Edinburgh School of Economics

Discussion Paper Series

Number 301

Tenable threats when Nash Equilibrium is the norm

József Sákovics
University of Edinburgh and Universitat de les Illes Balears

Françoise Forges
Université Paris-Dauphine

June 2021

Published by
School of Economics
University of Edinburgh
30 -31 Buccleuch Place
Edinburgh EH8 9JT
+44 (0)131 650 8361
https://www.ed.ac.uk/economics
@econedinburgh

Tenable threats when Nash Equilibrium is the norm*

Françoise Forges† and József Sákovics‡

June 11, 2021

Abstract

We formally assume that players in a game consider Nash Equilibrium (NE) the behavioral norm. In finite games of perfect information this leads to a refinement of NE: Faithful Nash Equilibrium (FNE). FNE is outcome equivalent to NE of the "trimmed" game, obtained by restricting the original tree to its NE paths. Thus, it always exists but it need not be unique. Iterating the norm ensures uniqueness of outcome. FNE may violate backward induction when subgame perfection requires play according to the SPE following a deviation from it. We thus provide an alternative view of tenable threats in equilibrium analysis.

JEL codes: C72, C73, D01, D83, D91.

1 Introduction

The role of game theory in social science is to explain/predict the behavior of people in situations of strategic interdependence. This is achieved by formulating cases as games,

*We are grateful for helpful comments from Roberto Burguet and (virtual) audiences at Università Bocconi and University of Edinburgh. Sákovics acknowledges financial support from the Spanish Government through a Beatriz Galindo grant (BG20/00079).

†Université Paris-Dauphine

‡Universitat de les Illes Balears and The University of Edinburgh

and proposing solution concepts that provide robust rules of behavior – preferably confirmed by empirical evidence. Thus, when analysts use Nash Equilibrium (NE) as the solution concept – as indeed the majority of them do – they implicitly posit that players normally behave according to some (any) NE. Taking them at their word, we hypothesize Nash behavior as the accepted norm and explore some of its behavioral consequences. In this paper, we do this in the context of finite extensive-form games of perfect information, without chance moves.

Apart from building on NE, we wish to construct our solution from bottom up, so we start with what the players – rather than the analyst – believe. Thus, we posit that each *player* believes that the others (also) are playing according to some (any) NE as long as he has no proof to the contrary. This is not a radical step, rather a natural extension of hypothesizing NE play, making explicit that the players are aware of the rules governing behavior.

We model the shared belief in NE play as a "faith".[1] If a player finds himself at an unexpected decision node – say, off the NE path he thought was being played –, he will maintain his belief that it was *some* NE play that led there. Rather than, say, supposing that another player has made a mistake. What differentiates faith in NE play from belief in a given NE is that the former is harder to lose[2] – as there are often multiple NE of a game. The fact that a given NE has not been followed does not imply that none of the NE has been followed.

Implicit in the above discussion is that faith in NE is an extensive-form concept: we are considering a version of sequential rationality that constrains the allowed – tenable, if you will – "threats" in the strategy profile. While his faith has not been contradicted, a player plays his part of (any) one of the NE he considers still feasible. That is, he behaves as if[3] best responding to a NE profile of the opponents that is consistent with the observed play. Thus, faith in NE is self-confirming: given their faith it is indeed a best response to behave according to the behavior prescribed by the faith. Only at decision nodes that are not reached by any NE does our player lose his faith, and remains without a guide on

[1] Battigalli and Siniscalchi (2002) refer to – a more technical version of – "faith" as "strong belief".

[2] See Blume, Brandenburger and Dekel (1991a, 1991b) for a formal analysis of lexicographic beliefs.

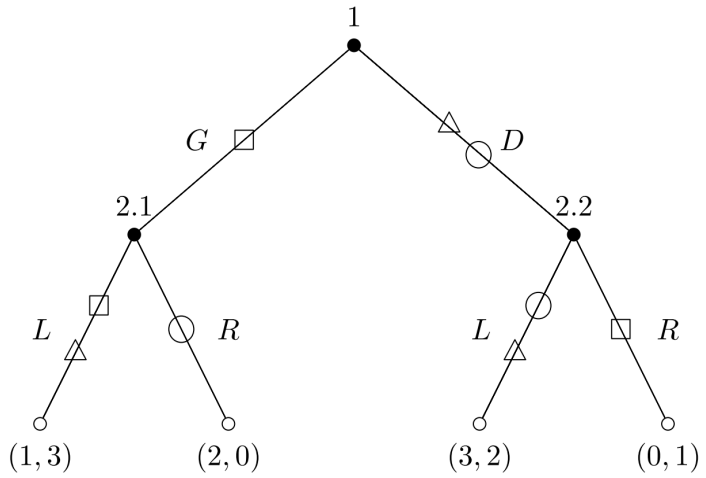
[3] We do not require that the player be explicitly/consciously optimizing at this point.

behavior. In this case, we see no clear justification to impose any restriction on NE (see Section 6.3 for a possible alternative).

