Respect, Iterative Admissibility, and Equilibrium Refinements.*

Mikhail Panov†

May 31, 2020

Abstract

I develop a theory of "interspection" for finite extensive-form games of perfect recall. This theory addresses the question of how sufficiently sophisticated players will behave in games if suboptimal actions are interpreted by other players as evidence of a player's lower sophistication level, rather than inconsequential one-time trembling mistakes.

I show that interspection theory has several attractive features. In particular, under interspection theory, normal-form analysis of games is sufficient: extensive-form games can be equivalently analyzed in the corresponding normal forms. Also, under interspection theory, any player's behavior is admissible: players do not select weakly dominated strategies.

I introduce the notion of interspected rationalizability and provide an axiomatic characterization of environments to which it applies. I also introduce the notion of interspected equilibrium, prove its existence, and discuss its properties.

*I am grateful to Piotr Dworczak, Yossi Feinberg, Michael Harrison, David Kreps, Erik Madsen, Joshua Mollner, Gleb Romanyuk, Ilya Segal, Andrzej Skrzypacz, Alexander Wolitzky, Pavel Zryumov, and especially Michael Ostrovsky, Sergey Vorontsov, and Robert Wilson for helpful comments, suggestions, and discussions.

†Graduate School of Business, Stanford University; mpanov@stanford.edu.

Contents

If you can keep your head when all about you
Are losing theirs and blaming it on you,
If you can trust yourself when all men doubt you,
But make allowance for their doubting too.

Rudyard Kipling

1 Introduction

This paper studies general finite extensive-form games of perfect recall. I develop a theory to address the following question:

> Which strategies will be chosen by sufficiently intelligent players if suboptimal actions are interpreted by other players as evidence of a player's lower sophistication level, rather than inconsequential one-time trembling mistakes?

The first part of the question: "which strategies will be chosen by sufficiently intelligent players" is a standard question that almost any theory of games aims to answer. However, a theory of games needs to describe the behavior of players both along the predicted path and in the case of observed deviations. To do so, a theory of games should also specify how players interpret possible deviations.

One way to interpret deviations is to attribute them to one-time inconsequential trembling-hand mistakes, following Selten (1975). In such a case, a deviation from the predicted path would be commonly perceived by other players as bearing no signal about the characteristics of the deviator.

The current paper considers an alternative way of treating deviations: the players think that trembling mistakes are impossible. In the absence of trembling-hand mistakes, suboptimal deviations must then signal deviators' lack of sophistication (c.f. Reny (1993)).

The emphasis of this paper is on *"interspection"* in games. The word "introspection" is defined in the dictionary as "the examination or observation of one's own mental and emotional processes." By analogy, the word "interspection" is understood here as "the examination or observation of other people's mental and emotional processes." Specifically, in a game, a player "interspects" when he thinks about what his opponents play and why, what they think about what their opponents play and why, what they think their opponents think about the thought processes of their opponents, and so on.

To model interspection, I introduce the concept of *interspecting type*. Each player has some interspecting type $t = (k; \sigma; \Omega^k; \boldsymbol{P})$. Here $k \geqslant 0$ denotes the player's finite *interspection level*; σ is a pure strategy the player intends to play; Ω^k is the player's subjective state-space, called the *interspection state-space*; and $\boldsymbol{P}_i$ is the player's subjective probability assessment on Ω^k. The state-space Ω^k encodes the combinations of the player's opponents' strategy profiles and their interspection levels for all interspection levels *lower* than k. A player with interspection level 0 is irrational and can play any strategy. If $k > 1$, the strategy σ intended by the player must be a best response to his assessment $\boldsymbol{P}$. A player with interspection level 1 forms an assessment on his opponents'

strategy profiles, but is incapable of interspecting their behavior. I.e., a level-1 interspecting player may form any full-support subjective probability assessment on his opponents' strategy profiles. A player with interspection level 2 forms a full-support assessment on his opponents' strategy profiles, distinguishing strategies played by opponents with interspection level 0 (irrational), and interspection level 1 (level-1 interspecting). And so on.

A player is *rational* if his intended strategy is a best response to his subjective probability assessment. Departing from the classical Bayesian model, I use the framework of *full-support* lexicographic probability systems (LPS) developed by Blume et al. (1991a). A *cautiously rational* player then is a player who plays a best-response strategy to some LPS with *full support* on his subjective state-space.

The model of cautious rationality has two attractive features. First, for the theory developed here, cautious rationality ensures the equivalence of normal-form and extensive-form analyses. I.e., with certain conditions on players' subjective state-spaces, a normal-form strategy that is *initially* optimal against a cautious assessment will be *sequentially* optimal against the Bayesian updates of the same assessment in the extensive-form game. Second, cautious rationality guarantees admissibility of selected strategies. I.e., a cautiously rational player never plays weakly dominated strategies.

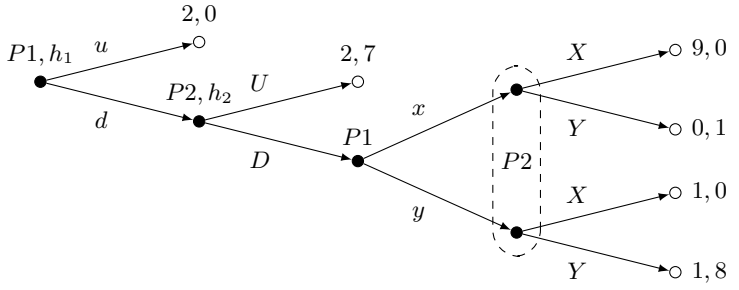
Finally, I put a restriction on how severely players, upon seeing a deviation, can downgrade the deviator's perceived interspection level. Informally, a player *respects* his opponents if he assesses that it is *infinitely more likely* that they have higher interspection levels rather than lower ones. I assume that it is common knowledge among the players that they respect each other (c.f. Battigalli (1996), Battigalli and Siniscalchi (2002)).

1.1 Main results

The main results of the paper can be divided into tree parts.

First, I show two *equivalence results*. The first equivalence result states that in interspection theory, extensive-form games can be completely analyzed using their normal forms only. I.e., for a cautiously rational player, with certain conditions on his subjective state-space, a strategy that is optimal initially is optimal sequentially. The second equivalence result translates the formal notion of Respect, expressed in terms of assessments on interspection state-spaces, into the language of assessments on strategy profiles. Thus, in interspection theory, instead of analyzing the epistemic model, one can work directly with primitives of the game.

Second, I give an explicit construction of *interspected rationalizability*. I provide an axiomatic characterization of *environments* in which the players can be expected to play precisely interspected rationalizable strategies. For two-player games interspected rationalizability coincides with iterative admissibility (elimination of weakly dominated strategies in rounds). However, for general multi-player games these two concepts are different. For perfect information games without chance nodes and without relevant ties, the unique interspected rationalizable outcome is the unique backward induction outcome.



(a) Extensive form Γ

	U	DX	DY
u	$2,0$	$2,0$	$2,0$
dx	$2,7$	$9,0$	$0,1$
dy	$2,7$	$1,0$	$1,8$

(b) Normal form $G(\Gamma)$

Figure 1: Example

Third, I introduce the concept of *interspected equilibrium* as a refinement of interspected rationalizability. For general finite games, interspected equilibria always exist and satisfy invariance, (weak) sequential rationality, and admissibility. For two-player games, interspected equilibria also satisfy a certain forward induction property: an interspected equilibrium of the full game corresponds to an interspected equilibrium of the reduced game after a one-round elimination of *all never-best-response* strategies of *one* of the players. For multi-player games, interspected equilibria only satisfy a weaker version of forward induction: an interspected equilibrium of the full game corresponds to an interspected equilibrium of the reduced game after elimination of *all interspected non-rationalizable strategies* of *all* of the players. The path of interspected equilibrium may be not supportable in a subgame perfect equilibrium even in perfect information games. Finally, interspected equilibria do not always satisfy on-the-path backward induction. I.e., an interspected equilibrium does not necessary induce an interspected equilibrium of a subgame reachable with positive probability under the equilibrium play.

1.2 Example

To illustrate informally the ideas behind interspection theory, consider the game from Myerson (1997), p. 192. The extensive-form game and its normal form are given in Figure 1a and Figure 1b.

The unique subgame perfect equilibrium (SPNE) of this game is $([u],[D],[y],[Y])$. This outcome may be interpreted as a predicted outcome under common knowledge of rationality, with deviations being attributed to inconsequential one-time trembling mistakes.

Let us now analyze this game from the perspective of interspection theory.

If the players are level-0 interspecting, i.e., irrational, they can play any strategy: Player 1 can play $\{u, dx, dy\}$ and Player 2 can play $\{U, DX, DY\}$.[1]

If the players are level-1 interspecting, they play best-response strategies to some full support LPS assessments on the sets of strategies of their opponents. In this example, this is equivalent to the play of best-response strategies to some full-support probability measures. I.e., cautious

[1]Each player is assumed to play only pure strategies. A mixed strategy is understood as a pure strategy to commit to a randomizing device. Allowing for mixed strategies would not change the predictions of interspection theory. C.f. Aumann (1987).

level-1 interspecting players should not play weakly dominated strategies. In particular, a level-1 interspecting Player 1 would never play strategy $[dy]$. Similarly, a level-1 interspecting Player 2 would never play strategy $[DX]$. Thus, in this game, level-1 interspecting players may only play strategies $\{u, dx\}$ and $\{U, DY\}$.

Suppose the players are level-2 interspecting. Then, they can calculate all possible plays of 1- and level-0 interspecting opponents. If the players respect their opponents, they assess that it is infinitely more likely that their opponents have interspection level 1, rather than 0. In particular, a level-2 interspecting Player 1 who respects Player 2 will play a best response to some full-support probability measure on $\{U, DY\}$. Similarly, a level-2 interspecting Player 2 who respects Player 1 will play a best response to a full-support probability measure on $\{u, dx\}$. Therefore, in this game, players who have at least two levels of interspection and who respect their opponents, may only play strategies $\{u\}$ and $\{U\}$.

Note that the prediction $\{u\} \times \{U\}$ generated by interspection theory prescribes action $[U]$ to Player 2 at the node h_2. This is different from the SPNE prescription $[D]$. The reason is that under these two theories, Player 2 is supposed to think differently if he observes a deviation $[d]$ from Player 1. Specifically, under the theory of subgame perfect equilibrium, Player 2 thinks:

> It is commonly known at the beginning of the game that we are supposed to play the equilibrium $([u], [D], [y], [Y])$. However, Player 1 deviated. Since I know that there is common knowledge of the equilibrium and of rationality, I think that Player 1 made a trembling mistake. Thus, I believe that he should know the continuation of the equilibrium and should believe in that I will conform to it. Since I think he is rational and would also find it optimal to conform, I expect him to play $[uy]$, which makes it optimal for me to play $[D]$.

On the contrary, under interspection theory, Player 2 thinks:

> Action $[d]$ is inconsistent with Player 1 having at least 2 interspection levels and respecting me. Since he did not make a trembling mistake and since he always respects me, I conclude that he is either irrational (level-0), or level-1 interspecting. Since I respect Player 1, I think it is infinitely more likely that he is level-1 interspecting rather than irrational. Therefore, I expect him to play $[dx]$. Against such belief I find it optimal to play $[U]$.

The thought process behind Player 2's reaction under interspection theory corresponds to the logic of forward induction. Myerson (1997) identifies one more outcome of this game corresponding to the forward induction logic: $\{dx\} \times \{U\}$. This outcome is eliminated in interspection theory since this outcome is inconsistent with Player 1 being cautious. Indeed, the play of $[u]$ gives the payoff 2 to Player 1 for sure. The play of $[d]$ is not certain. Under interspection theory, Player 1 expects that with probability 1, Player 2 will play $[U]$, giving the same payoff, 2, to Player 1. But the second most likely response is $[DY]$, which is strictly worse for Player 1. Thus, Player 1 would

prefer to get the payoff 2 for sure, rather than to pass the turn to Player 2. I.e., he would rather play $[u]$.

1.3 Related literature

Bernheim (1984) and Pearce (1984) were among the first to study rationalizable behavior in games. Specifically, Pearce (1984) proposed rationalizability concepts for both normal-form and extensive-form games. Battigalli (1996) showed that the logic of Pearce (1984)'s extensive-form rationalizability (EFR) can be expressed by the so-called *Best Rationalization Principle*: "a player should always believe that her opponents are implementing one of the 'most rational' (or 'least irrational') strategy profiles which are consistent with her information." Subsequently, Battigalli and Siniscalchi (2002), using the epistemic model of Battigalli and Siniscalchi (1999), provided a further characterization of EFR in terms of common strong belief in expected utility maximization. The ideas behind interspection theory and especially behind interspected rationalizability are closely related to the ideas expressed in that sequence of papers. However, interspection theory is different in the following aspects: First, EFR is formulated for usual-type rational players, while interspected rationalizability is formulated for cautiously rational players. This fact ensures equivalence of normal-form and extensive-form analysis under interspection theory. Second, in the model of Battigalli and Siniscalchi (1999), subjective assessments are modeled as lexicographic conditional probability systems (LCPSs) distributed on epistemic type-spaces, while under interspection theory, subjective assessments are lexicographic probability systems (LPSs) distributed on interspection state-spaces. This makes the model of interspection theory much more tractable: in finite games, players form assessments only on finite state-spaces. Third, the framework of interspection theory is also convenient for formalizing the equilibrium concept as well as the rationalizability concept.

Cautious rationality is modeled in interspection theory with the use of full-support LPSs of Blume et al. (1991a). In the companion paper, Blume et al. (1991b) showed that such classical equilibrium concepts as perfect and proper equilibrium can be equivalently characterized as equilibria played by cautiously rational players. Another application of full-support LPSs is Brandenburger et al. (2008), which provides epistemic conditions under which players will play iteratively admissible strategies.

An Interspected equilibrium is a special case of Reny (1992)'s weak sequential equilibrium. In a weak sequential equilibrium, players interpret a deviation as a signal that the deviator is not aware of the equilibrium, similar to the case of interspected equilibrium. Even more closely, the concept of interspected equilibrium is related to the concept of Reny (1992)'s explicable equilibrium. The main difference is that in an explicable equilibrium, upon seeing a deviation, players attribute to the deviator strategies that are still logically possible and can be played in *equilibria* of the highest order of explicability. In an interspected equilibrium, the behavior of the deviators is assumed to be of the highest order of *rationalizability*, but does not have to correspond to any equilibrium. Another difference is that unlike explicable equilibria, interspected equilibria are defined for cautiously rational players directly in the normal form of a game, rather than in the extensive form. This fact

makes the properties of interspected equilibria easier to study.

The concept of interspected equilibrium introduced here satisfies a certain forward induction property. The idea of forward induction was first introduced by Kohlberg and Mertens (1986), and later by Van Damme (1989). Besides Battigalli and Siniscalchi (2002) mentioned above, other papers have also studied possible ways to formalize the logic of forward induction. See for example, Cho and Kreps (1987), Govindan and Wilson (2009), and, for the discussion of a forward induction property of proper equilibrium, Blume et al. (1991b).

Finally, the idea that players may play strategies corresponding to finite levels of interspection is similar to level-k thinking introduced by Stahl and Wilson (1994), Stahl and Wilson (1995), and Nagel (1994). Thus, interspection theory may shed some additional light on their ideas.

2 Basic Concepts

In this section I review some basic concepts used in interspection theory.

2.1 Terminology and notation

Normal-form game. The definition of N-player normal-form game G is standard. The notation is as follows:

- the players are Player 1, 2,..., N;
- σ^i and μ^i are pure and randomized strategies of Player i;
- $\mathcal{S}_i$ is the set of all pure strategies of Player i;
- $u_i(\sigma^1, \sigma^2, ..., \sigma^N)$ is the payoff to Player i from the play of pure strategy profile $(\sigma^1, \sigma^2, ..., \sigma^N)$; correspondingly, $u_i(\mu^1, \mu^2, ..., \mu^N)$ is the expected payoff to Player i from the play of randomized strategy profile $(\mu^1, \mu^2, ..., \mu^N)$.

Extensive-form game of perfect recall. The definition of extensive-form game of perfect recall Γ is standard as well (c.f. Kuhn (1953)). The notation is the same as in Kreps and Wilson (1982).

Normal form corresponding to an extensive-form game. For each N-player extensive-form game of perfect recall Γ, the corresponding normal-form game $G(\Gamma)$ is defined as an N-player game in normal form. Normal-form strategies are strategies that prescribe actions to the player only in strategy-relevant (logically reachable) information sets.[2] The payoffs in the normal-form game $G(\Gamma)$ for a given strategy profile correspond to the terminal payoffs in Γ, possibly after taking objective expectation operator if Γ has chance nodes. This version of normal form is referred to as *reduced normal form* in Fudenberg and Tirole (2000). However, it is different from Kohlberg and Mertens (1986)'s reduced normal form.

[2]Following Kuhn (1953), an information node h is relevant to Player i's strategy σ^i if there is a strategy profile $\sigma^{-i} \in \mathcal{S}_{-i}$ for Player i's opponents such that h is reached with positive probability under the play of $(\sigma^i; \sigma^{-i})$

2.2 Cautious (admissible) rationality

A key concept in interspection theory is the notion of cautious (admissible) rationality. This notion was formally developed by Blume et al. (1991a, subsequently BBD1) with the use of full support lexicographic probability systems (LPSs). I start by defining cautious rationality using the framework of BBD1, which is then adapted to the setting of games.

Suppose a decision maker faces an exhaustive *finite* set of states Ω and a set of pure consequences $\mathcal{C}$. Let $\mathcal{P}$ denote the set of simple objective probability distributions on $\mathcal{C}$. The decision maker has (weak) preferences $\succeq$ over acts, which are maps from Ω into $\mathcal{P}$. In the terminology of Anscombe and Aumann (1963), acts are horse lotteries over states, that in turn contain roulette lotteries over pure consequences. The ωth coordinate of act x is denoted as x_ω. Nonempty subsets of Ω are termed events. For any event $S \subseteq \Omega$, x_S denotes the tuple $(x_\omega)_{\omega \in S}$. Also, x_{-S} denotes $x_{\Omega \setminus S}$. A constant act x maps each state into the same roulette lottery over pure consequences: $x_\omega = x_{\omega'}$ for all $\omega, \omega' \in \Omega$.

Consider the following axioms (the numbering as in BBD1):

Axiom CR 1 (Order). *$\succeq$ is a complete and transitive binary relation on $\mathcal{P}^\Omega$.*

Axiom CR 2 (Objective Independence). *For all $x, y, z \in \mathcal{P}^\Omega$ and $0 < \alpha \leqslant 1$, if $x \succ$ (respectively $\sim$) y, then $\alpha x + (1-\alpha)z \succ$ (respectively $\sim$) $\alpha y + (1-\alpha)z$.*

Axiom CR 3 (Nontriviality). *There are $x, y \in \mathcal{P}^\Omega$ such that $x \succ y$.*

Definition (Conditional Preferences). *Conditional preferences $\succeq_S$ on the event $S \subseteq \Omega$ are determined as follows:*

$$x_S \ \succeq_S \ y_S \iff \left(\exists z \in \mathcal{P}^{\Omega \setminus S}, (x_S, z_{-S}) \succeq (y_S, z_{-S})\right)$$

For one state event $\{\omega\}$, conditional preferences will be denoted as $\succeq_\omega$.

