliefs $(\mu^{i,h'}(\cdot|h''))_{h''\succeq h,h'}$ where each $\mu^{i,h'}(\cdot|h'')$ is the pushforward of $\mu^{i,h}(\cdot|h'')$ through ς.

Fact 1: $(\mu^{i,h'}(\cdot|h''))_{h''\succeq h,h'}$ is (part of) a CPS for the continuation game with root h'.

Fact 2: $\mu^{i,h'}(\cdot|h'')$ and $\mu^{i,h}(\cdot|h'')$ have the same continuation best replies for all types.[20]

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 $\mathcal{BR}^{k-1} | h = \mathcal{BR}^{h,k-1}$.

[20]Formally, the continuation best replies are elements of S_i^h and $S_i^{h'}$, so when $h \neq h'$ what we mean is that the continuation best replies, which only depend on the actions they prescribe from h'' onwards, coincide from h'' onwards.

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

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

Let $\mu^{i,h} = \left(\mu^{i,h}(\cdot|h')\right)_{h' \succeq h}$ be the CPS over $\Theta_0 \times \Theta_{-i} \times S_{-i}^h$ where, for each $h' \succeq h$, $\mu^{i,h}(\cdot|h')$ is 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, since $\mu^i(\cdot|h')$ and $\mu^{i,h}(\cdot|h')$ have the same continuation best replies for all types, 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 $\overline{s}_{-i}|\pi(\overline{h}) \notin S_{-i}^{\pi(\overline{h})}(h)$, but since $\overline{s}_{-i}|\overline{h} \in S_{-i}^{\overline{h}}(h)$, we must have $\overline{s}_{-i}|\pi(\overline{h}) \notin S_{-i}^{\pi(\overline{h})}(\overline{h})$. 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 $\bar{\mu}^i(\Theta_0 \times \Theta_{-i} \times S_{-i}(\overline{h})|\pi(\overline{h})) = 0$. 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 $(\bar{\theta}_0, \bar{\theta}_{-i}, s_{-i}) \in \Theta_0 \times \Theta_{-i} \times S_{-i}(h)$ 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 $\mu^i(\cdot|h') = \mu^i(\cdot|h)$;

[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.

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. Now we show that μ^i is viable for $\mathcal{BR}_i^k$. 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 $(\theta_{-i}, s_{-i}|h') \in \mathcal{BR}_{-i}^{k-1}|h'$. The same holds for every other h' by 3. and 4., so μ^i is viable for $\mathcal{BR}_i^k$. Finally, for each $h' \succeq h$, by 1. and (a), $\mu^i(\cdot|h')$ and $\mu^{i,h}(\cdot|h')$ have the same continuation best replies, therefore μ^i justifies $(\theta_i, s_i) \in \mathcal{BR}_i^k$ for some s_i such that $s_i|h = s_i^h$. ■

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 7. *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 $(\bar{\theta}_0, (\bar{\theta}_j, \widetilde{s}_j^h)_{j\neq i}) \in \Theta_0 \times \mathcal{B}\hat{\mathcal{R}}_{-i}^{h,K+k-1}$ such that $\widetilde{s}_j^h \simeq s_j^h$ 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.

Given an IPE (b, p), for each $i \in N$ and $t_i \in T_i$, we will formally describe the equilibrium beliefs of t_i as a CPS $\hat{\mu}^{t_i} = (\hat{\mu}^{t_i}(\cdot|h))_{h\in\mathcal{H}}$ over $\Theta_0 \times T_{-i} \times S_{-i}$, defined by the following recursive procedure, which uses the notion of *replacement plan*: 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 $s_i'(h') = s_i^{h'}(h)$ for every $h' \not\succ h$. [22]

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

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

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

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

[22] In words, the replacement plan of s_i at h is the strategy that specifies actions consistent with h on the path to h but coincides with s_i everywhere else, most importantly at histories that follow h.