Consider a game with multiple NE and a player who (initially) believes that one of those is being played. If he has faith in NE, it affects his behavior following a "deviation" of an opponent from the NE supposedly being played.[4] By the above, he must update his belief of which NE is being played and act according to the "new" equilibrium, effectively ruling the original NE profile out for including "non-tenable" threats.

To clarify ideas, consider the following example:

Example 1



This game has three (pure-strategy) NE, denoted by Δ*,* O *and* $\square$*. At player 1's decision node both available actions – G(auche) and D(roite) – are prescribed by some NE, so faith in NE does not restrict her choice of action. In contrast, player 2 is obliged by his faith to move L(eft) at both of his decision nodes, as the NE path(s) reaching either node prescribe that move. Since* O *and* $\square$ *prescribe R(ight) at nodes 2.1 and 2.2, respectively, they are*

[4] Note the tendentious terminology! Why should an unexpected move necessarily lead to the conclusion that the opponent has "misbehaved"? We suggest that it is perhaps more likely that the opponent *is* playing according to a NE, but not the one our player thought. After all, the coordination on a *specific* NE is the Achilles heel of the entire construct.

incompatible with faith in NE. Δ is the (only) NE that is compatible, as at node 2.1 (off-path) it prescribes L(eft) – the same move as the (only) NE whose path reaches 2.1 (Δ). Note that O is eliminated, despite having the same path as the equilibrium selected.

As the example shows, faith in NE can be instrumentalized without a need for a common belief of which equilibrium is being played, or in fact, for any belief. Neither is an iterative procedure needed. A player – or the analyst – can simply go through each NE profile in turn and eliminate those that include an action taken at a node on some NE path that does not agree with any of the on-path equilibrium actions at that node (in the example above, $>$). The remaining NE will be consistent with faith in NE. We call these Faithful Nash Equilibria (FNE). They satisfy – a version of – sequential rationality by ensuring that – whenever possible – play continues along an equilibrium path.

As it happens, the standard notion of sequential rationality for NE – Subgame-Perfect Equilibrium (SPE) – also selects $=$ in our example. However, its derivation could hardly be more different. Subgame perfection looks at subgames in isolation: nothing is "read into" the fact that play has arrived at a node. That is why it can be derived as the result of backward induction. This last observation could simply be taken as an advantage, what it indeed is. However, there is a price to be paid: we are led to ignore the fact that a deviation from a NE flies in the face of the assumption of rationality, the very basis for subgame-perfect play (c.f. Selten (1965)). In our view, it is not very congruent to require NE play in the subgame following a deviation from the NE in the entire game. Of course, we are not the first ones to point out this problem (see, for example, Rosenthal, 1981), but we do not know of any proposed method to overcome it in the refinement literature. FNE, on the other hand, does not impose any additional restriction once play veers off all possible NE paths (and thus faith in NE is lost), thus avoiding this critique.

Our refinement of NE is qualitatively different from what has previously been proposed in the literature (see van Damme, 2002 for the classical refinements, Kalai, 2020, for a "modern" one). It does not make assumptions about the players' rationality beyond their disposition to play Nash. Neither does it employ the analysis of perturbed games. In fact, it does not rely on anything but the players' knowledge of the set of NE profiles, not even the payoffs are needed beyond that. In a broad sense, it is a forward induction concept,

since it predicts the future based on past behavior. However, because of its "behavioral" – conditional on NE, payoff-independent – nature, it does so in a radically different way from the variants of forward induction proposed in the literature, starting with its "discovery" as a consequence of strategic stability in Kohlberg and Mertens (1986).[5]

Our contribution is closer to the epistemic strand of the literature. However, – unlike, say, Aumann and Brandenburger (1995)[6] – we are not looking for epistemic conditions that lead to NE, but the reverse: we take faith in NE as the "epistemic condition" and explore where that leads us.

As FNE is a backward-looking concept, it is intimately based on the NE of the entire game, and does not invoke optimality in all subgames, unlike not only SPE but even the different versions of extensive-form rationalizability (EFR) – see, Pearce (1984), Battigalli (1997), Battigalli and Siniscalchi (2002), Battigalli and Friedenberg (2012).[7] This difference is instrumental in making our most surprising result possible: FNE need not admit the backward induction solution.[8] This sets it apart from all solution concepts known to us (including EFR and its variants). We argue that this may (only) happen when at a node off the SPE path, faith in NE prescribes a different action from the subgame-perfect one. That is, exactly when SPE would restrict the continuation to NE behavior, despite play having contradicted it already. At the same time, FNE does not always eliminate SPE when the latter suffers from the "Rosenthal critique", rather it provides an alternative motivation for such a plan of play, which is not vulnerable to the same type of criticism.