Axiom CR 4′ (Conditional Archimedian Property). *For each $\omega \in \Omega$, if $x \succ_\omega y \succ_\omega z$, then there exist $0 < \alpha < \beta < 1$ such that $\beta x + (1-\beta)z \succ_\omega y \succ_\omega \alpha x + (1-\alpha)z$.*

Axiom CR 5′ (State Independence). *For all states $\omega, \omega' \in \Omega$ and for any two constant acts $x, y \in \mathcal{P}^\Omega$, $x \succ_\omega y$ if and only if $x \succ_{\omega'} y$.*

Cautious preferences over acts are defined as follows:

Definition (Cautious Preferences). *Preferences over acts $\mathcal{P}^\Omega$, which satisfy BBD Axioms CR 1,2,3,4′, and 5′, are called cautious preferences.*

As usual, a decision maker who choses an act optimal according to his preferences is called *rational.*

Definition (Cautious Rationality). *A rational decision maker with cautious preferences is called cautiously rational.*

A lexicographic probability system (LPS) is a K-tuple $\rho = (\lambda_1, ..., \lambda_K)$, for some natural K, of nonnegative nonzero measures on Ω. An LPS ρ has full support if Ω is covered by the supports of the measures from ρ:

$$\Omega = \bigcup_{k=1}^{K} supp(\lambda_k)$$

The following is a slightly restated result from BBD1, which provides a lexicographic representation of cautious preferences:

(BBD1) Corollary 3.1. *Preferences $\succeq$ over acts in $\mathcal{P}^\Omega$ are cautious preferences if and only if they can be represented by some affine function $u : \mathcal{P} \to \mathbb{R}$ and some LPS $\rho = (\lambda_1, ..., \lambda_K)$ on Ω. I.e., for all acts $x, y \in \mathcal{P}^\Omega$:*

$$x \succeq y \iff \Big(\sum_{\omega \in \Omega} \lambda_k(\omega) u(x_\omega) \Big)_{k=1}^{K} \geqslant_L \Big(\sum_{\omega \in \Omega} \lambda_k(\omega) u(y_\omega) \Big)_{k=1}^{K}$$

where $\geqslant_L$ is the standard lexicographic order. Furthermore, u is unique up to positive affine transformations. There is minimal $K \leqslant |\Omega|$. Among LPSs of minimal length K representing $\succeq$, each λ_k is unique up to linear combinations of λ_1,..., λ_k that assign positive weight to λ_k. Finally, for each representation ρ and for each ω, there is k such that $\lambda_k(\omega) > 0$.

Now, define the concept of subjective probability assessment. Take a cautiously rational decision maker with preferences $\succeq$ over acts in $\mathcal{P}^\Omega$. Fix any lexicographic representation (u, ρ) of $\succeq$. For any act $x \in \mathcal{P}^\Omega$ define the *imputed act* $x^{imp} \in \mathbb{R}^\Omega$ as $x^{imp} = \{u(x_\omega)\}_{\omega \in \Omega}$. LPS ρ then defines the preferences of the decision maker over imputed acts. These preferences then may be extended to the whole $\mathbb{R}^\Omega$. Call these preferences *imputed preferences* $\succeq^{imp}$ of the decision maker on $\mathbb{R}^\Omega$ determined by representation (u, ρ). Remarkably, imputed preferences do not depend on the particular lexicographic representation of $\succeq$.[3] Thus, for any cautiously rational decision maker, the corresponding imputed preferences are defined uniquely. Following the standard path in Bayesian decision theory, I identify the imputed preferences with the subjective probability assessment:

Definition (Subjective Probability Assessment)**.** *For a cautiously rational decision maker with preferences $\succeq$ over acts $\mathcal{P}^\Omega$, subjective probability assessment $\boldsymbol{P}[\succeq]$ on Ω is defined as the imputed preference relation $\succeq^{imp}$ on $\mathbb{R}^\Omega$.*

In the context of games, lexicographic subjective probability assessments on players' strategy profiles are closely related to test sequences used in the original definition of Selten (1975)'s perfect and Myerson (1978)'s proper equilibrium. Blume et al. (1991b, subsequently BBD2) provides a discussion of this issue. See also Appendix A.2 for additional details.

[3] See Appendix A.1 for the details.

2.3 Infinitely more likely events

I now define what it means for a cautiously rational decision maker to think that one event is infinitely more likely than another.[4]

Consider any cautiously rational decision maker with a finite state-space Ω and cautious preferences $\succeq$ over acts $\mathcal{P}^{\Omega}$. Fix any lexicographic representation $\rho = (\lambda_1, ..., \lambda_K)$ of the subjective probability assessment $\boldsymbol{P}[\succeq]$. For any nonempty event $S \subseteq \Omega$, define ρ_S as the beginning of ρ which covers S. I.e., $\rho_S = (\lambda_1, ..., \lambda_t)$, with $t = \min\left\{n \in \mathbb{N} : S \subseteq \bigcup_{i=1}^{t} supp\{\lambda_i\}\right\}$. Define the relation of "being infinitely more likely" as follows:

Definition (Infinitely More Likely)**.** *Given any cautious preferences $\succeq$ over acts $\mathcal{P}^{\Omega}$ and any two nonempty disjoint events A and B, event A is infinitely more likely than B, symbolically $A \gg_{\boldsymbol{P}[\succeq]} B$, if there exists some lexicographic representation ρ of $\boldsymbol{P}[\succeq]$ such that:*

$$supp\{\rho_A\} \cap B = \varnothing$$

In other words, for a given representation ρ, the event A is infinitely more likely than B, if the states in B start getting covered by ρ only after A is for the first time fully covered. It turns out that the relation $\gg_{\boldsymbol{P}[\succeq]}$ does not depend upon a particular representation of $P[\succeq]$. Thus, $\gg_{\boldsymbol{P}[\succeq]}$ is well defined. Also, $\gg_{\boldsymbol{P}[\succeq]}$ is an incomplete strict order on the set of nonempty events in Ω.

2.4 Complete Strategy Description Assumption (CSDA)

The strategies in a game must completely capture the the strategic situation. In particular, I require that for each player the following *Complete Strategy Description Assumption* (CSDA) holds:

Definition (Complete Strategy Description Assumption)**.** *In an N-player extensive-form game of perfect recall Γ, Complete Strategy Description Assumption holds for Player i if:*

1. *Player i is a decision maker who perceives his acts as subjective lotteries over simple objective lotteries over pure consequences;*
2. *Player i has in mind a finite subjective state-space Ω_i;*
3. *each state $\omega_i \in \Omega_i$ is labeled with a strategy profile of his opponents $\sigma^{-i} \in \mathcal{S}_{-i}$;*
4. *for each strategy profile of his opponents $\sigma^{-i} \in \mathcal{S}_{-i}$, there is at least one state $\omega_i \in \Omega_i$ labeled with σ^{-i};*
5. *each Player i's strategy $\sigma^i \in \mathcal{S}^{-i}$, considered as an act, induces from the point of view of Player i conditional on any state $\omega_i \in \Omega_i$ with label σ^{-i} precisely the objective lottery as specified in the rules of Γ for the play of $(\sigma^i; \sigma^{-i})$.*

[4] The notion of "infinitely more likely" considered here is different from the one discussed in BBD1, p. 69. However, the current version of "infinitely more likely" also admits a representation in terms of the original preferences $\succeq$. See Appendix A.3 for a more detailed exposition.

CSDA restricts how a player in a game may assess uncertainty. First, it states that a player must distinguish in his state-space between situations, in which his opponents play different strategies. Second, in his state-space he must consider all of the possible strategy profiles for his opponents. Third, from his subjective point of view, his play of strategy σ^i together with the play of his opponents σ^{-i} completely defines the resulting objective lottery.

CSDA in not an innocuous assumption. First, it assumes that each player is a BBD-type decision maker. I.e., the player's perception of his acts under uncertainty can be represented as a subjective lottery over simple objective lotteries over pure consequences. Second, under CSDA, the player views his own trembling mistakes as impossible. Third, if a game has objective chance nodes, under CSDA, the player does not distinguish states with different chance-node realizations. Instead, the player combines these realizations into states corresponding to his opponents' strategy profiles. This last requirement may sound overly restrictive. However, this requirement already seems to be implicitly present in the definition of chance nodes as *objective lotteries* with commonly known probability distributions *independent* of players' actions.

One implication of CSDA is that for a player, his subjective probability assessment on potentially rich state-space Ω_i may be equivalently represented as a subjective probability assessment directly on the set of his opponent's strategy profiles. I.e., there is a a subjective probability assessment on his opponent's strategy profiles, which induces the same preferences over his strategies. To obtain this equivalent assessment, one simply needs to merge subjective probabilities from the original assessment for states with the same label. This representation is standard in Bayesian game theory.

2.5 Interspecting types

Under interspection theory, each player has an *interspecting type*. For Player i, his interspecting type t_i is a tuple $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$. Here $k_i \geqslant 0$ is his *interspection level* (similar to level-k thinking); σ^i is the pure strategy he intends to play; $\Omega_i^{k_i}$ is his subjective state-space, called the *interspection state-space*; and $\boldsymbol{P}_i$ is his cautious assessment on the interspection state-space $\Omega_i^{k_i}$. The interspection state-space $\Omega_i^{k_i}$ encodes Player i's opponents' interspection levels and strategy profiles they intend to play, but not their assessments. Informally, Player i has interspection level k_i if he can analyze the play of his opponents up to $(k_i - 1)$-th level, but not further. The assessment $\boldsymbol{P}_i$ captures the probability assessment Player i may have on the combinations of his opponents' strategy profiles and their interspection levels *lower* than k_i. Also, if Player i has interspection level $k_i \geqslant 1$, then his strategy σ^i must be a best response to his assessment $\boldsymbol{P}_i$. Irrational types are types with level-0 interspection. They can play any strategy.

For each Player i with interspection level k_i, his interspection state-space $\Omega_i^{k_i}$ can be constructed as follows: Consider for a moment an intelligent outside advisor. The outside advisor knows for each Player j, with $j \neq i$, and for each interspection level $k_j < k_i$ the exact set of strategies $H_j^{k_j} \subseteq \mathcal{S}_j$ which Player j can play if he has interspection level k_j. For $k_j = 0$ the set H_j^0 is the set of all Player j's strategies $\mathcal{S}_j$. For $k \geqslant 1$ the set H_j^k may be any non-empty subset of $\mathcal{S}_j$. The outside advisor recommends to Player i to consider all of the *interspection events* $\mathcal{E}^{\boldsymbol{k}_{-i}}$, indexed by

a nonnegative multi-index $\boldsymbol{k}_{-i} = (k_1, k_2, .., k_{i-1}, ..., k_{i+1}, ..., k_N)$ with $\boldsymbol{k}_{-i} \leqslant \{k-1\}^{N-1}$.[5] The event $\mathcal{E}^{\boldsymbol{k}_{-i}}$ is the event that Player i's opponents have interspection levels corresponding to the multi-index $\boldsymbol{k}_{-i}$. Further, the outside advisor tells Player i that for each event $\mathcal{E}^{\boldsymbol{k}_{-i}}$, his opponents can play strategy profiles comprising precisely the set $E^{\boldsymbol{k}_{-i}} = H_1^{k_1} \times H_2^{k_2} \times ... \times H_{i-1}^{k_{i-1}} \times H_{i+1}^{k_{i+1}} \times ... \times H_N^{k_N}$. Following this advice, Player i constructs the interspection state-space $\Omega_i^{k_i}$ as the disjoint union of all such strategy profile sets $E^{\boldsymbol{k}_{-i}}$ (strategy profiles may have several copies in $\Omega_i^{k_i}$):

$$\Omega_i^{k_i} = \bigsqcup_{\boldsymbol{k}_{-i} \leqslant \{k-1\}^{N-1}} E^{\boldsymbol{k}_{-i}}$$

In the state-space $\Omega_i^{k_i}$, interspection events $\mathcal{E}^{\boldsymbol{k}_{-i}}$ are identified with the corresponding sets $E^{\boldsymbol{k}_{-i}}$.

In the actual game, Player i does not receive any recommendation from an outside advisor. Instead, Player i calculates the sets $E^{\boldsymbol{k}_{-i}}$ himself. The level of interspection then captures Player i's limiting abilities for this kind of calculations.

All possible interspecting types of a given game comprise the interspection type-space corresponding to this game:

Definition (Interspection Type-Space)**.** *The interspection type-space corresponding to a finite game is the set of all interspecting types possible in this game.*

Notably, for any finite game the corresponding interspection type-space is *unique*.

2.6 Respect

Another key concept in interspection theory is the concept of Respect. Informally, a player respects his opponents if he assesses that it is infinitely more likely that they have higher interspection levels than not. Yet, each player has only a finite interspection level and, therefore, cannot analyze opponents' interspection levels higher than his own. Formally:

Consider Player i with an interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$. The interspection state-space $\Omega_i^{k_i}$ is the disjoint union of interspection events $E^{\boldsymbol{k}_{-i}}$, for all nonnegative $\boldsymbol{k}_{-i} \leqslant \{k-1\}^{N-1}$.

Respect assumption requires that if some interspection event has higher interspection multi-index than another, then the former is infinitely more likely than the later:

Definition (Respect)**.** *In an N-player game Γ, an interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$ of Player i satisfies Respect if:*

$$\forall\ \boldsymbol{k}_{-i}, \hat{\boldsymbol{k}}_{-i} \leqslant \{k-1\}^{N-1} : \left(\boldsymbol{k}_{-i} > \hat{\boldsymbol{k}}_{-i}\right) \implies \left(E^{\boldsymbol{k}_{-i}} \gg_{\boldsymbol{P}_i} E^{\hat{\boldsymbol{k}}_{-i}}\right)$$

Player i then is said to respect his opponents.

[5] $\{a\}^n$ stands for the vector $\boldsymbol{a} = (a, ..., a) \in \mathbb{R}^n$. Vectors in $\mathbb{R}^n$ are compared componentwise: for two vectors $x, y \in \mathbb{R}^N$, $x \geqslant y$ if $\forall i \leqslant N$, $x_i \geqslant y_i$; and $x > y$ if $\forall i \leqslant N$, $x_i \geqslant y_i$ with at least one inequality being strict.

3 Equivalence Results

In this section I provide two equivalence results, which greatly simplify the construction of interspection theory. The first result states that under CSDA, the play of cautiously rational players in extensive-form games may be equivalently analyzed in the corresponding normal forms. More precisely, under CSDA, a player is initially rational against some cautious assessment if and only if he is sequentially rational against the same assessment. The second result is the equivalent reformulation of Respect as a property of assessments on the set of players' strategy profiles. This representation allows me to define interspection theory solution concepts in terms of primitives of the game, rather than in terms of the interspection type-space.

3.1 Sufficiency of normal-form analysis

Recall that under CSDA, any subjective assessment of a player may be equivalently represented as a subjective probability assessment on the set of his opponent's strategy profiles. It turns out, that under CSDA, the behavior of cautiously rational players in extensive-form games of perfect recall can be analyzed completely in the corresponding normal forms. I.e., if in an extensive-form game a player choses initially a best response strategy to some cautious assessment on his opponents' strategy profiles, then this strategy will be a sequential best response to the Bayesian updates of the same assessment in all of the strategy-relevant information sets. Formally:

Definition (Initial Rationality)**.** *In an extensive-form game of perfect recall* Γ*, Player* i *is initially rational if:*

1. *he has a cautious probability assessment* $\boldsymbol{P}_i$ *on the set of his opponents' strategy profiles* $\mathcal{S}_{-i}$*;*
2. *he plays a normal-form best response strategy to* $\boldsymbol{P}_i$*.*

In that case Player i *is said to be initially rational against* $\boldsymbol{P}_i$*.*

Assume for simplicity that the game has no chance nodes. Suppose that at the beginning of the game Player i has some cautious subjective probability assessment $\boldsymbol{P}_i$ on his opponents' strategy profiles $\mathcal{S}_{-i}$. If some of Player i's information sets h is reached during the play, then Player i will be able to logically exclude strategy profiles of his opponents that are not h-relevant. Denote their remaining strategy profiles through $\mathcal{S}^h_{-i}$. Let the *updated assessment* $\boldsymbol{P}^h_i$ be the assessment on $\mathcal{S}^h_{-i}$ obtained from $\boldsymbol{P}_i$ using Bayes rule. I.e., $\boldsymbol{P}^h$ is $\boldsymbol{P}$ conditioned on the event $\mathcal{S}^h_{-i}$. Sequential rationality is defined as cautious rationality against the updated assessment in each of the strategy-relevant information sets:

Definition (Sequential Rationality)**.** *In an extensive-form game of perfect recall* Γ*, Player* i *is sequentially rational if:*

1. *at the beginning of the game he has a cautious assessment* $\boldsymbol{P}_i$ *on the set of his opponents's strategy profiles* $\mathcal{S}_{-i}$*;*

2. *he plays a strategy that is a best response to the updated assessment* $\boldsymbol{P}_i^h$ *in each of his strategy-relevant information sets* h.

In that case Player i *is said to be sequentially rational against* $\boldsymbol{P}_i$.

This notion of sequential rationality is similar to *weak sequential rationality* of Reny (1992). The main difference is that Reny (1992)'s version is defined for usual-type rational players.

Naturally, if a player is sequentially rational against some cautious assessment, then he is initially rational against the same assessment. The following theorem establishes the equivalence of initial rationality and sequential rationality for cautiously rational players under CSDA. Thus, in interspection theory, it is sufficient to analyze games in their normal forms:

Theorem 1 (Normal-Form Sufficiency)**.** *Under CSDA, in an extensive-form game of perfect recall, a cautiously rational player is initially rational against a cautious probability assessment if and only if he is sequentially rational against the same assessment.*

Proof. See Appendix B.1 □

Theorem 1 holds for games with chance nodes as well: in that case one has to include into Bayesian updating elimination of logically impossible combinations of chance-node realizations and opponents' strategy profiles.

3.2 Respect in strategy form

The purpose of this paper is to provide a framework for analyzing games. The notion of Respect, however, is formulated for assessments on interspection state-spaces. I now provide an equivalent reformulation of Respect as a property of assessments on players' strategy profiles.

Definition (Hierarchy)**.** *A hierarchy* $\mathcal{H}$ *of order* $k \geqslant 1$ *on some nonempty finite set* S *is a non-decreasing sequence* $H^{k-1}, H^{k-2}, ..., H^0$ *of subsets of* S*, with the last element being equal to* S*:*

$$\mathcal{H} = \left\{H^i\right\}_{i=k-1}^{0}, \text{ where } H^{k-1} \subseteq H^{k-2} \subseteq ... \subseteq H^0 = S$$

The sets H^i in hierarchy $\mathcal{H}$ are termed *atoms*. For each atom H^i the number i is called *the index of atom* H^i in $\mathcal{H}$. Atoms in a hierarchy are allowed to be equal.

Definition (Product Hierarchy)**.** *Let* Γ *be a finite* N*-player extensive-form game of perfect recall. Suppose that for each Player* j*, with* $j \neq i$*, there is a given hierarchy* $\mathcal{H}_j$ *of order* $k_i \geqslant 1$ *on the set of Player* j*'s strategies* $\mathcal{S}_j$*. Then, the corresponding product hierarchy* $\mathcal{H}_{-i}$ *of order* k_i *is the collection of* $\mathcal{H}_j$*, for* $j \neq i$*:*

$$\mathcal{H}_{-i} = \left\{\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_{i-1}, \mathcal{H}_{i+1}, ..., \mathcal{H}_N\right\}$$

Suppose that some product hierarchy $\mathcal{H}_{-i}$ of order k_i is fixed for Player i. For each Player i's opponents' strategy profile $\sigma^{-i} \in \mathcal{S}_{-i}$, I define the corresponding *hierarchical index* as the highest multi-index of atoms in $\mathcal{H}_{-i}$ whose product contains σ^{-i}. Formally:

Definition (Hierarchical Index)**.** *Let Γ be a finite N-player extensive-form game of perfect recall. Let $\mathcal{H}_{-i}$ be a product hierarchy of order $k_i \geqslant 1$ constructed on the set $\mathcal{S}_{-i}$ of Player i's opponents' strategy profiles. For $\sigma^{-i} \in \mathcal{S}_{-i}$, the hierarchical index $Ind_{\mathcal{H}_{-i}}(\sigma^{-i})$ with respect to $\mathcal{H}_{-i}$ is defined as:*

$$Ind_{\mathcal{H}_{-i}}(\sigma^{-i}) = \max\left\{\boldsymbol{k}_{-i} = \{k_1, ..., k_{i-1}, k_{i+1}, ..., k_N\} \leqslant \{k_i - 1\}^{N-1} : \ \sigma^{-i} \in \bigtimes_{j \neq i} H_j^{k_j}\right\}$$

Given a product hierarchy $\mathcal{H}_{-i}$ and some cautious subjective probability assessment $\boldsymbol{P}$ on $\mathcal{S}_{-i}$, the assessment $\boldsymbol{P}$ conforms to the hierarchy $\mathcal{H}_{-i}$ if under $\boldsymbol{P}$, strategy profiles with higher hierarchical indices are infinitely more likely than are strategy profiles with lower indices. Formally:

Definition (Conforming to Hierarchy)**.** *Let Γ be a finite N-player extensive-form game of perfect recall. Let $\mathcal{H}_{-i}$ be a product hierarchy of order $k_i \geqslant 1$ constructed on the set $\mathcal{S}_{-i}$ of Player i's opponents' strategy profiles. Let $\boldsymbol{P}$ be a cautious lexicographic probability assessment on the state-space $\mathcal{S}_{-i}$. The assessment $\boldsymbol{P}$ is said to conform to $\mathcal{H}_{-i}$ if:*

$$\forall \sigma^{-i}, \hat{\sigma}^{-i} \in \mathcal{S}_{-i} : \ \left(Ind_{\mathcal{H}_{-i}}(\sigma^{-i}) > Ind_{\mathcal{H}_{-i}}(\hat{\sigma}^{-i})\right) \rightarrow \left(\{\sigma^{-i}\} \gg_{\boldsymbol{P}} \{\hat{\sigma}^{-i}\}\right)$$

Recall the definition of Player i's interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$. The construction of the corresponding interspection state-space $\Omega_i^{k_i}$ was given based on the recommendations of an outside advisor. In particular, the outside advisor was asked to name all of the sets $H_j^{k_j} \subseteq \mathcal{S}_j$ for all $j \neq i$ and $0 \leqslant k_j < k_i$, i.e., the sets of strategies that can be played by a level-k_j interspecting Player j. Suppose that those sets $H_j^{k_j}$ were coming from some product hierarchy $\mathcal{H}_{-i}$ of order k_i constructed on $\mathcal{S}_{-i}$. In that case, the interspecting type t_i is said to be *constructed upon the hierarchy* $\mathcal{H}_{-i}$.