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

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

"If" part) For each $i \in N$ and $t_i \in T_i$, define a CPS μ^{t_i} 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

$$\mu^{t_i}(\theta_0, \theta_{-i}, s_{-i}|h) = \hat{\mu}^{t_i}(\{\theta_0\} \times \vartheta_{-i}^{-1}(\theta_{-i}) \times \{s_{-i}\}\,|h),$$

where $\vartheta_{-i}^{-1}((\theta_j)_{j\neq i}) = \times_{j\neq i}\vartheta_j^{-1}(\theta_j)$. Thus, for each $s_i \in S_i$ such that $b_i(s_i|t_i) > 0$, s_i is sequentially rational for $\vartheta_i(t_i)$ given μ^{t_i}. At every history h, μ^{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 we construct the belief map. Fix $t_i = (\theta_i, s_i)$. By the fixed-point property of $\mathcal{BR}$ (Remark 2), there is 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 $s'_j|h = s_j|h$ for every $j \neq i$. Let $\tau_i(\cdot|t_i)$ be the pushforward of $\mu^{t_i}(\cdot|h^0)$ through $\xi_{h^0}^{t_i}$.

Now we construct an IPE (b, p) where, for every $i \in N$ and $(\theta_i, s_i) \in \mathcal{BR}_i$, there exists $t_i \in T_i$ such that $b_i(s_i|t_i) = 1$ and $\vartheta_i(t_i) = \theta_i$.

For each $i \in N$, define the strategy b_i as $b_i(s_i|t_i) = 1$ for each $t_i = (\theta_i, s_i)$. By construction of the type structure, b satisfies the desired condition.

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 $\hat{\mu}^{t_i}(\cdot|h^0)$ with equation 2. Now fix $h \succ h^0$ and suppose that $\hat{\mu}^{t_i}(\cdot|\pi(h))$ was defined. If $\hat{\mu}^{t_i}(\Theta_0 \times T_{-i} \times S_{-i}(h)|\pi(h)) > 0$, derive $\hat{\mu}^{t_i}(\cdot|h)$ 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 $\hat{\mu}^{t_i}(\cdot|h)$ 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 $\hat{\mu}^{t_i}(\Theta_0 \times T_{-i} \times S_{-i}(h)|\pi(h)) = 0$. Then, for each $\omega = (\theta_0, (t_j)_{j\neq i}, (s_j)_{j\neq i})$, we have

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

For each $\omega' = (\theta_0, (\theta_j)_{j\neq i}, (s'_j)_{j\neq i}) \in \left(\xi_h^{t_i}\right)^{-1}(\theta_0, (t_j)_{j\neq i})$, for every $j \neq i$, we have $\theta_j = \vartheta_j(t_j)$ and $s'_j|h = b_j(t_j)|h$, therefore if $s_j = \varrho_{j,h}(b_j(t_j))$, $s'_j|h = s_j|h$. Hence, $\hat{\mu}^{t_i}(\cdot|h)$ and $\mu^{t_i}(\cdot|h)$ induce the same belief over payoff-relevant types and continuation strategies. So, since s_i is a continuation best reply to $\mu^{t_i}(\cdot|h)$, it is also optimal under the (candidate) equilibrium belief $\hat{\mu}^{t_i}(\cdot|h)$. The same is true at every other history h as $\hat{\mu}^{t_i}$ 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. and De Vito, N. (2021), 'Beliefs, plans, and perceived intentions in dynamic games', *Journal of Economic Theory* pp. 1–43.

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.

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

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

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.

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

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.

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. (2022), 'Efficient full implementation via transfers: Uniqueness and sensitivity in symmetric environments', *AEA papers and proceedings* **112**, 438–443.

Ollár, M. and Penta, A. (2023), 'A network solution to robust implementation: The case of identical but unknown distributions', *Review of Economic Studies* .

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

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

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

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

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.

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* .