[5] Other important references where forward induction is formalized include van Damme (1989), Stalnaker (1998), Battigalli and Siniscalchi (2002), Hillas and Kohlberg (2002), Govindan and Wilson (2009), Battigalli and Friedenberg (2012) and Catonini (2021).

[6] As we are considering certainty rather than knowledge of NE play, we are closer in spirit to, say, Ben-Porath (1997).

[7] This is also why FNE is not a τ theory *à la* Gul (1996). Otherwise, our approach shares a lot with his. Similarly, we differ from the theory of social situations of Greenberg (1990).

[8] This is all the more surprising in view of Reny (1992) and Battigalli (1997)'s results, which show that in games with perfect information and no relevant ties, EFR yields the SPE outcome (see Perea (2018) for a recent account). At the same time, the result is in line with the observation of Balkenborg and Winter (1997), that forward knowledge of rationality – something we do not assume (or imply) – is a necessary and sufficient condition for backward induction.

In the remainder of this paper – after introducing some notation – we derive the precise consequences of faith in NE, leading to our refinement, FNE. We demonstrate that it always exists for finite extensive form games with perfect information. We display examples showing that it need not be unique and that it can be at odds with backward induction. We also show that – in generic games – FNE strictly reduces the number of NE paths (when there are more than one) and as a result an iterative use of the concept leads to a unique prediction. A discussion of various extensions concludes.

2 Games of perfect information: some notation

Let us fix Γ, a finite game of perfect information – played by players $i \in \{1, 2, ..., I\}$ – without chance moves. Assume, w.l.o.g., that Γ has $K+1$ stages: the first K are decision stages and at stage $K+1$ are all the terminal nodes.[9] $\mathcal{N}_k$ denotes the set of nodes at stage k, $k = 1, ..., K+1$, with $\mathcal{N} = \cup_{k=1}^{K+1} \mathcal{N}_k$. $\mathcal{N}_1$ contains a single node: the root. At every (decision) node $n \in \mathcal{N}_k$, $k \leq K$, a unique player chooses one of finitely many actions, each leading to a different node in $\mathcal{N}_{k+1}$. A path is a sequence of connected nodes $n_1, ..., n_{K+1}$, from the root to a terminal one, with $n_k \in \mathcal{N}_k$, $k = 1, ..., K+1$.[10]

H^i is the set of decision nodes player i controls. Player i's (pure) strategy, s^i associates an action to each of her decision nodes, and a strategy profile is $s = (s^1, ..., s^I)$. Each strategy profile leads down a path. A strategy profile s is a NE if and only if s^i is a best response to s^{-i}. For simplicity, we consider exclusively pure-strategies and with NE we refer to pure-strategy NE.[11]

Let $\mathcal{E}$ denote the set of nodes that are on some NE path of Γ.[12] $\mathcal{E} \neq \emptyset$, since Γ has at least one NE. It will be useful to define the Trimmed Game (TG) resulting from the intersection between $\mathcal{N}$ and $\mathcal{E}$: eliminating from Γ the nodes not in $\mathcal{E}$ (and of all actions

[9] If some player can "finish" play earlier, insert a chain of decision nodes in later stages with a single available action at each.

[10] Nodes n_k and n_{k+1} are connected if the player moving at n_k has an action leading to n_{k+1}.

[11] Given that we analyze games of perfect information, disregarding mixed strategies has no major consequence.

[12] When no confusion can arise, we will also refer to the set of NE paths of Γ as $\mathcal{E}$.

leading to any node not in $\mathcal{E}$). It is straightforward to verify that TG is a well-defined $K+1$-stage game.

3 Characterization

Let us start by giving a precise definition of NE behavior combined with faith in NE.[13]

Definition 1 *A* Faithful Nash Equilibrium *(FNE) is a NE that at every node on some NE path prescribes play according to (any)one of the NE paths that reach that node. That is, a strategy profile, s, is a FNE if and only if it is a NE profile and for all $i \in \{1,2,...,I\}$ and $n \in \mathcal{E} \cap H^i$, there exists some NE profile, $\widehat{s}$, that has n on its path and satisfies $s^i(n) = \widehat{s}^i(n)$.*

In other words, as long as they can rationalize history by some NE, the players play their part of one of those equilibria. This differs from the standard (sequential) rationality postulate – maximization of expected payoffs conditional on strategic beliefs – in three ways. First, it is as if the players had a *set-valued* conditional – on history – conjecture about the strategy profiles that might be played by the other players.