Consider two different tasks. Task 1 is to predict the behavior of a player who has an interspecting type constructed upon some product hierarchy, and who respects his opponents. Task 2 is to predict the behavior of this player, provided he is cautiously rational and has a cautious assessment on his opponents' strategy profiles conforming to the same product hierarchy.

The following theorem, effectively, establishes the equivalence of Task 1 and Task 2:

Theorem 2 (Respect in Strategy Form)**.** *Let Γ be a finite N-player extensive-form game of perfect recall. Let $\mathcal{H}_{-i}$ be a product hierarchy of order $k_i \geqslant 1$ constructed on the set $\mathcal{S}_{-i}$ of Player i's opponents' strategy profiles. Then:*

1. *for any interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$ constructed upon $\mathcal{H}_{-i}$ and satisfying Respect, there exists a cautious probability assessment $\tilde{\boldsymbol{P}}_i$ on $\mathcal{S}_{-i}$ conforming to $\mathcal{H}_{-i}$ such that $\boldsymbol{P}_i$ and $\tilde{\boldsymbol{P}}_i$ induce the same preference relation on $\mathcal{S}_i$;*

2. *for any cautious probability assessment $\tilde{\boldsymbol{P}}_i$ on $\mathcal{S}_{-i}$ conforming to $\mathcal{H}_{-i}$, there exists an interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$ constructed upon $\mathcal{H}_{-i}$ and satisfying Respect such that $\boldsymbol{P}_i$ and $\tilde{\boldsymbol{P}}_i$ induce the same preference relation on $\mathcal{S}_i$.*

Proof. See Appendix B.2 □

Theorem 2 allows one to switch from Task 1 to Task 2 in the construction of interspection theory. Using this equivalence, in the remainder of the paper, I work directly with assessments on strategy profiles that conform to product hierarchies, rather than with interspecting types satisfying Respect.

4 Interspected Rationalizability

In the previous section, I established two equivalences in interspection theory: the equivalence between normal-form and extensive-form analyses, and the equivalent representation of Respect in terms of primitives of the game. Using these equivalences, I proceed by working directly with normal forms and players' assessments on sets of their opponents' strategy profiles. I now introduce two solution concepts in interspection theory: interspected rationalizability and interspected equilibrium. The current section deals with interspected rationalizability. The next section discusses interspected equilibrium.

In this section, I first give a constructive definition of the notion of interspected rationalizability. I then provide an axiomatic characterization of "environments" to which this notion applies. Finally, I compare interspected rationalizability with two familiar solution concepts: iterative admissibility and backward induction.

4.1 Interspected rationalizability: construction

In this subsection I provide a constructive definition of interspected rationalizability.

Consider a finite N-player extensive-form game of perfect recall Γ. The corresponding normal form is $G(\Gamma)$.

Assume CSDA holds. Then players' assessments may be equivalently represented as subjective probability assessments on the set of their opponents' strategy profiles. For any Player i's cautious assessment $\boldsymbol{P}_i$ on the state-space $\mathcal{S}_{-i}$, denote by $B(\boldsymbol{P}_i)$ the set of Player i's pure strategies that are best responses to $\boldsymbol{P}_i$.

Recall the definition of product hierarchy from subsection 3.2. Let $\mathcal{H}_{-i}$ be a product hierarchy $\mathcal{H}_{-i}$ of order k_i on the set $\mathcal{S}_{-i}$. Denote through $\mathcal{A}(\mathcal{H}_{-i})$ the space of all cautious assessments on $\mathcal{S}_i$ conforming to the hierarchy $\mathcal{H}_{-i}$. Player i's strategy is said to be an *$\mathcal{H}_{-i}$-never-best response* if it is not a best response to any cautious assessment conforming to $\mathcal{H}_{-i}$. Formally:

Definition ($\mathcal{H}$-Never-Best Response)**.** *Given a product hierarchy $\mathcal{H}_{-i}$ on $\mathcal{S}_{-i}$, a strategy $\sigma^i \in \mathcal{S}_i$ is said to be an $\mathcal{H}_{-i}$-never-best response if:*

$$\forall\ \boldsymbol{P}_i \in \mathcal{A}(\mathcal{H}_{-i}) : \quad \sigma^i \notin B(\boldsymbol{P}_i)$$

Suppose it is known to us that Player i is cautiously rational and that his cautious assessment conforms to some product hierarchy $\mathcal{H}_{-i}$ of order k_i on $\mathcal{S}_{-i}$. Then, based on this information

alone, we can only conclude that Player i will not play a strategy that is $\mathcal{H}_{-i}$-never-best response. In other words, we can only predict that he will play some strategy from $\mathcal{H}_{-i}$-*best response set* $B(\mathcal{H}_{-i})$, defined as follows:

Definition ($\mathcal{H}$-Best-Response Set)**.** *Given a product hierarchy $\mathcal{H}_{-i}$ on $\mathcal{S}_{-i}$, the $\mathcal{H}_{-i}$-best-response set $B(\mathcal{H}_{-i})$ is defined as the complement in $\mathcal{S}_i$ of the set of $\mathcal{H}_{-i}$-never-best response strategies.*

As $G(\Gamma)$ is finite, $B(\mathcal{H}_{-i})$ is nonempty for any hierarchy $\mathcal{H}_{-i}$. One can show that in a finite game, the set $B(\mathcal{H}_{-i})$ can be computed in a finite algorithm. Thus, $B(\mathcal{H}_{-i})$ is defined *constructively.*

The sequence of interspected hierarchies is defined as follows:

Definition (Interspected Hierarchies)**.** *In a finite N-player extensive-form game of perfect recall Γ, the sequences of interspected hierarchies $\{\mathcal{H}_i^k\}_{k\geqslant 0}$ for $i = 1, ..., N$, are defined recursively in k and simultaneously for all players as follows:*

1. $\mathcal{H}_i^0 = \{\mathcal{S}_i\}$, *for* $i = 1, ..., N$;
2. $\mathcal{H}_i^{k+1} = \{B(\mathcal{H}_{-i}^k); \mathcal{H}_i^k\}$ *for* $i = 1, ..., N$ *and* $k \geqslant 0$.

For each $k \geqslant 0$, the set of strategy profiles $H^k = \underset{i=1}{\overset{N}{\times}} H_i^k$ consisting of the first atoms in the k-th step interspected hierarchies $\mathcal{H}^k = \left(\mathcal{H}_1^k, ..., \mathcal{H}_N^k\right)$ is called the set of *k-th step interspected strategies*, or the set of *k-th step predicted strategies.*

By induction, $H^{k+1} \subseteq H^k$, for each $k \geqslant 0$.[6] Also, if $H^{m+1} = H^m$ for some $m > 0$, then $H^{k+1} = H^k$ for all $k \geqslant m$. I.e., the sequence of the sets of k-th step predicted strategies $\{H^k\}_{k\geqslant 0}$ is weakly decreasing. Moreover, if this sequence stabilizes for the first time, then it is constant ever after. As the game Γ is finite, there exists some finite $k \geqslant 0$ such that $H^{k+1} = H^k$. The smallest such k is called *the interspection index* of the game Γ:

Definition (Interspection Index)**.** *For a finite extensive-form game of perfect recall Γ, the interspection index $Ind(\Gamma)$ is the first number k at which the sequence of the sets of k-th step interspected strategies stabilizes:*

$$Ind(\Gamma) = \min_{k\geqslant 0} \left(H^{k+1}(\Gamma) = H^k(\Gamma)\right)$$

The set of *interspected rationalizable strategies* then is defined as the set of strategies predicted at the moment the construction of the interspected hierarchies stabilizes:

Definition (Interspected Rationalizability)**.** *For a finite extensive-form game of perfect recall Γ with interspection index $Ind(\Gamma)$, the set of interspected rationalizable strategies is defined as the set of $Ind(\Gamma)$-th step interspected strategies:*

$$R(\Gamma) = H^{Ind(\Gamma)}(\Gamma)$$

[6] Indeed, assuming that atoms in $\mathcal{H}^k$ contain atoms from $\mathcal{H}^{k-1}$, i.e. $\mathcal{H}^{k-1} \subseteq \mathcal{H}^k$, we get that any assessment that conforms to $\mathcal{H}_{-i}^k$ also conforms to $\mathcal{H}_{-i}^{k-1}$. Therefore, $B(\mathcal{H}_{-i}^k) \subseteq B(\mathcal{H}_{-i}^{k-1})$, and so $\mathcal{H}^k \subseteq \mathcal{H}^{k+1}$.

The strategies that are not interspected rationalizable are called *interspected non-rationalizable.*

Fully interspected hierarchies are the hierarchies obtained at the moment the construction of the interspected hierarchies stabilizes:

Definition (Fully Interspected Hierarchies)**.** *For a finite extensive-form game of perfect recall* Γ *with interspection index* $Ind(\Gamma)$*, fully interspected hierarchies* $\mathcal{H}^\infty$ *are identified with the* $Ind(\Gamma)$*-th step interspected hierarchies:*

$$\mathcal{H}^\infty = \mathcal{H}^{Ind(\Gamma)}$$

Strictly speaking, the construction of interspected hierarchies never stabilizes. However, in each step after the $Ind(\Gamma)$-th step, only a copy of the set of interspected rationalizable strategies is attached to the front of the previous-step hierarchies. Such attachment is inconsequential. Thus, the hierarchies after the $Ind(\Gamma)$-th step may be identified with fully interspected hierarchies $\mathcal{H}^\infty$.

4.2 Interspected rationalizability: axiomatic characterization

In the previous subsection, I gave a constructive definition of interspected rationalizability. It is natural to ask when we expect this notion to apply. In this subsection I provide axiomatic characterization of *environments* in which players' behavior conforms to interspected rationalizability. The axioms are as follows:

Axiom 1 (Common Knowledge of the Game Form)**.** *The set of normal-form strategies of the game is common knowledge.*

Axiom 2 (Common Knowledge of the Payoffs)**.** *The payoffs in the normal form are common knowledge.*

Axiom 3 (Cautious Rationality)**.** *Each player is cautiously rational.*

Axiom 4 (Sufficient Interspection)**.** *Each player has an interspecting type with interspection level at least as high as the interspection index of the game.*

Axiom 5 (Common Knowledge of Impeccability)**.** *It is common knowledge that each action in the game is intentional.*

Axiom 6 (Common Knowledge of Respect)**.** *It is common knowledge that the players respect their opponents.*

Common Knowledge of the Game Form and of the Payoffs imply that the normal form of the game is common knowledge. Sufficient Interspection ensures that each player can fully grasp the game. Impeccability is the logical negation of Selten's type trembling-hand mistakes. Common Knowledge of Respect is a substitute for Common Knowledge of Rationality. If trembling-hand mistakes are commonly viewed as impossible, deviations can be explained by a breakdown of Common Knowledge of Rationality. One possible way to restrict players' interpretations of these breakdowns is to impose Common Knowledge of Respect.

I can now state and prove the main result of this section:

Theorem 3 (Axiomatic Characterization)**.** *Given a finite extensive-form game of perfect recall* Γ*, if Axioms* 1-6 *are satisfied, then the set of strategies the players can be expected to play is precisely the set of interspected rationalizable strategies* $R(\Gamma)$*.*

Proof. See Appendix B.3 □

Thus, interspected rationalizability is a solution concept for environments satisfying Axioms 1-6. Other solution concepts in interspection theory can be obtained as refinements of interspected rationalizability, provided there is some *secondary* or *refining* principle in addition to Axioms 1-6.

Consider, for example, a secondary principle that the played strategies should be consistent with players' assessments about each other. This is the *refining principle of equilibrium.* Suppose some tentative equilibrium is believed in by the players. The players then may also contemplate the possibility that their opponents do not conform to the equilibrium. If the principle of equilibrium is secondary, then at any point each player believes that it is infinitely less likely that his opponents are not sufficiently interspecting, rather than the fact that they do not play according to the equilibrium. In other words, the hypothesis that their opponents are playing according to the equilibrium is the first hypothesis the players would reject if they observe a deviation.

Another notable refinement principle is Harsanyi's doctrine: the assumption that assessments of the players come from the common prior.

4.3 Interspected rationalizability and iterative admissibility

Brandenburger et al. (2008) provided epistemic conditions under which players can be expected to play strategies surviving iterative elimination of weakly dominated strategies in rounds (iterative admissibility). I now compare interspected rationalizability and iterative admissibility.

For finite two-player games these two concepts coincide:

Theorem 4 (Iterative Admissibility)**.** *For a finite two-player extensive-form game of perfect recall* Γ*, the set of interspected rationalizable strategies coincides with the set of iteratively admissible strategies. Moreover, the construction of fully interspected hierarchies* $(\mathcal{H}_1^\infty, \mathcal{H}_2^\infty)$ *corresponds to the rounds of iterative admissibility. I.e., for each Player* i*,* $i = 1, 2$*, the set* H_i^j *is precisely the set of Player* i*'s strategies surviving* j *rounds of elimination of weakly dominated strategies.*

Proof. See Appendix B.4 □

Remarkably, for games with more than two players, interspected rationalizability does not coincide with iterative admissibility:

Consider, for example, the three-player game Γ depicted in Figure 2.[7] At the beginning of the game, Players 1 and 2 each choose independently and simultaneously whether to move "left" or "right." If both of them move "right" the game ends. If at least one of them choses "left" the turn is passed to Player 3. He then select either "left" or "right" himself. Yet Player 3 does not

[7] The extensive form of Γ is borrowed from Kreps and Ramey (1987).

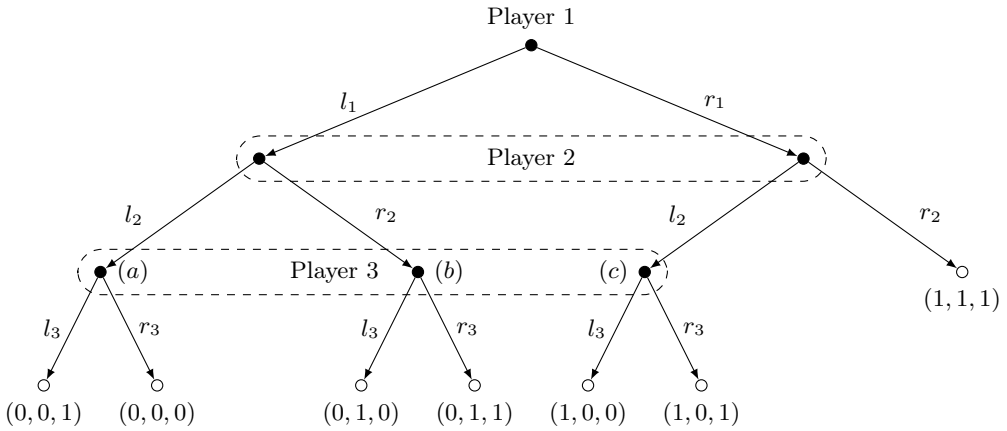


Figure 2: Interspected rationalizability ≠ iterative admissibility.

observe which of his opponents has chosen "left". After Player 3's move the game ends. The payoffs are such that for both Player 1 and Player 2, choosing "right" strictly dominates choosing "left." Player 3, however, does not have a dominated strategy. Thus, the set $IA(\Gamma)$ of iteratively admissible strategies in this game is:

$$IA(\Gamma) = \{r_1\} \times \{r_2\} \times \{l_3, r_3\}$$

What about interspected rationalizability? For Players 1 and 2, the result is still the same, as dominated strategies cannot be cautious best responses. Consider the situation from the point of view of Player 3. Upon getting a chance to move, Player 3 knows that at least one (and possibly both) of his opponents (and possibly both) are not rational. However, if Player 3 respects his opponents, he must conclude that it still is infinitely more likely that only one of them is irrational, rather than that both are irrational. Then, Player 3 would think that the node (a) is infinitely unlikely. His unique best response then will be to choose "right." Thus, the set of interspected rationalizable strategies is:

$$R(\Gamma) = \{r_1\} \times \{r_2\} \times \{r_3\} \neq IA(\Gamma)$$

In other words, under interspection theory, players think respectfully about their opponents even after observing deviations. Under the theory of iterative admissibility, players may form arbitrary assessments of deviators' behavior.

4.4 Interspected rationalizability and backward induction in PI games

Finally, I discuss the properties of interspected rationalizability in perfect information games (PI games). The result is that for any PI game without chance nodes and without relevant ties, the unique terminal node reachable by playing interspected rationalizable strategies is the unique backward induction outcome of the game.

What is the logic of backward induction in PI games? Following Arieli and Aumann (2013), for any PI extensive-form game Γ *without chance nodes*, each decision node h is labeled in inductive labeling procedure by a set Z_h of terminal outcomes. The labeling procedure starts by labeling all terminal nodes with their own outcomes. Then, the procedure gradually moves towards the root of the tree, from sons to their parenting nodes. The precise procedure is the following. Suppose that for some node h, the labels have been attached to all of his sons. For each of his sons s with the label Z_s, define $max\, s$ and $min\, s$ as the maximum and minimum payoffs to the father h from the outcomes in Z_s. Call a son node s inferior if it has a brother s' with $min\, s' > max\, s$. The label of the father h then is the union of the labels of all of his non-inferior sons. The set of *backward induction outcomes* $BI(\Gamma)$ of the game Γ is the label attached to its root.

Following Battigalli (1997), a PI game *without chance nodes* is said to be *without relevant ties* if for any two terminal nodes, the player moving at the last common predecessor of these nodes is not indifferent between them. In particular, PI games with generic payoffs have no relevant ties.

If a PI game without chance nodes has no relevant ties, then its backward induction outcome is unique. Interspected rationalizability predicts the same outcome:

Theorem 5 (Backward Induction for PI Games)**.** *For any perfect information game without chance nodes and without relevant ties, the unique interspected rationalizable terminal outcome is the unique backward induction outcome.*

Proof. See Appendix B.5. □

The proof of Theorem 5 is an adaptation of the proofs of Reny (1992)'s Proposition 3 and Battigalli (1997)'s Theorem 4. The proof relies on the properties of Kohlberg and Mertens (1986)'s fully stable sets.

The question of how interspected rationalizability compares with backward induction for non-generic PI games remains open. Also, with regards to equilibria, subsection 6.3 contains an example of a non-generic PI game that has an interspected equilibrium with the path that cannot be supported by any subgame perfect equilibrium.

5 Interspected Equilibrium

In the previous section, I studied the notion of interspected rationalizability. In this section, I introduce the notion of interspected equilibrium, another solution concept in interspection theory. Interspected equilibrium may be viewed as a refinement of interspected rationalizability. Interspected equilibria correspond to situations in which, before observing any deviation, players believe that their opponents follow the equilibrium strategies. Moreover, after observing deviations, players believe that their opponents play strategies corresponding to the highest interspection levels that are still logically possible for them.

I start by considering the simpler and more tractable case of two-player games. I then move to the general case of N-player games. Finally, I provide an interpretation of the interspected

equilibrium concept and contrast it with Selten (1975)'s trembling-hand perfect equilibrium.

5.1 Interspected equilibrium in two-player games

Consider a two-player extensive-form game of perfect recall Γ. The players are Player 1 and Player 2. Denote through (μ^1, μ^2) a pair of randomized strategies for Player 1 and Player 2, and through $(\boldsymbol{P}_1, \boldsymbol{P}_2)$ their cautious assessments on $\mathcal{S}_2$ and $\mathcal{S}_1$ correspondingly. The following is an adapted version of the definition from BBD2:

Definition (Lexicographic Nash equilibrium, $N = 2$)**.** *In a two-player normal-form game G, a lexicographic Nash equilibrium $(\mu^1, \mu^2; \boldsymbol{P}_1, \boldsymbol{P}_2)$ is a pair of players' randomized strategies (μ^1, μ^2) and a pair of lexicographic assessments $(\boldsymbol{P}_1, \boldsymbol{P}_2)$ on $(\mathcal{S}_2, \mathcal{S}_1)$ such that:*

1. *each pure strategy in the support of μ^i is a best response to $\boldsymbol{P}_i$, for $i = 1, 2$;*
2. *the first measure λ_1^i in any LPS representation $\rho_i = (\lambda_1^i, ..., \lambda_{K_i}^i)$ of $\boldsymbol{P}_i$ is proportional to the opponents' randomized strategy:*

$$\mu^{-i} = \alpha_i \cdot \lambda_1^i, \text{ for some } \alpha_i > 0 \text{ and for } i = 1, 2$$

Recall the construction of fully interspected hierarchies $(\mathcal{H}_1^\infty, \mathcal{H}_2^\infty)$. Interspected equilibrium is defined as a lexicographic Nash equilibrium with cautious assessments conforming to fully interspected hierarchies:

Definition (Interspected Equilibrium, $N = 2$)**.** *In a finite two-player extensive-form game of perfect recall Γ, an interspected equilibrium $(\mu^1, \mu^2; \boldsymbol{P}_1, \boldsymbol{P}_2)$ is a pair of randomized strategies (μ^1, μ^2) and a pair of cautious assessments $(\boldsymbol{P}_1, \boldsymbol{P}_2)$ on $(\mathcal{S}_2, \mathcal{S}_1)$ such that:*

1. *$(\mu^1, \mu^2; \boldsymbol{P}_1, \boldsymbol{P}_2)$ is a lexicographic Nash equilibrium of $G(\Gamma)$;*
2. *assessments $(\boldsymbol{P}_1, \boldsymbol{P}_2)$ conform to fully interspected hierarchies $(\mathcal{H}_2^\infty, \mathcal{H}_1^\infty)$.*

Interspected equilibria correspond to situations in which, before observing any deviation, each player believes that his opponent follows the equilibrium strategy. Moreover, after observing a deviation, a player believe that his opponent plays a strategy corresponding to the highest interspection level that is still logically possible for the opponent.

Note that in the definition of lexicographic Nash equilibrium assessments are not required to be cautious. I.e., they do not have to have the full support. Yet, for the interspected equilibrium the full support of the assessments is a requirement.

For any finite two-player game at least one interspected equilibrium exists:

Theorem 6 (Equilibrium Existence, $N = 2$)**.** *For any finite two-player extensive-form game of perfect recall there exists at least one interspected equilibrium.*

Proof. See Appendix B.6 □

The proof of Theorem 6 is similar to the proof of the necessity part for LPS representation of proper equilibrium of BBD2 Proposition 5, p.88. Essentially, one needs to replace "respect preferences" by "conforms to fully interspected hierarchies". In turn, proper equilibrium may be defined as an equilibrium, with assessments conforming to hierarchies constructed by the expected payoffs against the equilibrium play.

For the existence of interspected equilibrium it is important that fully interspected hierarchies are constructed by iteratively dismissing never-best response strategies. Lexicographic Nash equilibria with assessments conforming to arbitrary hierarchies do not in general exist.

As any interspected equilibrium involves only the play of interspected rationalizable strategies,[8] the immediate corollary of Theorem 6 is:

Corollary 1. *If a game has the unique interspected rationalizable outcome, then this outcome is the unique outcome supportable in any interspected equilibrium.*

Next, I want to discuss the properties of the interspected equilibrium concept.

Each equilibrium induces some distribution on the payoff outcomes. If some equilibrium concept generates the same set of payoff distributions for equivalent games, this concept is called invariant:

Definition (Invariance)**.** *An equilibrium concept satisfies invariance if for any two extensive-form games with the same reduced normal form, the distribution on payoff outcomes induced by an equilibrium of the first game is the distribution on payoff outcomes induced by some equilibrium of the second game.*

Invariance is defined with respect to the same Kohlberg and Mertens (1986)'s reduced normal form, not normal form. I.e., this notion of invariance coincides with the one from Kohlberg and Mertens (1986).

I define the forward induction property for equilibrium concepts as follows:

Definition (Forward Induction)**.** *An equilibrium concept satisfies forward induction if an equilibrium of the full game corresponds to an equilibrium of the reduced game after a one-round elimination of all never-best response (weakly dominated) strategies of one of the players.*

In contrast, Kohlberg and Mertens (1986)'s notion of forward induction (or iterated dominance) requires that an equilibrium of the full game remains an equilibrium of the reduced game after elimination of a single never-best response strategy, not all of them. Also, Kohlberg and Mertens (1986) considers a second type of iterated dominance: with respect to elimination of strategies that are not best responses to the considered equilibrium. For a discussion of this second type of iterated dominance see Govindan and Wilson (2009) and BBD2, p. 89.

[8]The fact that the strategies played in interspected equilibrium are interspected rationalizable is directly built into the definition of interspected equilibrium.

Recall that in interspection theory, sequential rationality is cautious rationality in all of the strategy-relevant information sets. Sequential rationality for equilibrium concepts is defined as follows:

Definition (Sequential Rationality)**.** *An equilibrium concept satisfies sequential rationality if in any equilibrium a player following the prescribed strategy is sequentially rational against the assessment specified in the equilibrium.*

Admissibility for equilibrium concepts is defined as usually:

Definition (Admissibility)**.** *An equilibrium concept satisfies admissibility if any equilibrium does not involve the play of weakly dominated strategies.*

The following proposition summarizes the main properties of the interspected equilibrium concept in two-player games:

Theorem 7 (Equilibrium Properties, $N = 2$)**.** *For finite two-player extensive-form games of perfect recall, the concept of interspected equilibrium satisfies sequential rationality, admissibility, invariance, and forward induction.*

Proof. See Appendix B.7 □

Notably, interspected equilibrium does not in general satisfy subgame perfection. Section 6 contains examples.

5.2 Interspected equilibrium in general N-player games

Having considered the simpler case of two-player games, I now introduce the notion of interspected equilibrium for the general case of N-player games. In two-player games interspected equilibrium was introduced as a pair of players' strategies and their subjective assessments. In the general case of N players, interspected equilibrium is defined as a pair of players' strategies and a *common prior* cautious assessment on the set of strategy profiles of *all* of the players at once.

To capture independence of players' actions in interspected equilibrium, one needs to specify a notion of independence for lexicographic assessments. BBD1, p. 73 provides a discussion of this issue. Here, I briefly review some of their points.

There may be several ways to define independence for probability assessments on product spaces. I will focus on two of them: stochastic independence and strong independence. For classical probability measures these two notions are equivalent. However, for lexicographic assessments stochastic independence is strictly weaker than strong independence.

Suppose a state-space Ω is a direct product of N finite spaces $\Omega = \bigtimes_{i=1}^{N} \Omega_i$. Let $\boldsymbol{P}$ be a cautious probability assessment on Ω. Stochastic independence is defined as follows:

Definition (Stochastic Independence)**.** *A cautious probability assessment $\boldsymbol{P}$ on a finite product space $\Omega = \bigtimes_{i=1}^{N} \Omega_i$ satisfies stochastic independence, if for all $i = 1, ..., N$ conditional assessments $marg_i \, \boldsymbol{P} = \boldsymbol{P}_{\{\omega_i, *\}}$ (i.e. marginals) do not depend on $\omega_i \in \Omega_i$.*

Following BBD2, a *nested convex combination* $r \square \rho$ is defined for any LPS $\rho = (\lambda_1, \lambda_2, ..., \lambda_K)$ and a vector $r \in (0,1)^{K-1}$ as the following classical measure:

$$r \square \rho = (1 - r_1)\lambda_1 + r_1\big[(1 - r_2)\lambda_2 + r_2[(1 - r_3)\lambda_3 + ... + r_{K-1}\lambda_k]...\big]$$

Limiting nested convex combination $\big\{r(n) \square \rho\big\}_{n \in \mathbb{N}}$ is a sequence of nested convex combinations with $r(n) \to 0$, as $n \to \infty$.

Strong independence is defined as follows:

Definition (Strong Independence)**.** *A cautious probability assessment* $\boldsymbol{P}$ *on a finite product space* $\Omega = \underset{i=1}{\overset{N}{\times}} \Omega_i$ *satisfies strong independence if* $\boldsymbol{P}$ *has an LPS representation* ρ *for which there exists a limiting nested convex combination* $\big\{r(n) \square \rho\big\}_{n \in \mathbb{N}}$ *with* $r(n) \to 0$ *such that for each* $n \in \mathbb{N}$ *the measure* $r(n) \square \rho$ *is a product measure on* Ω *in the classical sense.*

I now define the concept of interspected equilibrium for general N-players games:

Definition (Interspected Equilibrium)**.** *In a finite* N*-player extensive-form game of perfect recall* Γ*, an interspected equilibrium* $(\mu; \boldsymbol{P})$ *is a pair of independently randomized strategies* $\mu = \underset{i=1}{\overset{N}{\times}} \mu^i$ *and a common prior cautious assessment* $\boldsymbol{P}$ *on the set of players' strategy profiles* $\mathcal{S} = \underset{i=1}{\overset{N}{\times}} \mathcal{S}_i$ *such that:*

1. *each pure strategy in the support of* μ^i *is a best response to the corresponding marginal* $marg_i$ $\boldsymbol{P}$*, for all* $i = 1, ..., N$*;*
2. *assessment* $\boldsymbol{P}$ *conforms to fully interspected hierarchies* $\mathcal{H}^\infty$*;*
3. *assessment* $\boldsymbol{P}$ *satisfies strong independence;*
4. *the first measure* λ_1 *in any LPS representation* $\rho = (\lambda_1, ..., \lambda_K)$ *of* $\boldsymbol{P}$ *is proportional to* μ*:*

$$\mu = \alpha \cdot \lambda_1, \text{ for some } \alpha > 0$$

Note that if a strongly independent cautious assessment $\boldsymbol{P}$ conforms to fully interspected hierarchies $\mathcal{H}^\infty$, then for all $i = 1, ..., N$, the marginal assessment $marg_i$ $\boldsymbol{P}$ conforms to $\mathcal{H}^\infty_{-i}$. Yet, the first measure in any LPS representation of a strongly independent assessment is a product measure. Thus, in any interspected equilibrium, players' subjective assessments conform to the corresponding fully interspected hierarchies. Also, the prescribed strategies are assessed as being independently randomized.

Note that for the case of $N = 2$ players, the current definition of interspected equilibrium and the definition used in the previous subsection are equivalent.

Interspected equilibria always exist in general finite games:

Theorem 8 (Equilibrium Existence)**.** *For any finite extensive-form game of perfect recall there exists at least one interspected equilibrium.*

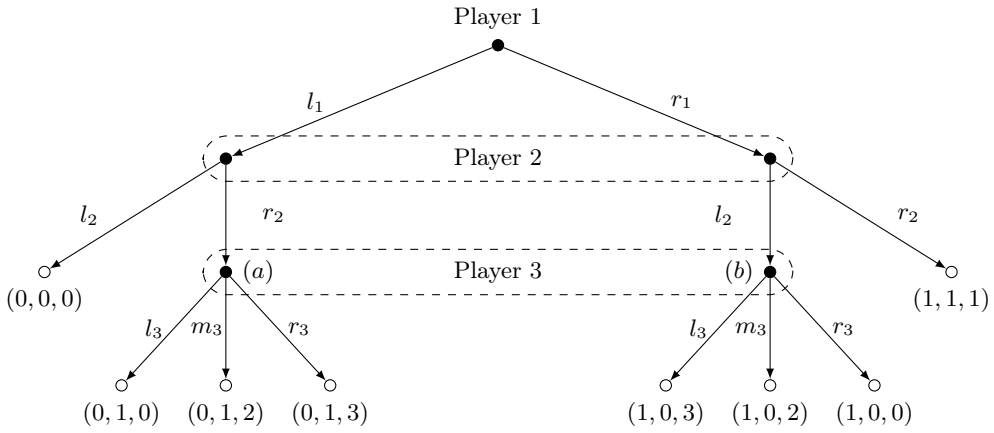


Figure 3: Interspected equilibrium does not satisfy strong version of forward induction.

Proof. See Appendix B.8 □

Recall the forward induction property for equilibrium concepts: an equilibrium of the full game must correspond to an equilibrium of the reduced game obtained after a one-round elimination of all never-best-response strategies of one of the players. This strong forward induction property can be weakened as follows:

Definition (Weak Forward Induction)**.** *An equilibrium concept satisfies weak forward induction if an equilibrium of the full game corresponds to an equilibrium of the reduced game after elimination of all interspected non-rationalizable strategies of all of the players.*

In finite N-player games, the concept of interspected equilibrium has the following properties:

Theorem 9 (Equilibrium Properties)**.** *In finite extensive-form games of perfect recall, the concept of interspected equilibrium satisfies sequential rationality, admissibility, invariance, and weak forward induction.*

Proof. See Appendix B.7. □

Note that in multi-player games, the concept of interspected equilibrium does not satisfy the strong version of forward induction. As an illustration, consider the three-player game Γ depicted in Figure 3. At the beginning of the game, Players 1 and 2 each choose independently and simultaneously whether to move "left" or "right." If both of them move into the same direction the game ends. If one of them choses "left" and the other choses "right" the turn is passed to Player 3. He then selects where to move: "left," "middle," or "right." Yet Player 3 does not observe which of his opponents has chosen "left." After Player 3's move the game ends. The payoffs are such that for both Player 1 and Player 2, choosing "right" strictly dominates "left." Thus, the strategies $[l_1]$ and $[l_2]$ do not survive the first round of interspection. Player 3, however, does not have a dominated strategy. The set of interspected rationalizable strategies is:

$$R(\Gamma) = \{r_1\} \times \{r_2\} \times \{l_3, m_3, r_3\}$$

Γ has an interspected equilibrium with the outcome $\{r_1\} \times \{r_2\} \times \{m_3\}$. Indeed, this outcome may be supported, for example, by the following common prior LPS assessment ρ:

$$\rho = \Big\{[r_1, r_2, m_3]; \quad \big([l_1, r_2, m_3] + [r_1, l_2, m_3] + [r_1, r_2, l_3] + [r_1, r_2, r_3]\big);$$
$$\big([l_1, l_2, m_3] + [r_1, l_2, l_3] + [r_1, l_2, r_3] + [l_1, r_2, l_3] + [l_1, r_2, r_3]\big); \quad \big([l_1, l_2, l_3] + [l_1, l_2, r_3]\big)\Big\}$$

Symbolically ρ can be written as:

$$\rho = \Big([r_1] + \epsilon \cdot [l_1]\Big) \times \Big([r_2] + \epsilon \cdot [l_2]\Big) \times \Big([m_1] + \epsilon \cdot [l_3] + \epsilon \cdot [r_3]\Big)$$

where ϵ stands for an infinitesimal. Clearly, ρ satisfies strong independence and conforms to fully interspected hierarchies. If the turn is passed to player Player 3, then under the assessment ρ, he will believe that he is equally likely to be either in the node (a) or (b). Under such a belief, $[m_3]$ will be a best response. Thus, ρ, indeed, corresponds to an interspected equilibrium.

However, the outcome $\{r_1\} \times \{r_2\} \times \{m_3\}$ does not survive the strong version of forward induction. Indeed, suppose we eliminate the strategy $[l_1]$ from the game. I.e., the unique never-best-response strategy of Player 1. In that case, in the reduced game, if the turn is passed to Player 3, he will know that he is in the node (b). Then, he will prefer to deviate to $[l_3]$ rather than to play $[m_3]$. Similarly, $[m_3]$ is not a best response for Player 3 if $[l_2]$ is eliminated from the full game.

Yet, the proposed equilibrium satisfies weak forward induction. Indeed, after elimination of *both* $[l_1]$ and $[l_2]$, Player 3 becomes indifferent among his strategies. In particular, $[m_3]$ remains a best response.

5.3 Interspected equilibrium: an interpretation

In this subsection, I provide an interpretation of interspected equilibrium. I also contrast interspected equilibrium to another equilibrium concept, Selten (1975)'s trembling-hand perfect equilibrium. Consider two different scenarios of how a game can be played:

Suppose first that before the game is played, the players meet and *coordinate* their actions with each other. I.e., the players establish common knowledge of the (extensive-form) game to be played, as well as common knowledge of the intended strategies. Effectively, the players write down an informal contract specifying how they should act in any contingency . For an agreed-upon contract to be actually executed, it should be self-enforcing: no player should find it profitable for himself to deviate if he believes that his opponents play according to the contract. In such a case, how can the players treat deviations from the agreed-upon self-enforcing contract? A deviation can probably not be interpreted as a signal about the deviator's intentions. Indeed, if the deviator had any disagreement with the contract, he would have a chance to express it before the game. Thus, the

players may think that the likeliest cause of deviations are inconsequential one-time trembling-hand mistakes. Therefore, in this case, Selten (1975)'s trembling-hand perfect equilibria may approximate self-enforcing informal contracts.

Suppose now that the players in the game do not *pre-coordinate* their actions. Instead, the players are drawn from an almost-homogenous society in which an almost-commonly known norm prescribes how the game should be played. In such a case, a player who knows the norm believes that his opponents most likely also know this norm and play accordingly. Deviations then can be interpreted as signals that the deviators are unaware of the prevailing social norm. Further, if the society is commonly respectful, i.e., if the members in the society believe in common knowledge of respect even after deviations, then the players would interspect the behavior of deviators. In such cases, interspected equilibria may be viewed as approximations of self-enforcing social norms that may persist in such societies. Note also that a player who knows the prevailing social norm does not have to check himself that this norm is self-enforcing. This player should believe that the description behind the norm corresponds to the true picture of the world and he should play a best-response strategy to that description. In other words, a player conforming to an interspected equilibrium is not required to be an expert in game theory, only in decision theory.

To sum up, the proposed interpretation suggests that a trembling-hand perfect equilibrium may serve as an approximation of a self-confirming *pre-coordinated informal contract*, while an interspected equilibrium may suit as a story of a self-enforcing *social norm played uncoordinatedly* in almost-homogeneous commonly respectful societies.

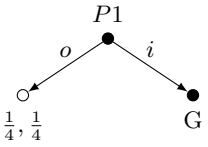
6 Examples

In the previous section, I introduced the concept of interspected equilibrium and studied its main properties. In this section, I provide several examples illustrating additional features of interspected equilibria. The first example shows that an interspected equilibrium may be not subgame perfect. The second example demonstrates that interspected equilibria do not satisfy on-the-path backward induction: an interspected equilibrium of the full game may not induce an interspected equilibrium in a subgame, even when the subgame is reached with positive probability on the equilibrium path of play. The last example reinforces the point made by the first one: the path of an interspected equilibrium may be not supportable in a subgame perfect equilibrium even in perfect information games.

6.1 The Prejudice Game: interspected equilibrium is not SPNE

The example in this subsection show that there may be interspected equilibria whose path cannot be supported in any SPNE. This example is quite similar to the example of explicable and non-subgame perfect equilibrium of Reny (1992). The reason I provide the alternative is twofold. First, in the example I am about to show. the off-equilibrium subgame has a pure strategy Nash equilibrium, which is Pareto superior. Second, the game may be loosely interpreted as a story of how some

prejudices can be persistent in societies, even with mostly sophisticated and respectful members. I.e., the game might be of some interest on its own. Also, in Subsection 6.3, I provide an additional example showing that a path of an interspected equilibrium may be not supportable in in SPNE even in a perfect information game.



(a) Extensive form Γ

	C	I	R
c	$4, \quad 4$	$1, \quad 1$	$-1, 1$
e	$2, -5$	$2, -1$	$-1, 1$
a	$1, \quad 1$	$1, \quad 1$	$0, 0$

(b) Subgame G

	C	I	R
o	$\frac{1}{4}, \quad \frac{1}{4}$	$\frac{1}{4}, \quad \frac{1}{4}$	$\frac{1}{4}, \frac{1}{4}$
$i\,c$	$4, \quad 4$	$1, \quad 1$	$-1, 1$
$i\,e$	$2, -5$	$2, -1$	$-1, 1$
$i\,a$	$1, \quad 1$	$1, \quad 1$	$0, 0$

(c) Normal form $G(\Gamma)$

Figure 4: Prejudice game

Consider the two-player game Γ in Figure 4. At the beginning of the game, Player 1 may either take outside option $[o]$, or to play $[i]$ and enter the subgame G. The subgame G is a simultaneous move game. In the subgame G Player 1 can "cooperate" $[c]$, "exploit" $[e]$, or "apologize" $[a]$. At the same time Player 2 may "Cooperate" $[C]$, "Ignore" $[I]$, or "Rebut" $[R]$.

The subgame G alone has two Nash equilibria: one in pure strategies $\big([c];[C]\big)$ with the payoffs $(4;4)$; and one in randomized strategies $\big(\frac{1}{3}[e]+\frac{2}{3}[a];\frac{1}{2}[I]+\frac{1}{2}[R]\big)$ with the payoffs $(\frac{1}{3};\frac{1}{3})$. Note that in both of these equilibria Player 1 is strictly better off than he is by taking initially the outside option with the payoffs $(\frac{1}{4};\frac{1}{4})$. Thus, the whole game Γ has two subgame perfect equilibria, and in both of them Player 1 enters G. Both of these equilibria are also interspected equilibria.

However, Γ has an interspected equilibrium $\mathcal{E}$ with the outcome $\mu^1 \times \mu^2 = \{o\} \times \{R\}$. Indeed, the play of $\mu^1 \times \mu^2$ can be supported, for instance, by the following cautious assessments: $\rho_1 = \big([R];\frac{1}{3}[C]+\frac{1}{3}[I]+\frac{1}{3}[R]\big)$ and $\rho_2 = \big([o];\frac{1}{10}[ic]+\frac{8}{10}[ie]+\frac{1}{10}[ia]\big)$. These assessments conform to fully interspected hierarchies, since in Γ, fully interspected hierarchies are trivial $\big(\mathcal{H}_1^\infty;\mathcal{H}_2^\infty\big) = \big(\{\mathcal{S}_1\};\{\mathcal{S}_2\}\big)$. Also, $[o]$ and $[R]$ are lexicographic best responses to ρ_1 and ρ_2. Thus, $(\mu^1,\mu^2;\rho_1,\rho_2)$ is indeed an interspected equilibrium.

The equilibrium $\mathcal{E}$ can be interpreted as a story of a persistent prejudice. Suppose that a social norm dictates that Player 1 must always choose the decent outside option $[o]$, yielding a small but positive payoff, rather than to outrageously enter the subgame G. An individual opting to play in G will be perceived as a villain intending to exploit $[e]$ a naive Player 2, who choses to ignore $[I]$. The only appropriate response to such belief for Player 2 is to rebut the nonconformist with $[R]$. Should Player 2 interspect further and figure out that if Player 1 were aware of the equilibrium to be played, then Player 1 would rather apologize $[a]$, instead of exploiting $[e]$? But had Player 1

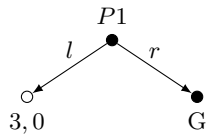
in fact been aware of the equilibrium, he would have never chosen to play $[i]$ in the first place. The play of $[i]$ requires either a Player 1 who is unaware of the equilibrium norm, or an irrational mistake. If the later is thought to be much less likely, then no further equilibrium interspection for Player 2 is possible. In view of this prospect, Player 1, at the beginning of the game, is forced to play according to the restrictive social norm, instead of, say, the Pareto efficient equilibrium $\big([ic];[C]\big)$. In the presence of prejudice, even good intentions of Player 1 may be interpreted falsely and then rebuffed by Player 2. It may seem that only coordinated efforts may overturn the bad outcome in such a case.

6.2 Interspected equilibrium does not satisfy on-the-path backward induction

An equilibrium concept satisfies *on-the-path backward induction*, if the restriction of an equilibrium on any proper subgame reachable with positive probability under equilibrium play induces an equilibrium of the subgame. Remarkably, interspected equilibrium does not satisfy on-the-path backward induction.

	b_1	b_2	b_3	b_4
a_1	$5,3$	$4,3$	$0,2$	$0,2$
a_2	$2,6$	$2,0$	$2,0$	$2,1$
a_3	$0,0$	$0,2$	$0,3$	$7,2$

(a) Subgame G



(b) Extensive form Γ

	b_1	b_2	b_3	b_4
l	$3,0$	$3,0$	$3,0$	$3,0$
ra_1	$5,3$	$4,3$	$0,2$	$0,2$
ra_2	$2,6$	$2,0$	$2,0$	$2,1$
ra_3	$0,0$	$0,2$	$0,3$	$7,2$

(c) Normal form $G(\Gamma)$

Figure 5: No on-the-path backward induction.

Consider, for example, the two-player game Γ given in Figure 5. At the beginning of Γ, Player 1 decides whether to enter the subgame G or to take the outside option yielding the payoffs $(3;0)$. The subgame G is the simultaneous move game given in Figure 5a. The corresponding normal form of the full game is given in Figure 5c.

On the one hand, the construction of fully interspected hierarchies for the subgame G is the following:

$$\Big(\Big\{\{a_1,a_2,a_3\}\Big\};\Big\{\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{a_1,a_2,a_3\}\Big\};\Big\{\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{a_1,a_2\},\{a_1,a_2,a_3\}\Big\};\Big\{\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{a_1,a_2\},\{a_1,a_2,a_3\}\Big\};\Big\{\{b_1\},\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{a_1\},\{a_1,a_2\},\{a_1,a_2,a_3\}\Big\};\Big\{\{b_1\},\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)$$

Therefore, the unique interspected equilibrium outcome for the subgame G is $\{a_1\}\times\{b_1\}$, corresponding to the payoff outcome $(5;3)$. On the other hand, fully interspected hierarchies for the full game Γ are constructed as follows:

$$\Big(\Big\{\{l,ra_1,ra_2,ra_3\}\Big\};\Big\{\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{l,ra_1,ra_3\},\{l,ra_1,ra_2,ra_3\}\Big\};\Big\{\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{l,ra_1\},\{l,ra_1,ra_2\},\{l,ra_1,ra_2,ra_3\}\Big\};\Big\{\{b_2,b_3\},\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{l,ra_1\},\{l,ra_1,ra_2\},\{l,ra_1,ra_2,ra_3\}\Big\};\Big\{\{b_2\},\{b_2,b_3\},\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)\rightarrow$$
$$\Big(\Big\{\{ra_1\},\{l,ra_1\},\{l,ra_1,ra_2\},\{l,ra_1,ra_2,ra_3\}\Big\};\Big\{\{b_2\},\{b_2,b_3\},\{b_1,b_2,b_3\},\{b_1,b_2,b_3,b_4\}\Big\}\Big)$$

Thus, the unique interspected equilibrium outcome in the full game Γ is $\{ra_1\}\times\{b_2\}$. The subgame G is reached with probability 1. But in the subgame, the equilibrium play does not coincide with the equilibrium outcome for the subgame alone. In particular, the full game equilibrium leads to the payoff outcome $(4;3)$, different from the subgame equilibrium outcome $(5;3)$.

In both the subgame alone and the full game, Player 1 is expected to play strategy $[a_1]$, which makes Player 2 first-order indifferent between playing $[b_1]$ or $[b_2]$. Further comparison of these two strategies depends upon which out-of-equilibrium play, $[a_2]$ or $[a_3]$, seems to be less unlikely from the point of view of Player 2. In the subgame G alone, only two steps of interspection are needed to exclude $[a_3]$ from the play of Player 1. I.e., Player 1 will never play $[a_3]$ if he is at least level-1 interspecting and respects Player 2. To eliminate $[a_2]$, Player 2 needs to be sure that Player 1 is much more sophisticated and respectful: Player 1 has to have at least three levels of interspection and believe in at least three levels of Common Knowledge of Respect. Therefore, if the subgame G is played alone, a cautiously rational Player 2 can be expected to play $[b_2]$ rather than $[b_1]$. However, if the full game Γ is played, and Player 1 has moved inside of G, then the presence of an outside option makes it extremely unlikely that Player 1 would play $[a_2]$. Indeed, he has to be irrational to do that, or he has to make a trembling mistake. To ensure sure that he does not play $[a_3]$,we still need to require that he is has at least one level of interspection and respects his opponent. In particular, Player 1 should be very certain that Player 2 will not play the dominated strategy $[b_4]$,

so as not to be tempted to play $[a_3]$ for the prize of 7. Thus, the outside option provides additional assurance that Player 1 would not play $[a_2]$, changing the preferences of Player 2 and leading to the outcome $\{ra_1\} \times \{b_2\}$.

	b_1	b_2	b_3	b_4
a_0	$4,2$	$6,2$	$0,2$	$0,2$
a_1	$5,3$	$4,3$	$0,2$	$0,2$
a_2	$2,6$	$2,0$	$2,0$	$2,1$
a_3	$0,0$	$0,2$	$0,3$	$7,2$

Figure 6: Subgame $\hat{G}$

The fact that Player 2 is first-order indifferent between $[b_1]$ or $[b_2]$ is not a necessary feature of the example above. For instance, one may consider the game $\hat{\Gamma}$ obtained from Γ by replacing the subgame G with the subgame $\hat{G}$ given in Figure 6. The unique interspected equilibrium payoff outcome of $\hat{\Gamma}$ is $(5;3)$, whereas the unique interspected equilibrium payoff outcome of $\hat{G}$ is $(6;2)$, so neither player is first-order indifferent between these outcomes.

6.3 The interspected equilibrium path is not an SPNE path in PI game

The following example illustrates that the interspected equilibrium path may be not supportable in any SPNE even in perfect information games.

Consider the PI game Γ in Figure 7, played by Player 1 and Player 2. At the beginning of the game, at the node h_1, Player 1 can either take the outside option $[l_1]$ and finish the game or to play $[r_1]$, passing the turn to Player 2. In the later case, at h_2, Player 2 can either take the outside option $[L_1]$ and finish the game, or play $[R_1]$ passing the turn back to Player 1. In the later case, the subgame Γ_{h_3} is played. It will be helpful to establish the following three properties of Γ.

Property 1: The strategy $[R_1\, R_2\, L_3]$ is the unique strategy of Γ, which is not interspected rationalizable. Indeed, $[R_1\, R_2\, L_3]$ is dominated by $[L_1]$, and therefore is not interspected rationalizable.

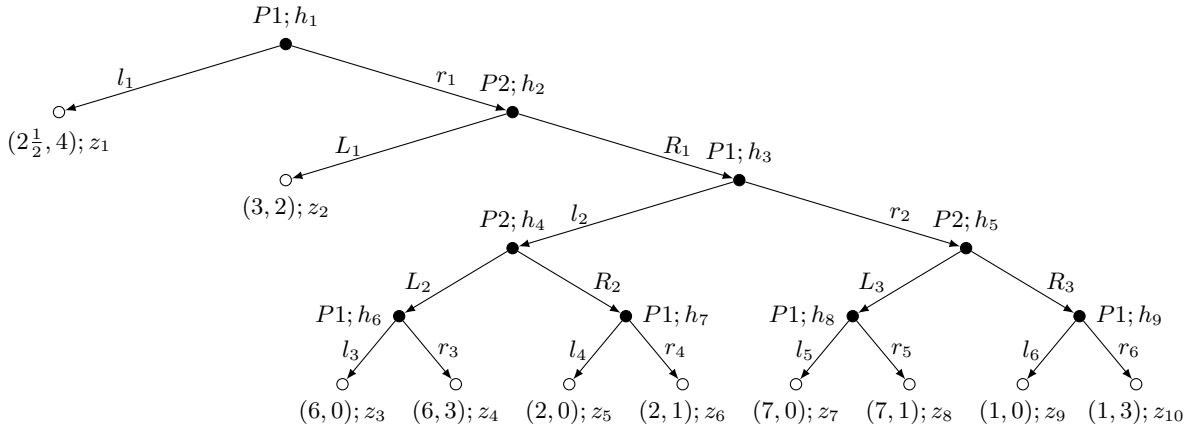


Figure 7: CGRT equilibrium path is not SPNE path in PI Game

Check now that all of the other strategies are interspected rationalizable. Start with Player 2's strategies first. Assuming that all of Player 1 strategies are interspected rationalizable, $[L_1]$ is a best response to any assessment, which upon reaching h_2, puts probability 1 on Player 1 playing $[l_2\, *]$ at the nodes $\{h_6, h_7, h_8, h_9\}$. The strategy $[R_1\, L_2\, L_3]$ is a best response to any assessment, which upon reaching h_2, puts the first probability measure on $[l_2\, r_3\, l_4]$ and the second measure on $[r_2\, r_5\, l_6]$. Analogously, $[R_1\, L_2\, R_3]$ and $[R_1\, R_2\, R_3]$ are interspected rationalizable. Check now that for Player 1, any strategy is interspected rationalizable, provided the set of Player 2's interspected rationalizable strategies is $\big\{L_1;\ R_1\, L_2\, L_3;\ R_1\, L_2\, R_3;\ R_1\, R_2\, R_3\big\}$. Indeed, as I will show below, $[l_1]$ is interspected rationalizable as it is played in an interspected equilibrium. Any strategy of type $[r_1\, r_2\, *]$ is a best response to any assessment starting with $[L_1]$ as a first measure and $[R_1\, L_2\, L_3]$ as a second measure. Any strategy $[r_1\, l_2\, *]$ is a best response to any assessment starting with $[L_1]$ as a first measure and $[R_1\, L_2\, R_3]$ as a second measure. Q.E.D.

Property 2: Any SPNE of Γ *does not involve the play of* $[l_1]$. Consider first the subgame Γ_{h_5} starting from the node h_5. Note that in any SPNE of Γ_{h_5} yielding to Player 2 more than 1 in expectation, Player 2 always plays $[R_3]$. Thus, in any such SPNE the payoff to Player 1 is precisely 1. Analogously, in any SPNE of Γ_{h_4} yielding to Player 2 more than 1, the payoff to Player 1 is precisely 6. Take a look now at the subgame Γ_{h_3}. The claim is that in any SPNE of Γ_{h_4} yielding to Player 2 at least 2, the payoff to Player 1 is at least 3. Indeed, in any SPNE of Γ_{h_4} that reaches h_5 with positive probability, the payoff to Player 2 conditional on reaching h_5 is at most 1. Otherwise, Player 1 would receive 1 from playing $[r_2]$, so that he would strictly prefer to play $[l_1]$ and get at least 2. But conditional on reaching h_4, the payoff to Player 2 is at most 3. Thus, in any SPNE of Γ_{h_4} yielding at least 2 to Player 2, both statements are true: h_4 is reached with at least 50% probability, and at h_4 Player 2 plays only L_2. Thus, in any such SPNE, Player 1 receives at least 3. Look now at the subgame Γ_{h_2}. From the above analysis it follows that in any an SPNE of Γ_{h_2}, Player 1 receives the payoff of at least 3. But then in any SPNE of the full game Γ, at the node h_1, Player 1 strictly prefers to play $[r_1]$ rather than $[l_1]$. Q.E.D.

Property 3: There exists an interspected equilibrium in Γ *with the path equal to* $[l_1]$. For instance, take the following equilibrium $\mathcal{E}$. The equilibrium strategies are $\mu^1 = [l_1]$ and $\mu^2 = [R_1\, R_2\, R_3]$. The supporting cautious assessments are $\rho_1 = \big\{1 \cdot R_1\, R_2\, R_3;\ \frac{1}{4} \cdot \big(L_1 + R_1\, L_2\, L_3 + R_1\, L_2\, R_3 + R_1\, R_2\, R_3\big);\ U(\mathcal{S}_2)\big\}$ and $\rho_2 = \big\{l_1;\ 1 \cdot r_2\, r_5\, r_6;\ 1 \cdot l_2\, l_3\, r_4;\ U(\mathcal{S}_1)\big\}$, where $U(\mathcal{S}_i)$ is the uniform distribution on the set of all Player i's strategies. By Property 1, these assessments conform to fully interspected hierarchies, so that $\mathcal{E}$ is indeed an interspected equilibrium. Q.E.D.

Thus, the path of the interspected equilibrium $\mathcal{E}$ cannot be represented as an SPNE path.

Lastly, one may argue that the equilibrium $\mathcal{E}$ may not be robust to seemingly natural impulses guiding human behavior, such as gratitude and retribution. Indeed, take a closer look at the beliefs of Player 2 in $\mathcal{E}$. In particular, at h_5, his decision to play $[R_3]$ is based on the expectation that Player 1 plays $[l_5\, r_6]$. But think about the subgame Γ_{h_5} from a different perspective. At the node h_5, Player 2 decides whether to give to Player 1 a good or a bad payoff. After observing his decision, Player 1 then may reward or punish Player 2 at no cost. Then should not Player 2 reasonably expect

that Player 1 would reward him in the case of the good payoff for Player 1, and retaliate otherwise? If so, the belief for Player 2 should be $[r_5\, l_6]$, and playing $[R_3]$ becomes suboptimal. However, one may still construct an example similar to $\mathcal{E}$ that may be robust to the behavior considerations of the sort mentioned above. For instance, take the game Γ and add to the node h_5 an outside option O_3 leading to the payoffs $(-100; \frac{1}{2})$, i.e., a horrible outcome for Player 1. Would it still be reasonable to expect retaliation from Player 1, even after Player 2 saved Player 1 by selecting the middle option $[R_3]$? The answer does not seem to be obvious anymore.

7 Discussion

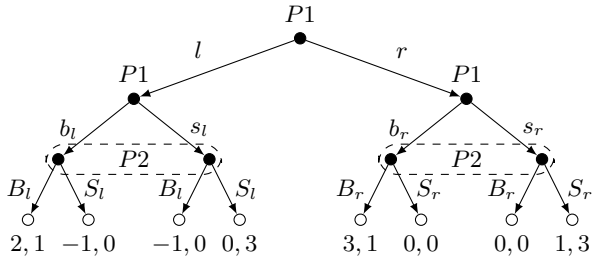
In the previous sections, I provided a general exposition of the theory of interspection in games. In this section, I discuss some additional issues related to interspection theory.

7.1 K-step interspected equilibrium

For any game Γ, fully interspected hierarchies are constructed by iteratively dismissing hierarchical never-best responses. Let L be the interspection index of Γ, i.e., the number of steps needed to reach fully interspected hierarchies. An interspected equilibrium then is a lexicographic Nash equilibrium with a strongly independent common prior assessment conforming to the L-th step interspected hierarchies. By analogy, for any $K < L$ one can consider K-step interspected equilibria, which are lexicographic Nash equilibria with strongly independent common prior assessments conforming to the K-th step interspected hierarchies. Informally, an equilibrium is a K-step interspected equilibrium if it can survive K steps of interspection. The existence and the main properties of the concept of K-step interspected equilibrium are the same as for interspected equilibrium.

Naturally, any interspected equilibrium is a K-step interspected equilibrium for any $K \geqslant 0$, but not otherwise. For instance, a 0-step interspected equilibrium is just a perfect equilibrium in normal form. The concept of K-step interspected equilibrium may correspond to the following story: Whenever logically possible, each player believes that his opponents play according to the equilibrium. If not, then the player expects the opponents to conform to at least K levels of interspection. An equilibrium that is $(K-1)$-step but not K-step interspected cannot persist in environments in which players have at least K levels of interspection and $(K-1)$ levels of Common Knowledge of Respect and of Impeccability. In other words, the more interspection steps a proposed equilibrium survives, the more robust it may be to strategic manipulations. In this sense, an interspected equilibrium may be viewed as being perfectly robust to strategic manipulations.

Consider, for example, the Van Damme (1989) Money Burning game given in Figure 8. This game has four Nash equilibrium outcomes: $\big([rb_r]; [B_r]\big)$, $\big([rs_r]; [S_r]\big)$, $\big(\frac{3}{4} \cdot [rb_r] + \frac{1}{4} \cdot [rs_r]; \frac{1}{4} \cdot [B_r] + \frac{3}{4} \cdot [S_r]\big)$, and $\big([lb_l]; [B_l]\big)$. The unique outcome supportable in interspected equilibrium is $\big([rb_r]; [B_r]\big)$. The outcome $\big([lb_l]; [B_l]\big)$ can be supported in a two-step equilibrium, but not in a three-step equilibrium. The outcomes $\big([rs_r]; [S_r]\big)$ and $\big(\frac{3}{4} \cdot [rb_r] + \frac{1}{4} \cdot [rs_r]; \frac{1}{4} \cdot [B_r] + \frac{3}{4} \cdot [S_r]\big)$ are supportable only in zero-step equilibria. Thus, the best outcome for Player 1 $\big([rb_r]; [B_r]\big)$ is the



(a) Extensive form Γ

	$B_l\, B_r$	$B_l\, S_r$	$S_l\, B_r$	$S_l\, S_r$
$l\ b_l$	$2,1$	$2,1$	$-1,0$	$-1,0$
$l\ s_l$	$-1,0$	$-1,0$	$0,3$	$0,3$
$r\, b_r$	$3,1$	$0,0$	$3,1$	$0,0$
$r\, s_r$	$0,0$	$1,3$	$0,0$	$1,3$

(b) Normal form $G(\Gamma)$

Figure 8: The Battle of Sexes with Money Burning

unique equilibrium outcome surviving three steps of interspection.

7.2 Respect as a representation of ϵ-doubt Respect

The notion of Respect that is introduced in this paper may seem very extreme. Indeed, it requires that respectful players treat lower interspection levels of opponents as *infinitely* less likely than higher levels. One may wonder if Respect may be relaxed to allow for lower interspection levels to be treated as still less likely, but not infinitely less likely. For instance, take Player i with interspection level k_i and interspection state-space $\Omega_i^{k_i}$. Let P be any classical probability measure with full support on $\Omega_i^{k_i}$. Define ϵ-doubt Respect as follows:

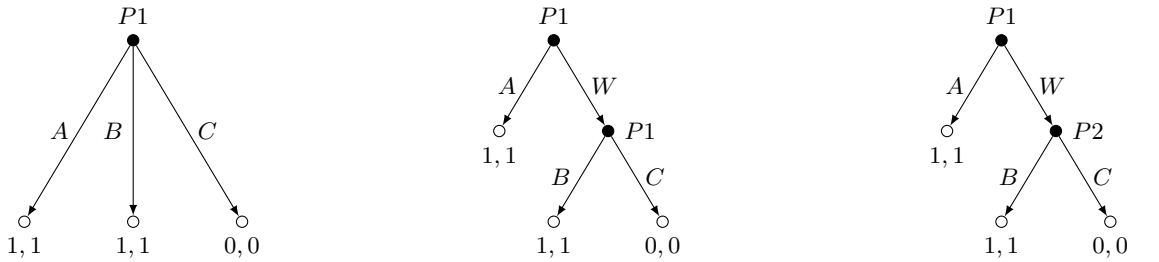
Definition (ϵ-Doubt Respect)**.** *A probability measure P, with full support on the interspection state-space $\Omega_i^{k_i}$, satisfies ϵ-doubt Respect for $\epsilon > 0$, if for any two states $\omega_1, \omega_2 \in \Omega_i^{k_i}$ such that the interspection index of the interspection event containing ω_1 is higher than the index of the event containing ω_2, the following holds:*

$$P(\omega_2) < \epsilon \cdot P(\omega_1)$$

Thus, Respect may be viewed as the limiting case of ϵ-doubt Respect with $\epsilon \to 0$. Indeed, using the results from Appendix A.2, one can show that in finite games, if Player i's preferences can be represented by some ϵ-doubt Respect assessment P_ϵ for any ϵ close to 0, then these preferences can be represented by a cautious lexicographic assessment $\boldsymbol{P}$ satisfying Respect. The reverse is also true: if Player i's preferences can be represented by a cautious lexicographic assessment $\boldsymbol{P}$ satisfying Respect, then for some $\epsilon > 0$, they can be represented by an ϵ-doubt Respect assessment P_ϵ.

As in finite games, there are only finitely many strategies and only finitely many steps in the construction of fully interspected hierarchies, for each finite game Γ, there exists a constant $\epsilon(\Gamma)$ such that the result of Theorem 3 (Axiomatic Characterization) still holds if Common Knowledge of Respect is replaced by Common Knowledge of $\epsilon(\Gamma)$-doubt Respect. In other words, informally:

> Interspected rationalizability is the prediction under Cautious Rationality, Sufficient Interspection, and Common Knowledge of the Game Form, of the Payoffs, of Impeccability, and of Low-Doubt Respect.



(a) One period game Γ_1 (b) Two period game Γ_2 (c) Two player game Γ_3

Figure 9: Neglect of own mistakes

7.3 Cautious behavior and players' own mistakes

The model of interspection theory is one of the possible ways to model cautiously rational behavior in games. Under interspection theory, players take into account the possibility that their opponents may play any strategy. However, players are not cautious against their own trembling-hand mistakes. Even if such mistakes occur, the assumption is that the players treat them as being so unlikely that there is no practical reason to account for them in the future.

Consider, for example, the games in Figure 9. In the game Γ_1 in Figure 9a, Player 1 selects $[A]$, $[B]$, or $[C]$ in one period. The set of interspected rationalizable strategies is $\{A, B\}$. The game Γ_2 in Figure 9b is logically equivalent to Γ_1. However, in Γ_2, Player 1 makes his decision in two periods; first he choses $[A]$ or to wait $[W]$, and if he waits, he then further choses $[B]$ or $[C]$. Again, interspected rationalizability predicts logically the same set $\{A, WB\}$. Now, the game Γ_3 in Figure 9c is played by two players. The preferences of Player 1 and Player 2 are perfectly aligned. For this game, the unique interspected rationalizable outcome, however, is $\{A\}$. The games Γ_1 and Γ_2 are treated in interspection theory as equivalent: they are both played by Player 1, who disregards the possibility of his own mistakes. In the game Γ_3, however, a cautious Player 1 would not wait $[W]$. He would not wait not because he thinks Player 2 might make a trembling mistake, but because he thinks Player 2 might be irrational. Note also that in Γ_2, the outcome $\{WB\}$ can be supported in interspected equilibrium. However, this outcome is not consistent with perfectly cautious behavior: it seems that a truly cautious player should also factor in the possibility of his own mistakes. Allowance for such kind of incautious behavior is the price to pay for keeping interspection theory simple.

Finally, even though players completely disregard the possibility of trembling mistakes under interspection theory, the theory still provides a guidance to a player in case he actually makes one. Specifically, if he makes a mistake, a player does not have to change his assessment of the opponents' behavior. Indeed, he is the only player who knows that his action was a mistake. Therefore, the next time this player acts, he should simply calculate a best response continuation strategy to his Bayesian updated assessment. Analogue of Theorem 1 guarantees sequential rationality of that continuation strategy. In other words, after an extremely unlikely one-time trembling-hand mistake, the mistaken player should make a one-time correction.

7.4 Independence or correlation

In the analysis of games with more than two players an issue arises whether players' assessments of the opponents' behavior should be modeled as independent or correlated.

Interspected rationalizability assumes that players form correlated assessments. Alternatively, it is possible to restrict attention on strongly independent assessments. However, in that case the sets of hierarchical best responses might be not computable.[9] Yet, the notion of Respect entails some weak requirement of independence: in an assessment satisfying Respect, the fact that one of the players has a low interspection level does not affect the likelihood that other players have low interspection levels.

Interspected equilibrium assumes that players' behavior is strongly independent. Even the behavior of deviators, who are unaware of the equilibrium. This is despite the fact that fully interspected hierarchies on deviators' strategies are constructed with correlated interspected rationalizability. This may seem like a logical flaw. However, this construction is consistent if an interspected equilibrium is understood as a description of an environment with independent players in which deviators, who are unaware of the equilibrium, assess the game in a correlated fashion.

7.5 Rationality and bounded interspection

Bounded interspection is not always a deviation from perfect rationality. In fact, a rational sophisticated player may find it sufficient to think only on few interspection levels. For instance, if a player attributes a high probability to low interspection levels of his opponents, then thinking beyond a few levels may be unnecessary, since additional interspection would not change his decision.

Under interspection theory, a sophisticated player believes in Common Knowledge of Respect. He then desires to think on as many levels of interspection as will suffice to fully comprehend the game. In environments with commonly respectful players, the desired level of interspection rises as in an arms race. The converse is also true: a player who does not deeply respect his opponents may often decide not to overthink his play against them.

8 Conclusion

Interspection theory developed in this paper addresses the question of how sufficiently sophisticated players will behave in games if suboptimal actions are interpreted by other players as evidence of a player's lower sophistication level, rather than inconsequential one-time trembling mistakes. This theory employs the notion of cautious rationality, a version of Bayesian rationality against full-support lexicographic probability systems. Interspection theory has the following attractive features:

1. interspection theory deals with general finite extensive-form games;

[9] The issue is that one would need to check responses to all of the *strongly independent* assessments conforming to a given product hierarchy.

2. extensive-form games can be equivalently analyzed in their normal forms;

3. for finite games, players' probability assessments are distributed on finite state-spaces;

4. the epistemic model of interspection theory admits a tractable reformulation in terms of primitives of the game;

5. interspection theory incorporates both rationalizable behavior and equilibrium;

6. players' behavior is admissible;

7. interspection theory captures a form of forward induction logic.

The main solution concepts in interspection theory are interspected rationalizability and interspected equilibrium.

The notion of interspected rationalizability applies to environments with sufficiently interspecting cautiously rational players, provided there is Common Knowledge of the Game Form, of the Payoffs, of Impeccability, and of Respect. Interspected rationalizability coincides with iterative admissibility in two-player games. In multi-player games, these two concepts are different. In perfect information games without chance nodes and without relevant ties, the unique interspected rationalizable outcome coincides with the unique backward induction outcome.

Interspected equilibria correspond to situations in which, before observing any deviation, players believe that their opponents follow the equilibrium strategy. Moreover, after observing a deviation, players believe that their opponents play a strategy corresponding to the highest logically possible interspection level. Interspected equilibria exist in general finite extensive form games and satisfy invariance, (weak) sequential rationality, and admissibility. Also, in two-player games, interspected equilibria satisfy forward induction: an interspected equilibrium of the full game corresponds to an interspected equilibrium of the reduced game after a one-round elimination of *all never-best-response* strategies of *one* of the players. In multi-player games, interspected equilibria only satisfy weak forward induction: an interspected equilibrium of the full game corresponds to an interspected equilibrium of the reduced game after elimination of *all interspected non-rationalizable strategies* of *all* of the players. In general, interspected equilibria are not subgame perfect.

Finally, the concept of interspected equilibrium may be interpreted as a story of self-enforcing *social norms played uncoordinatedly* in almost-homogeneous commonly respectful societies. In contrast, the concept of trembling-hand perfect equilibrium may serve as an approximation of self-confirming *pre-coordinated informal contracts*.

A Properties of Lexicographic Assessments

A.1 Lexicographic probability assessments

In this subsection, I show that lexicographic subjective probability assessments are well defined, i.e., they do not depend on particular lexicographic representations of decision makers' preferences.

The results in this subsection can be extended to the case of LPSs without full support on Ω.

Let $\succeq$ be cautious lexicographic preferences on acts in $\mathcal{P}^{\Omega}$, where Ω is a finite state-space. Let (u, ρ) be some lexicographic representation of $\succeq$, where $u : \mathcal{P} \to \mathbb{R}$ is an in-state utility function and $\rho = (\lambda_1, \lambda_2, ..., \lambda_K)$ is an LPS with full support on Ω. Let $\succeq_{\rho}^{imp}$ be the imputed preferences on $\mathbb{R}^{\Omega}$ induced by representation (u, ρ).

The following lemma is similar to Lemma 2 from BBD1. It shows that $\succeq_{\rho}^{imp}$ does not depend on particular representation (u, ρ) of $\succeq$:

Lemma 1. *For any cautious lexicographic preferences on $\mathcal{P}^{\Omega}$ subjective probability assessment $\boldsymbol{P}[\succeq]$ defined as the inputed preferences $\succeq_{\rho}^{imp}$ on $\mathbb{R}^{\Omega}$ for some representation (u, ρ) of $\succeq$ does not depend on the particular representation (u, ρ).*

Proof. Fix some in-state utility function u, and consider all representations for $\succeq$ of the form (u, ρ). By (BBD) Axiom 3 (Nontriviality) and (BBD) Axiom 5′ (State Independence), the image of the set of all acts $\mathcal{P}^{\Omega}$ after taking in-state expectation by u contains some non-empty cube C in $\mathbb{R}^{\Omega}$. Preferences $\succeq$ then define uniquely projected preferences in C. We know that these preferences can be represented by some LPS ρ. But then these preferences can be described as follows:

- in the first approximation, C is divided by some parallel hyper-spaces into first-order indifference surfaces;
- in the second approximation, each first-order indifference surface is divided by parallel hyper-spaces into second order indifference curves;
- and so on.

In other words, the induced preferences in C can be represented as a geometrical object uniquely pinned down by $\succeq$. This object (collection of indifference surfaces) can be extended to the whole $\mathbb{R}^{\Omega}$ by translation. Note that for fixed u any other LPS representing the preferences on C must generate the same geometrical object, i.e. the same inputed preferences on $\mathbb{R}^{\Omega}$.

To finish the proof it remains to notice that u is unique up to positive affine transformation, and the geometrical object remains invariant to translations and positive proportional rescaling. Q.E.D. □

A.2 Lexicographic assessments and test sequences

In this subsection, I relate lexicographic preferences to test sequences.

Following BBD2, *nested convex combination* $r \square \rho$ is defined for any LPS $\rho = (\lambda_1, \lambda_2, ..., \lambda_K)$ and vector $r \in (0, 1)^{K-1}$ as the following measure on Ω:

$$r \square \rho = (1 - r_1)\lambda_1 + r_1\big[(1 - r_2)\lambda_2 + r_2[(1 - r_3)\lambda_3 + ... + r_{K-1}\lambda_k]...\big]$$

For any vector $x \in \mathbb{R}^{\Omega}$ and nested convex combination $r \square \rho$ define the expectation $E_{r \square \rho}(x)$ of x against measure $r \square \rho$ as usual:

$$E_{r\square\rho}(x) = \sum_{\omega\in\Omega} \big(r\square\rho\big)(\omega)\cdot x_\omega$$

Limiting nested convex combination $\big\{r(n)\square\rho\big\}_{n\in\mathbb{N}}$ is a sequence of nested convex combinations with $r(n)\to 0$, as $n\to\infty$.

The following lemma is similar to Proposition 1 from BBD2 and it states that lexicographic preferences on $\mathbb{R}^\Omega$ induced by LPS ρ coincide with preferences induced by the tail of any limiting nested convex combination $r(n)\square\rho$:

Lemma 2. *Let ρ be LPS distributed on finite state-space Ω, $r(n)\square\rho$ be some limiting nested combination, and $\succeq_\rho^{imp}$ be imputed preferences induced by ρ on $\mathbb{R}^\Omega$. Then for any vectors $x,y\in\mathbb{R}^\Omega$ the following two equivalences hold:*

1. $\Big(x\sim_\rho^{imp} y\Big)\leftrightarrow\Big(\forall n\in\mathbb{N}:\ \ E_{r(n)\square\rho}(x)=E_{r(n)\square\rho}(y)\Big)$
2. $\Big(x\succ_\rho^{imp} y\Big)\leftrightarrow\Big(\exists M\in\mathbb{N}:\forall n>M:\ \ E_{r(n)\square\rho}(x)>E_{r(n)\square\rho}(y)\Big)$

Proof. Part 1 is trivial. Part 2 is just a slightly restated version of Proposition 1 from BBD2. □

The following is a restated version of Proposition 2 from BBD2.

(BBD2) Proposition 2. *For any sequence $p(n)$ of nonnegative measures on a finite state-space Ω there exists a subsequence $p(n_m)$, which can be represented as a limiting nested convex combination $p(n_m)=r(m)\square\rho$ for some LPS $\rho=(\lambda_1,\lambda_2,...,\lambda_K)$ and $r\in(0,1)^{K-1}$ with $r(n)\to 0$.*

The above can be informally summarized as follows:

Statement. *From any test sequence of nonnegative measures on finite state-space Ω one can select a subsequence, which in the limit induces preferences on $\mathbb{R}^\Omega$ equivalent to preferences induced by some LPS.*

Note also that if we are interested in player's lexicographic preferences over the set of his strategies in a finite game, then these preferences can be represented by some nested convex combination $r\square\rho$ without going to the limit $r(n)\to 0$. This fact is shown in BBD2 Proposition 1. Also, a similar idea implies for finite games the following corollary :

Corollary 2. *Suppose that in a finite game Player i's preferences over his strategies are represented by LPS assessment ρ on the set of his opponents' strategy profiles $\mathcal{S}_{-i}$. Then these preferences can also be represented by some LPS $\tilde\rho$ on $\mathcal{S}_{-i}$ such that consecutive measures in $\tilde\rho$ have increasing support.*

A.3 Infinitely more likely events

In this subsection, I provide a formal treatment of the "infinitely more likely" relation.

Let $\succeq$ be cautious lexicographic preferences over acts in $\mathcal{P}^{\Omega}$, where Ω is a finite state-space. Let (u, ρ) be some lexicographic representation of $\succeq$, where $u : \mathcal{P} \to \mathbb{R}$ is an in-state utility function and $\rho = (\lambda_1, \lambda_2, ..., \lambda_K)$ is an LPS with full support on Ω. Let $\boldsymbol{P}[\succeq]$ be the induced subjective probability assessment on $\mathbb{R}^{\Omega}$.

For each nonempty event $S \subseteq \Omega$ define ρ_S as the beginning of ρ, which covers S. I.e., $\rho_S = (\lambda_1, ..., \lambda_t)$, with $t = \min\left\{n \in \mathbb{N} : S \subseteq \bigcup_{i=1}^{t} supp\{\lambda_i\}\right\}$. For any two disjoint nonempty events A and B say that A is infinitely more likely in $\boldsymbol{P}[\succeq]$ than B, symbolically $A \gg_{\boldsymbol{P}[\succeq]} B$, if elements in B start getting covered by ρ only after A is covered:

$$\Big(A \gg_{\boldsymbol{P}[\succeq]} B\Big) \leftrightarrow \Big(supp\{\rho_A\} \cap B = \varnothing\Big)$$

Alternative way to think about $P[\succeq]$ is the following. Take any LPS representation ρ of $P[\succeq]$. Then ρ induces complete weak order on elements of Ω : ω_1 precedes ω_2, if ω_1 is covered by ρ strictly before ω_2. Event A then is infinitely more likely than event B if and only if each element in A precedes each element in B. Such interpretation immediately implies the following two properties of $P[\succeq]$:

$$\Big(A \gg_{\boldsymbol{P}[\succeq]} C\Big) \& \Big(B \gg_{\boldsymbol{P}[\succeq]} C\Big) \Rightarrow \Big(A \cup B \gg_{\boldsymbol{P}[\succeq]} C\Big)$$

$$\Big(A \gg_{\boldsymbol{P}[\succeq]} B\Big) \& \Big(A \gg_{\boldsymbol{P}[\succeq]} C\Big) \Rightarrow \Big(A \gg_{\boldsymbol{P}[\succeq]} B \cup C\Big)$$

Of course, for $\gg_{\boldsymbol{P}[\succeq]}$ to be well defined, it should not depend on particular LPS representation ρ of $P[\succeq]$. One way to show this is to restate $\gg_{\boldsymbol{P}[\succeq]}$ directly in terms of the original preferences $\succeq$. In order to do this additional definitions will be useful.

I say that for some nonempty event $A \subseteq \Omega$ act $z \in \mathcal{P}^{\Omega}$ *A-dominates* act $z' \in \mathcal{P}^{\Omega}$ if for all $\omega \in A$: $z_w \succeq z'_w$, and for at least one $\omega \in A$: $z_w \succ z'_w$. In that case I write $z \; Dom_A(\succeq) \; z'$. Robust preferences conditional on event A are defined as follows:

Definition (Robust Preferences)**.** *Given cautious preferences $\succeq$ on $\mathcal{P}^{\Omega}$ and some non-empty finite event $A \subseteq \Omega$, the robust preferences conditional on A over acts in $\mathcal{P}^{\Omega}$, symbolically $\succ_A^{\boldsymbol{R}}$, are defined as follows:*

$$\forall x, y \in \mathcal{P}^{\Omega} : \Big(x \succ_A^{\boldsymbol{R}} y\Big) \leftrightarrow \Big(\exists z, z' \in \mathcal{P}^{\Omega} \; \exists \alpha \in (0,1) : \big(z \; Dom_A(\succeq) \; z'\big) \wedge \big(\alpha{\cdot}x + (1-\alpha){\cdot}z' \; \succ_A \; \alpha{\cdot}y + (1-\alpha){\cdot}z\big)\Big)$$

Informally, x is robustly better than y on A if x is better than y plus some act "positive" on A.

Robust preferences have a representation derived from a lexicographic representation of the original preferences:

Lemma 3. *Suppose that $\succeq$ are cautious lexicographic preferences over acts $\mathcal{P}^\Omega$, where Ω is a finite state-space. Let (u, ρ) be any lexicographic representation of $\succeq$, and A a nonempty event in Ω. Then the robust preference $\succ_A^{\boldsymbol{R}}$ on $\mathcal{P}^A$ have lexicographic representation $(u, \rho_A(A))$, where $\rho_A(A)$ is the Bayesian update of ρ_A conditional on A.*

Proof. First, suppose that for two acts $x, y \in \mathcal{P}^A$: $x \succ_{(u,\rho_A(A))} y$. Let $\omega \in A$ be a state, which is covered last by ρ among the elements in A. Take two acts $z, z' \in \mathcal{P}^A$ such that $z_{\hat{\omega}} = z'_{\hat{\omega}}$ for all $\hat{\omega} \in A$ with $\hat{\omega} \neq \omega$, and $z \succ_\omega z'$. Then $z \; Dom_A(\succeq) \; z'$. Also, for $\alpha \in (0,1)$ sufficiently close to 1 it will be that $\alpha \cdot x + (1-\alpha) \cdot z' \; \succ_A \; \alpha \cdot y + (1-\alpha) \cdot z$. Thus, $x \succ_A^{\boldsymbol{R}} y$.

Second, suppose that for two acts $x, y \in \mathcal{P}^A$: $x \sim_{(u,\rho_A(A))} y$. Then, for any two acts $z, z' \in \mathcal{P}^A$ with $z \; Dom_A(\succeq) \; z'$ and any $\alpha \in (0,1)$ we will have that $\alpha \cdot x + (1-\alpha) \cdot z' \; \prec_{(u;\rho_A(A))} \; \alpha \cdot y + (1-\alpha) \cdot z$, and hence $\alpha \cdot x + (1-\alpha) \cdot z' \; \prec_A \; \alpha \cdot y + (1-\alpha) \cdot z$. Thus, it will be not the case that $x \succ_A^{\boldsymbol{R}} y$.

Therefore, $(u, \rho_A(A))$ indeed represents $\succ_A^{\boldsymbol{R}}$. Q.E.D. □

Lexicographic representation of robust preferences allows to define "infinitely more likely" relation in terms of the original preferences:

Proposition 1 (Invariant Definition)**.** *Suppose that $\succeq$ is cautious lexicographic preferences over acts $\mathcal{P}^\Omega$, where Ω is a finite state-space. Let $A, B \subseteq \Omega$ be two non-empty disjoint events. The following are equivalent:*

- *A is infinitely more likely than B according to $\succeq$: $A \gg_{\boldsymbol{P}[\succeq]} B$;*
- $\forall x, y \in \mathcal{P}^\Omega : \left(x \succ_A^{\boldsymbol{R}} y \right) \implies \left(x \succ_{A \sqcup B}^{\boldsymbol{R}} y \right)$.

The proof of Proposition 1 is left as an exercise.

B Proofs

B.1 Proof of Theorem 1 (Normal-Form Sufficiency)

Proof. The proof is given for the simpler case of games without chance nodes. The idea of this proof can be directly extended to the general case.

Suppose that Player i initially forms a cautious subjective assessment $\boldsymbol{P}_i$ on the strategy profiles of his opponent. Let σ^1 be an initial best response to $\boldsymbol{P}_i$. We need to show that σ^1 remains a best response to $\boldsymbol{P}_i^h$ in all of Player i's σ^1-relevant information sets h.

Suppose, on the contrary, that there exists σ^1-relevant information set h for Player i and some continuation strategy $\tilde{\sigma}_h^1$ that is strictly better against $\boldsymbol{P}_i^h$ than the continuation of σ^1. Consider then strategy $\hat{\sigma}^1$ in the initial normal form that is obtained from σ^1 by replacing the actions after h to the actions prescribed by $\tilde{\sigma}_h^1$. Then $\hat{\sigma}^1$ is strictly better than σ^1 against $\boldsymbol{P}_i$, as $\boldsymbol{P}_i$ has full support. Thus, σ^1 is not an initial best response. Contradiction. Q.E.D. □

B.2 Proof of Theorem 2 (Respect in Strategy Form)

Proof of Part 1. Consider any interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$ constructed upon a product hierarchy $\mathcal{H}_{-i}$ and satisfying Respect. We need to show that $\boldsymbol{P}_i$ can be equivalently represented by some cautious assessment $\tilde{\boldsymbol{P}}_i$ on $\mathcal{S}_{-i}$ conforming to $\mathcal{H}_{-i}$. Indeed. Take any LPS representation $\rho_i = (\lambda_1, ..., \lambda_m)$ of $\boldsymbol{P}_i$. Take any measure λ_j from ρ_i and construct the combined measure $\tilde{\lambda}_j$ on $\mathcal{S}_{-i}$ that is obtained from λ_j by merging probability masses corresponding to the same strategy profiles. Consider the resulting LPS $\tilde{\rho}_i = (\tilde{\lambda}_1, ..., \tilde{\lambda}_m)$. By CDSA, $\tilde{\rho}_i$ and ρ_i induce the same preferences on Player i's strategies.

It remains to show that $\tilde{\boldsymbol{P}}_i$ conforms to $\mathcal{H}_{-i}$. Consider any two strategy profiles σ^{-i} and $\hat{\sigma}^{-i}$ such that $Ind_{\mathcal{H}_{-i}}(\sigma^{-i}) > Ind_{\mathcal{H}_{-i}}(\sigma^{-i})$. Return to the assessment $\boldsymbol{P}_i$ on the interspection state-space $\Omega_i^{k_i}$. Consider the interspection event $E^{Ind_{\mathcal{H}_{-i}}(\sigma^{-i})}$. As type t_i satisfies Respect, a copy of σ^{-i} from $E^{Ind_{\mathcal{H}_{-i}}(\sigma^{-i})}$ is covered by an LPS representation ρ_i of $\boldsymbol{P}_i$ strictly before all the copies of $\hat{\sigma}^{-i}$. But then in $\tilde{\rho}_i$, the profile σ^{-i} is covered strictly before $\hat{\sigma}^{-i}$. Thus, $\tilde{\rho}_i$ conforms to $\mathcal{H}_{-i}$. Q.E.D.

Proof of Part 2. Consider any cautious assessment $\tilde{\boldsymbol{P}}_i$ on $\mathcal{S}_{-i}$ conforming to a product hierarchy $\mathcal{H}_{-i}$. We need to find some interspecting type $t_i = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$, which is constructed upon $\mathcal{H}_{-i}$, satisfies Respect, and induces the same preferences on $\mathcal{S}_i$ as $\tilde{\boldsymbol{P}}_i$. Fix any LPS $\tilde{\rho}_i$ representing $\tilde{\boldsymbol{P}}_i$.

I now provide an algorithm, which takes as input an LPS $\tilde{\rho}_i = (\tilde{\lambda}_1, ..., \tilde{\lambda}_l)$ with full support on $\mathcal{S}_{-i}$ conforming to hierarchy $\mathcal{H}_{-i}$, and produces as output an LPS ρ_i with full support on the interspection state-space $\Omega_i^{k_i}$ constructed upon $\mathcal{H}_{-i}$ such that ρ_i induces the same preferences on Player i's strategies and ρ_i satisfies Respect.

Algorithm:

The algorithm stores data in two stacks: Input Stack and Output Stack. During the work of the Algorithm Input Stack will be filled with measures from $\tilde{\rho}_i = (\tilde{\lambda}_1, ..., \tilde{\lambda}_m)$, one by one. In parallel, Output Stack will be filled with measures $(\nu_1, ..., \nu_l)$, which will eventually comprise the resulting LPS ρ_i.

Recall that interspection state-space $\Omega_i^{k_i}$ is the disjoint union of k_i^{N-1} interspection events $E^{\boldsymbol{k}}$, for all multi-indices $\{0\}^{N-1} \leqslant \boldsymbol{k} \leqslant \{k_i - 1\}^{N-1}$.

During the work of the Algorithm each of the interspection events will be exactly in one of three possible states: *Intact*, *Active*, or *Treated*. At the beginning all events are Intact. As the Algorithm proceeds each of the events will be changing its status to Active and then Treated. At the end all events will be Treated. Also, once some event becomes Treated, it remains Treated forever.

At any during the work of the Algorithm the condition for interspecting an event to remain Intact is precisely the following:

- interspection event $E^{\boldsymbol{k}}$ remains Intact if and only if at the current stage there is at least one interspection event $E^{\boldsymbol{k}'}$ with higher index, $\boldsymbol{k}' > \boldsymbol{k}$, which is still not Treated.

That is for any Active or Treated event all events with higher interspection indices are Treated.

An LPS ρ distributed on $\mathcal{S}_{-i}$ (not necessary with full support) *covers* interspection event $E^{\boldsymbol{k}}$ if all strategy profiles from $E^{\boldsymbol{k}}$ are covered by the support of ρ. Note, that if an interspection

state-space is constructed upon some product hierarchy and ρ covers some interspection event $E^{\boldsymbol{k}}$, then ρ also covers all interspection events with indices higher than $\boldsymbol{k}$.

The Algorithm proceeds in large steps. Each large step consists of several small steps as follows:

Beginning of the Algorithm

Step 0*:* Make the highest index interspection event $E^{(k_i-1;k_i-1;...;k_i-1)}$ Active. Empty Input and Output Stacks.

Step j, for all $1 \leqslant j \leqslant m$, where m is the number of measures in input LPS $\tilde{\rho}_i$:

1. add $\tilde{\lambda}_j$ to Input Stack, so as Input Stack becomes $\rho(j) = (\tilde{\lambda}_1, ..., \tilde{\lambda}_j)$;
2. denote through $C(j)$ the set of interspection events covered by $\rho(j)$
3. repeat the following Procedure, until all events in $C(j)$ become Treated:
   (a) calculate which events are currently Active;
   (b) for $s = 1$ to j:
   - take $\tilde{\lambda}_s$;
   - construct measure ν on $\Omega_i^{k_i}$ by splitting each probability mass from λ_s equally among all of the copies of corresponding strategy profile in events, which are currently Active or Treated;
   - add ν to the end of Output Stack

   (c) change the status of all currently Active events in $C(j)$ to Treated;

Last Step: After the end of Step m return as output the LPS contained in Output Stack.
End of the Algorithm

Two things might potentially go wrong during the work of the Algorithm.

First, at some stage during Step j, 3 (b) it may become impossible to find some strategy profile from $\tilde{\lambda}_s$ inside of currently Treated or Active events. Suppose that this happens. Then, it must be precisely at the measure $\tilde{\lambda}_j$, as in the previous Step $(j-1)$ the Algorithm was able to find combinations in events which were Active or Treated. These events are, moreover, Active or Treated right now. Let σ^{-i} be a strategy profile, which is covered by $\tilde{\lambda}_s$, but does not belong to any currently Active or Treated event. Take the highest index event that still contains σ^{-i}. Let this event be $E^{\boldsymbol{k}}$. As this event is currently Intact, there exist event $E^{\boldsymbol{k}'}$ with higher index, which is not yet Treated. By the construction of the Algorithm, $E^{\boldsymbol{k}'}$ then was not covered by the $(j-1)$-th step Input Stack LPS. Thus, the strategy profile σ^{-i} with index $\boldsymbol{k}$ is covered by $\tilde{\rho}_i$ no later than some strategy profile $\hat{\sigma}^{-i}$ from $E^{\boldsymbol{k}'}$. This contradicts to $\tilde{\rho}_i$ conforming to $\mathcal{H}_{-i}$. Thus, this problem cannot occur during the work of the Algorithm.

Second, at some Step j the repetition of Procedure 3 (a,b,c) might freeze. Note, however, that for any event covered by $\rho(j)$, the events with higher indices are also covered by $\rho(j)$. Thus, if after a repetition of the Procedure there are events in $C(j)$, which are not yet Treated, then at least

one of those events is Active. Thus, after each repetition of the Procedure at least one event in $C(j)$ becomes newly Treated. As there are finitely many events in $C(j)$, the Procedure stopes after finitely many steps.

Therefore, the Algorithm will work well for any input LPS conforming to a product hierarchy.

To finish the proof of Part 2 it remains to notice that:

- after each Step j the Output Stack LPS induces preferences of $\mathcal{S}_i$, which are the same as those induced by $\rho(j) = (\tilde{\lambda}_1, ..., \tilde{\lambda}_j)$. Thus, at the end the output LPS ρ_i induces the same preferences as $\tilde{\rho}_i$;
- each interspection event becomes Treated precisely at the moment it is covered by the output LPS;
- eventually the whole $\Omega_i^{k_i}$ is covered by ρ_i. I.e., ρ_i corresponds to some interspecting type $t = (k_i; \sigma^i; \Omega_i^{k_i}; \boldsymbol{P}_i)$;
- moreover, such interspecting type t satisfies Respect, as. by construction, lower index interspection events begin being covered in ρ_i (i.e., become Active) only after all higher index events are covered in ρ_i (i.e., become Treated).

Q.E.D.

B.3 Proof of Theorem 3 (Axiomatic Characterization)

Consider any finite extensive-form game Γ with the normal form $G(\Gamma)$. Suppose Axioms 1-6 hold.

Prove first that CSDA is satisfied. Indeed, under Cautious Rationality, the fact that Player i knows his own Impeccability and the Game Form, implies that the states in his subjective state-space are labeled by his opponents' strategy combinations (chance-node realizations do not label states as chance nodes are viewed as objective lotteries). Also, the fact that Player i knows his payoffs is interpreted as follows. First, Player i knows his in-state objective expected utility function. Second, he thinks that any of his strategies generates the same objective lottery in states with the same label. Thus, CSDA holds for Player i.

Prove next the following Statement:

> $B(\mathcal{H}_{-i}^k)$ is the set of strategies that can be played by a level-k interspecting Player i.

Indeed:

Suppose Player i has level-1 of interspection. By CSDA, his assessment can be represented as a full-support LPS on his opponents' strategy profiles. Therefore, he plays some strategy from $B(\mathcal{H}_{-i}^0) = B(\mathcal{S}_{-i})$. Also, by definition of hierarchical best response set, any strategy from $B(\mathcal{H}_{-i}^0)$ can potentially be played by a level-1 interspecting Player i.

Suppose now that Player i is level-2 interspecting. He considers events corresponding to his opponents having interspection levels 0 and 1. As he knows their Impeccability, he additionally

infers that for $j \neq i$, a level-1 interspecting Player j can play strategies precisely from $B(\mathcal{H}^0_{-j})$. Player i respects his opponents. By Theorem 2, his assessment then is equivalent to an assessment conforming to $\mathcal{H}^1_{-i}$. Thus, Player i plays some strategy from $B(\mathcal{H}^1_{-i})$. Again, by definition of hierarchical best response set and by Theorem 2, any strategy from $B(\mathcal{H}^0_{-i})$ can potentially be played by a level-2 interspecting Player i.

Suppose now that Player i is level-3 interspecting. He considers events corresponding to his opponents having interspection levels 0, 1, 2. Player i knows their Impeccability and knows that they also know the Impeccability of their opponents. Also, Player i knows that his opponents respect their opponents. Thus, he additionally concludes that for $j \neq i$, a level-2 interspecting Player j can play strategies precisely from $B(\mathcal{H}^1_{-j})$. Player i respects his opponents. By Theorem 2, his assessment then is equivalent to an assessment conforming to $\mathcal{H}^2_{-i}$. Thus, he plays some strategy from $B(\mathcal{H}^2_{-i})$. Also, by the same logic as above, any strategy from $B(\mathcal{H}^2_{-i})$ can potentially be played by a level-2 interspecting Player i.

Repeating this argument for higher levels of interspection and using more levels of Common Knowledge of Respect and of Impeccability, we will inductively prove the above Statement.

In particular, the Statement implies, that under Sufficient Interspection, the set of strategies the players can play is precisely the set of interspected rationalizable strategies. Q.E.D.

B.4 Proof of Theorem 4 (Iterative Admissibility)

Consider a two-player game Γ with interspection index L. Fully interspected hierarchies are $(\mathcal{H}^\infty_1, \mathcal{H}^\infty_2) = (\{H^j_1\}^0_{j=L}, \{H^j_2\}^0_{j=L})$. It is sufficient to show that for each $i = 1, 2$, and $l \geqslant 0$, the set H^j_i is precisely the set of Player i's strategies surviving j rounds of WDS elimination. The prove is by induction:

For $l = 0$ this is true.

Suppose this fact holds for all $l \leqslant M$. Prove for $l = M + 1$.

By definition, the set H^l_i is the hierarchical best response to $\mathcal{H}^{l-1}_i$. I.e., H^l_i is the set of strategies which are best responses to some full-support LPS assessments conforming to $\mathcal{H}^{l-1}_i = \{H^j_{-i}\}^0_{j=l-1}$.

Call an LPS assessment ρ_i a *simple assessment* on $\mathcal{H}^{l-1}_{-i}$, if ρ_i is a concatenation of LPS assessments with full supports on H^{l-1}_{-i}; H^{l-2}_{-i},..., H^0_{-i}.

In finite games, a strategy is a best response to an LPS conforming to $\mathcal{H}^{l-1}_{-i} = \{H^j_{-i}\}^0_{j=l-1}$, if and only this strategy is a best response to some simple LPS on $\mathcal{H}^{l-1}_{-i}$.

Suppose that σ^i is a best response to a simple LPS $(\rho^{l-1}_i; ...; \rho^0_i)$ on $\mathcal{H}^{l-1}_{-i}$. Then, σ^i is a $\mathcal{H}^{l-1}_{-i}$-best response. By induction assumption, σ^i survives $l-1$ rounds of WDS elimination. Yet, σ^i is a best response to ρ^{l-1}. Thus, σ is not weakly dominated on H^{l-1}_{-i}. Then, σ survives l rounds of WDS elimination.

On the other hand, suppose σ^i survives l rounds of WDS elimination. Then, there exist a full support measure ρ^{l-1}_i on $\mathcal{H}^{l-1}_{-i}$ such that σ^i is a best response to ρ^{l-1}_i. By induction assumption, there also exists a simple LPS $(\rho^{l-2}_i; ...; \rho^0_i)$ on $\mathcal{H}^{l-2}_{-i}$ such that σ^i is a best response to it. Then, σ^i is a best response to $(\rho^{l-1}_i; ...; \rho^0_i)$. Thus, σ^i is a $\mathcal{H}^{l-1}_{-i}$-best response. Q.E.D.

B.5 Proof of Theorem 5 (Backward Induction)

Note first that if a PI game Γ has no relevant ties, then interspected rationalizability predicts the unique terminal outcome (c.f. Battigalli (1997), Lemma 5). Indeed, consider the tree $I(\Gamma)$ consisting of all the nodes reachable under the play of interspected rationalizable strategies. If $I(\Gamma)$ has no decision nodes with more than one action, then it has the unique terminal outcome. Suppose there are adecision nodes with at least two actions. Call such nodes *nontrivial*. Consider any nontrivial decision node h in $I(\Gamma)$ such that all of his successors in $I(\Gamma)$ are trivial. The player acting in h then with sufficiently high interspection level predicts the outcomes reachable after each of his actions at h. As the game has no relevant ties, he will strictly prefer only one action. This contradict the definition of $I(\Gamma)$.

Now, as any interspected equilibrium implies interspected rationalizable outcome, it is sufficient to show that in any PI game Γ without chance nodes and without relevant ties, the outcome of any interspected equilibrium is the unique backward induction outcome.

By Corollary 3 from Appendix B.8, any Kohlberg and Mertens (1986)'s fully stable set of a game contains at least one interspected equilibrium. But, as is shown in the proof of Battigalli (1997)'s Theorem 4, in PI games without relevant ties the unique fully stable outcome coincides with the unique backward induction outcome. Q.E.D.

B.6 Proof of Theorem 6 (Equilibrium Existence, $N = 2$)

Introduce first the notion of constrained equilibrium:

Definition (Constrained Equilibrium)**.** *In a finite two player extensive-form game of perfect recall Γ a pair of randomized strategies (μ^1, μ^2) and a pair of LPS assessments (ρ_1, ρ_2) is called constrained equilibrium if:*

1. *each strategy in the support of μ^i is a constrained best response against assessment ρ_i, where the choice is constrained to the set of interspected rationalizable strategies $R_i(\Gamma)$;*
2. *first beliefs in the assessments coincide with the strategies actually played by the opponent:*

$$\lambda_1^i = \mu^{-i}, \text{ for } i = 1, 2$$

3. *assessments (ρ_1, ρ_2) conform to fully interspected hierarchies $(\mathcal{H}_2^\infty; \mathcal{H}_1^\infty)$.*

The difference between constrained and interspected equilibrium is that in a constrained equilibrium the players are only allowed to choose among interspeced rationalizable strategies.

Lemma 4. *Any constrained equilibrium is interspected equilibrium.*

Proof. Let $(\mu^1, \mu^2; \rho_1, \rho_2)$ be a constrained equilibrium. We need to show that any $\sigma^i \in supp(\mu^i)$ is a best response to ρ_i. Suppose not. Then there exist $\tilde{\sigma}^i \in \mathcal{S}_i$ such that $\tilde{\sigma}^i \succ_{\rho_i} \sigma^i$. But since σ^i is

a best response to ρ_i among rationalizable strategies, then $\tilde{\sigma}^i \notin \mathcal{R}_i$. On the other hand, $\tilde{\sigma}^i$ is not a best response to ρ_i, because ρ_i conforms to $\mathcal{H}_{-i}$. Thus, there exists $\hat{\sigma}^i \in \mathcal{R}_i$ such that $\hat{\sigma}^i \succ_{\rho_i} \tilde{\sigma}^i$. This leads to $\hat{\sigma}^i \succ_{\rho_i} \sigma^i$. Contradiction. Q.E.D. □

Thus, to prove Theorem 6 we only need to show that for any finite two-player extensive-form game of perfect recall, there exists at least one constrained equilibrium.

Consider any normal form G. Consider the pair of fully interspected hierarchies $(\mathcal{H}_1^\infty; \mathcal{H}_2^\infty)$. Let $\mathcal{H}_i^\infty = \{H_i^j\}_{j=L}^0$, for $i = 1, 2$, where L is the interspection index of the game G. Take $\epsilon > 0$. Consider the constrained game $G(\epsilon)$, in which players are only allowed to play randomized strategies (μ^1, μ^2) supported on rationalizable sets $(R_1(G), R_2(G)) = (H_1^L, H_2^L)$. The payoffs in $G(\epsilon)$ for the strategy pair $(\mu^1, \mu^2) \in \Delta(\mathcal{R}_1) \times \Delta(\mathcal{R}_2)$ are defined as payoffs in G for the strategy pair $\big(\mu^1(\epsilon), \mu^2(\epsilon)\big) \in \Delta(\mathcal{S}_1) \times \Delta(\mathcal{S}_2)$, where:

$$\mu^i(\epsilon) = \big(1 - \delta^i(\epsilon)\big) \cdot \mu^i + \sum_{j=L}^{0} \epsilon^{L-j+1} \cdot \left|H_i^j\right| \cdot U(H_i^j)$$

with $\left|H_i^j\right|$ being the number of elements in H_i^j, $U(H_i^j)$ being the uniform distribution on H_i^j, and $\delta^i(\epsilon) = \sum_{j=L}^{0} \epsilon^{L-j+1} \cdot \left|H_i^j\right|$. For sufficiently small $\epsilon > 0$, the game $G(\epsilon)$ is well defined.

Take a sequence $\epsilon_n \to 0+$ such that all $G(\epsilon_n)$ are well defined. By Nash's theorem, for each $n \in \mathbb{N}$, there exists at least one Nash equilibrium $\mathcal{E}\big(G(\epsilon_n)\big)$ in $G(\epsilon_n)$. Select one $\mathcal{E}_n$ for each n. From the sequence of $\mathcal{E}_n$ select a converging subsequence $\mathcal{E}_{n'} \to \mathcal{E}$. Then, $\mathcal{E}$ is a Nash equilibrium in G.

Show now that $\mathcal{E} = (\mu^1, \mu^2)$ corresponds to a constrained equilibrium $(\mu^1, \mu^2; \rho_1, \rho_2)$ of G. Consider the sequence of full-support measures $\big(\mu^1(\epsilon_{n'}), \mu^2(\epsilon_{n'})\big)$ corresponding to $\mathcal{E}_{n'}$. By Proposition 2 from BBD2, p. 84, we can further select a subsequence $\{n''\}$ such that there is an LPS $\hat{\rho}_1 = (p_1^1, ..., p_{K_1}^1)$ with full support on $\mathcal{S}_2$ and that $\mu^2(\epsilon_{n''})$ can be written as a nested convex combination $\mu^2(\epsilon_{n''}) = r_1(n'') \square \hat{\rho}_1$ for a sequence $r_1(n'') \in (0,1)^{K_1 - 1}$ with $r_1(n'') \to 0$. Applying this proposition one more time we can further select a subsequence $\{n'''\}$ such that there is an LPS $\hat{\rho}_2 = (p_1^2, ..., p_{K_2}^2)$ with full support on $\mathcal{S}_1$ and that $\mu^1(\epsilon_{n'''})$ can be written as a convex combination $\mu^1(\epsilon_{n'''}) = r_1(n''') \square \hat{\rho}_1$, for a sequence $r_2(n''') \in (0,1)^{K_2 - 1}$ with $r_2(n''') \to 0$.

Consider the subsequence $\{n'''\}$. Note first that any strategy $\sigma^i \in supp(\mu^i)$ is a constrained best response to all $r_i(n'') \square \hat{\rho}_i$ for sufficiently large n'''. As $r_i(n''') \to 0$, the strategy σ^i then is a constrained best response to $\hat{\rho}_i$. Thus, $\mathcal{E}$ can be represented as a constrained lexicographic Nash equilibrium $(\mu^1, \mu^2; \hat{\rho}_1, \hat{\rho}_2)$.

It remains to show that for each $i = 1, 2$, the LPS $\hat{\rho}_i$ conforms to fully interspected hierarchy $\mathcal{H}_{-i}$. Suppose that a strategy $a_{-i} \in \mathcal{S}_{-i}$ has a higher hierarchical index in $\mathcal{H}_{-i}$ than a strategy $b_{-i} \in \mathcal{S}_{-i}$. By definition of $\mu^{-i}(\epsilon_{n'''})$, for any n''' we get $\mu^{-i}(\epsilon_{n'''})[b_{-i}] < \epsilon_{n'''} \cdot \mu^{-i}(\epsilon_{n'''})[a_{-i}]$. Thus:

$$\frac{\mu^{-i}(\epsilon_{n'''})[b_{-i}]}{\mu^{-i}(\epsilon_{n'''})[a_{-i}]} \to 0, \text{when } n''' \to +\infty$$

But since $\mu^{-i}(\epsilon_{n'''}) = r_i(n''') \square \hat{\rho}_i$ with $r_i(n''') \to 0$, the strategy a_{-i} is covered in $\hat{\rho}_i$ strictly earlier

than b_{-i}. Q.E.D.

B.7 Proof of Theorems 7 and 9 (Equilibrium Properties)

Sequential Rationality. Any strategy played in an interspected equilibrium is a best response to a cautious assessment. Thus this strategy is initially rational. Sequential rationality of this strategy then follows from Theorem 1.

Admissibility. Any strategy played in an interspected equilibrium is a best response to a full-support LPS, an therefore is not weakly dominated.

Invariance. Note that if Γ and Γ' have the same reduced normal form, then $G(\Gamma')$ may be obtained from $G(\Gamma)$ by a finite sequence of the following four operations: Add/Delete a strategy payoff-equivalent to a given strategy in the game, Add/Delete a strategy payoff-equivalent to a convex combination of two strategies in the game. Thus, to prove that invariance holds, we only need to check that for any interspected equilibrium of Γ, there will be an outcome-equivalent interspected equilibrium in game Γ' obtained from Γ by a one-time application of one of the above operations.

Let Γ be an N-player game with normal form $G(\Gamma)$. Let $\mathcal{E}$ be an interspected equilibrium of Γ represented by a strongly independent common prior cautious assessment $\boldsymbol{P}$ on players' strategy profiles. Check one-by-one that for any of the mentioned above four operations there will be an interspected equilibrium in the new game Γ' inducing the same distribution over payoff-outcomes:

Case 1: Add a payoff-equivalent strategy. By strong independence of $\boldsymbol{P}$, there exists an LPS representation ρ of $\boldsymbol{P}$ and a limiting nested convex combination $\left\{r(n)\square\rho\right\}_{n\in\mathbb{N}}$ with $r(n)\to 0$ such that for each n, the measure $r(n)\square\rho=\underset{i=1}{\overset{N}{\times}} p_i(n)$ is a product measure on $\mathcal{S}=\underset{i=1}{\overset{N}{\times}}\mathcal{S}_i$. Suppose that Γ' is obtained from Γ by adding a duplicate $\hat{\sigma}^i$ for some strategy σ^i of Player 1. For each measure $r(n)\square\rho$, construct a new measure $\widehat{r(n)\square\rho}$ on $\mathcal{S}'=\underset{i=1}{\overset{N}{\times}}\mathcal{S}'_i$ by taking probability mass $p_i(n)(\sigma^i)$ and splitting it in half between σ^i and $\hat{\sigma}^i$. Take the resulting limiting sequence. By Proposition 2 from BBD2, select a subsequence n' such that $\widehat{r(n')\square\rho}$, with $n\geqslant 1$, can be represented as a limiting nested combination $\left\{r'(n')\square\rho'\right\}_{n\in\mathbb{N}}$. Then, the assessment $\boldsymbol{P'}$ on the set of strategy profiles, induced by ρ' in the new game Γ', satisfies strong independence and conforms to fully interspected hierarchies. Thus, $\boldsymbol{P'}$ represents an interspected equilibrium with the same distribution over payoff-outcomes as the equilibrium represented by $\boldsymbol{P}$.

Case 2: Delete a payoff-equivalent strategy. The proof is similar to Case 1.

Case 3: Add a strategy payoff-equivalent to a convex combination of two strategies. The proof is similar to Case 1. Only now one has to be careful to guarantee that the new assessment conforms to fully interspected hierarchies of the new game: a convex combination of two strategies may have interspection index lower than the smallest interspection index of those two strategies.

Case 4: Delete a strategy payoff-equivalent to a convex combination of two strategies. This case is similar to Case 2.

Forward induction. Let Γ be a two-player game with normal form $G(\Gamma)$. Let Γ' be the simplified game obtained from Γ by deleting all weakly dominated strategies of Player $-i$. Let $\mathcal{E} = (\mu^1, \mu^2; \rho_1, \rho_2)$ be an interspected equilibrium of Γ. Take the LPS $\hat{\rho}_i$, the beginning of ρ_i that covers all strategies in $B(\mathcal{H}_i^0(\Gamma))$. Note that $\hat{\rho}_i$ has full support on $\mathcal{S}'_{-i}$ in Γ'. But then $\hat{\rho}_i$ conforms to $\mathcal{H}_{-i}^{\infty}(\Gamma')$. Thus, $\mathcal{E}' = (\mu^i, \mu^{-i}; \hat{\rho}_i, \rho_{-i})$ is the interspected equilibrium induced by $\mathcal{E}$ in the simplified game Γ'.

Weak Forward Induction. Consider a multi-player game. Let $\boldsymbol{P}$ be a common prior assessment corresponding to an interspected equilibrium of the full game. Represent $\boldsymbol{P}$ by a limiting convex combination in which each measure is a product measure. Selecting subsequences if necessary, make this limiting nested convex combination consist of products of measures that are limiting nested convex combinations themselves. Truncate each of those measures by considering only the beginnings of LPSs covering rationalizable strategies. Notice that the product of these truncated limiting nested convex combinations corresponds to an interspected equilibrium of the reduced game. Q.E.D.

B.8 Proof of Theorem 8 (Equilibrium Existence)

The proof of equilibrium existence for the case of two-player games given in Appendix B.6 may be directly adapted for the current case. However, I present here a slightly different and less direct version of the proof. This version then is used in the proof of Theorem 5.

Consider an N-player extensive-form game of perfect recall Γ with normal form G. Consider fully interspected hierarchies $(\mathcal{H}_1^{\infty}; \mathcal{H}_2^{\infty})$. Let $\mathcal{H}_i^{\infty} = \{H_i^j\}_{j=L}^0$, for $i = 1, ..., N$, where L is the interspection index of the game G.

For any sufficiently small ϵ define *simple ϵ-perturbation* $G^S(\epsilon)$ of the game G by perturbing slightly the strategies in G similar to how it is done in the proof of Theorem 6. Namely, each strategy σ^i of each Player i in game G is replaced in $G^S(\epsilon)$ by strategy $\tilde{\sigma}^i(\epsilon)$:

$$\tilde{\sigma}^i(\epsilon) = \big(1 - \delta^i(\epsilon)\big) \cdot \sigma^i + \sum_{j=L}^{0} \epsilon^{L-j+1} \cdot \left|H_i^j\right| \cdot U(H_i^j)$$

with $\left|H_i^j\right|$ being the number of elements in H_i^j, $U(H_i^j)$ being the uniform distribution on H_i^j, and $\delta^i(\epsilon) = \sum_{j=L}^{0} \epsilon^{L-j+1} \cdot \left|H_i^j\right|$.

Thus, in any simply perturbed game $G^S(\epsilon)$ all strategies correspond to full-support perturbations of strategies in the original game.

Theorem 8 follows immediately from the following lemma. This lemma is also is useful in the proof of Theorem 5:

Lemma 5. *Let $\{G^S(\epsilon_n)\}$ be a sequence of simple perturbations of G with $\epsilon_n \to 0$. Let $\{\mathcal{E}_n\}$ be any sequence of Nash equilibria of the corresponding games $\{G^S(\epsilon_n)\}$. Then $\{\mathcal{E}_n\}$ contains a subsequence $\{\mathcal{E}_{\tilde{n}}\}$ converging to an interspected equilibrium of the unperturbed game G.*

Proof. The proof is similar to the proof from Appendix B.6. Therefore, I only present a sketch here.

Take any sequence of Nash equilibria $\{\mathcal{E}_n\}$ of simply perturbed games $\{G^S(\epsilon_n)\}$ with $\epsilon_n \to 0$. Each $\mathcal{E}_n$ induces an independently mixed-strategy profile $\mu_n = \overset{N}{\underset{i=1}{\times}} \mu_n^i \in \Delta^0\big(\overset{N}{\underset{i=1}{\times}} \mathcal{S}_i\big)$ in G. Select a subsequence n' such that $\{\mu'_n\}$ can be represented as a limiting nested convex combination corresponding to a strongly independent common prior assessment $\boldsymbol{P}$. Select consecutively for each of the players further subsequences such that the players' marginals are also represented by limiting nested convex combinations. Denote the final subsequence $\{\mu_{\tilde{n}}\}$.

Consider $\{\mu_{\tilde{n}}\}$. For each Player i, denote through $\mathcal{D}_i$ the set of strategies that are selected in $\{\mathcal{E}_{\tilde{n}}\}$ infinitely often. Prove now that all strategies in $\mathcal{D}_i$, $i = 1, ..., N$ are interspected rationalizable. Indeed, suppose not. Take a strategy with the lowest interspection index among all strategis from all of $\mathcal{D}_i$, $i = 1, ..., N$. Let this strategy be σ^i. Let the index of σ^i be $k_i < L$, where L is the interspection index of the game. Note then that in the limiting sequence $\{\mu'_n\}$ the marginal of Player i conforms to the k_i-step interspected hierarchy. As σ^i is a best response infinitely often to assessments from $\{\mathcal{E}_{\tilde{n}}\}$, σ^i is a hierarchical best response to $\mathcal{H}_{-i}^{k_i}$. But then the index of σ^i is at least $k_i + 1$. Contradiction.

Thus, all strategies selected in $\{\mathcal{E}_{\tilde{n}}\}$ infinitely often are interspected rationalizable. Then the limiting assessment $\boldsymbol{P}$ conforms to fully interspected hierarchies. Also, $\boldsymbol{P}$'s first measure μ_∞, the limit of $\{\mu_{\tilde{n}}\}$, is a best response to $\boldsymbol{P}$. I.e., the sequence $\{\mathcal{E}_{\tilde{n}}\}$ converges to $(\mu_\infty; \boldsymbol{P})$, an interspected equilibrium of G. Q.E.D.

□

Lemma 5 has an immediate corollary:

Corollary 3. *Any Kohlberg and Mertens (1986)'s fully stable set of a finite game contains at least one interspected equilibrium.*

References

Anscombe, F. J. and R. J. Aumann (1963). A definition of subjective probability. *The Annals of Mathematical Statistics 34*(1), 199–205.

Arieli, I. and R. J. Aumann (2013). The logic of backward induction. Working Paper.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica 55*(1), 1–18.

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

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

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

Battigalli, P. and M. Siniscalchi (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory 106*, 356–291.

Bernheim, D. B. (1984). Rationalizable strategic behavior. *Econometrica 52*(4), 1007–1028.

Blume, L., A. Brandenburger, and D. Dekel (1991a). Lexicographic probabilities and choice under uncertainty. *Econometrica 59*(1), 61–79.

Blume, L., A. Brandenburger, and D. Dekel (1991b). Lexicographic probabilities and equilibrium refinements. *Econometrica 59*(1), 81–98.

Brandenburger, A., A. Friedenberg, and J. Keisler (2008). Admissibility in games. *Econometrica 76*(2), 307–352.

Cho, I.-K. and D. M. Kreps (1987). Signaling games and stable equilibria. *Econometrica 102*(2), 179–222.

Fudenberg, D. and J. Tirole (2000). *Game Theory*. Massachusetts Institute of Technology.

Govindan, S. and R. Wilson (2009). On forward induction. *Econometrica 77*(1), 1–28.

Kohlberg, E. and J.-F. Mertens (1986). On the strategic stability of equilibria. *Econometrica 54*(5), 1003–1037.

Kreps, D. M. and G. Ramey (1987). Structural consistency, consictency, and sequential rationality. *Econometrica 55*(6), 1331–1348.

Kreps, D. M. and R. Wilson (1982). Sequential equilibria. *Econometrica 50*(4), 863–894.

Kuhn, H. (1953). *Extensive games and the problem of information. In Contributions to the Theory of Games, Vol. 2*. Princeton, NJ: Princeton University Press. 193-216.

Myerson, R. B. (1978). Refinements of the nash equilibrium concept. *International Journal of Game Theory 7*(2), 73–80.

Myerson, R. B. (1997). *Game Theory. Analysis of Conflict.* Cambridge, MA; Longon, England: Harvard University Press.

Nagel, R. (1994). Unraveling in guessing games: an experimental study. *American Economic Review 85*, 1313–1326.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica 52*(4), 1029–1050.

Reny, P. J. (1992). Backward induction, normal form perfection and explicable equilibria. *Journal of Economic Theory 60*(3), 627–649.

Reny, P. J. (1993). Common belief and the theory of games with perfect information. *Journal of Economic Theory 59*, 257–274.

Selten, R. (1975). Re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory 4*, 25–55.

Stahl, D. O. and P. W. Wilson (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization 25*, 309–327.

Stahl, D. O. and P. W. Wilson (1995). On player's models of other players: theory and experimental evidence. *Games and Economic Behavior 10*, 218–254.

Van Damme, E. (1989). Stable equilibria and forward induction. *Journal of Economic Theory 48*, 476–496.