Remark 1 *We do not require our players to have a probability distribution over the set of NE reaching each node. If they did have such beliefs, then we could construct a refinement of FNE where players would be required to choose among the available NE paths using, say, maximum likelihood (cf. Ortoleva, 2012)*[14] *or maximizing their expected payoff. However, we consider it a plus that FNE need not rely on such a distribution.*

The second novelty is that our postulate applies selectively: once faith in NE is lost – that is, outside $\mathcal{E}$ – FNE does not constrain the NE actions that can be chosen by a

[13] Recall that we restrict attention to finite games of perfect information (without chance nodes) and to pure-strategy NE.

[14] He proposes a hypothesis testing model over the paths leading to a non-trivial information set, so his method would not directly apply for the games of perfect information analyzed here.

player.[15] Nevertheless, when there are multiple NE – and thus a refinement is relevant – FNE does restrict behavior at some off-path nodes of a NE (the ones on the path of another NE). Thus, faith in NE does impose a degree of dynamic consistency.

Remark 2 *At a decision node on a NE's path, playing his part of that NE profile is of course a best response for a player, also conditional on having reached that subgame. That is, at every node in $\mathcal{E}$, FNE play implies NE play in the subgame starting with that node. This is a characteristic shared with SPE. However, it is not the result of backward induction, it is simply the consequence of having started with NE. Consequently, these NE of subgames need not be subgame perfect. See Section 5.*

Our definition lends itself to a calculation-free implementation – retaining a behavioral flavor – once the set of NE strategy profiles is identified. Thus, imposing faith in NE does not increase the complexity of the players' task: A player (or the analyst) can check, for each NE in turn, whether there is any decision node where some player chooses an action that is not taken in any of those NE whose path reaches that node.

Our first result provides a characterization of the set of FNE, by relating it to the NE of TG. As TG always has a NE, the proposition also proves existence of FNE.

Proposition 1 *The set of FNE paths of Γ is the set of NE paths of TG:*

i) *Restricted to $\mathcal{E}$, every FNE profile of Γ constitutes a NE of TG.*

ii) *Every NE profile of TG can be extended to $\mathcal{N}$ so that the extended profile is a FNE of Γ.*

This result is powerful, as it says that in order to predict the outcome of strategic interaction in Γ – according to FNE – it suffices to identify the NE paths of TG. This is non-trivial, as despite $\mathcal{E}$ corresponding to the union of NE paths of Γ, not all of these

[15] This characteristic avoids the critique of some solution concepts, including SPE, that assume "rational" behavior after histories incompatible with "rational" play. At the same time, it can also make FNE look "weak", see the discussion in Section 6.3.

paths constitute a NE path of TG. The reason is that in Γ a NE path may be supported by off-path actions taking play outside $\mathcal{E}$.

Proof. Suppose a FNE profile restricted to $\mathcal{E}$ is not a NE of TG. Then some player has a profitable deviation from it in TG, which is also available in Γ. In Γ, by faithfulness, all the other players will keep play within $\mathcal{E}$. That is, the deviation leads to the same outcome in Γ. Contradiction (to the profile being a (F)NE of Γ).

For ii), first note that every NE profile in TG, trivially, keeps play within $\mathcal{E}$, so if its extension is a NE it is a FNE. Now take a NE profile of TG, s, and extend it to s^* in Γ in the following way:

Note that $\mathcal{N} \setminus \mathcal{E}$ corresponds to the union of subgames that start with a node outwith $\mathcal{E}$, whose parent node is in $\mathcal{E}$. For each such starting node, there exists a NE path that reaches its parent node. Let s^* prescribe play in the corresponding subgame according to such a NE profile. Suppose there exists a profitable deviation by j with deviation path D. D must leave $\mathcal{E}$, since s is a NE in TG. Let the node at which D leaves $\mathcal{E}$ be $n_s \in H^j$ (the others, playing s^*_{-j}, keep play in $\mathcal{E}$).

In the ensuing subgame, s^* prescribes play according to a NE of Γ (with n_s on its path), call it R. Then j cannot profit from not following R at n_s. Thus, starting from n_s, we can replace the continuation of D with the remaining path of R, without decreasing j's payoff. Since the new deviation path is a NE path, it is in $\mathcal{E}$. Contradiction (to s being a NE of TG). ■

It is important to note that – in generic games – FNE is a strict refinement of NE: the number of FNE paths is strictly smaller than the number of NE paths (as long as the latter is larger than one). Put differently, despite being made up of NE paths of Γ, TG has strictly fewer NE paths than Γ. We prove – a more general version of – this result in Section 4 (Lemma 2).

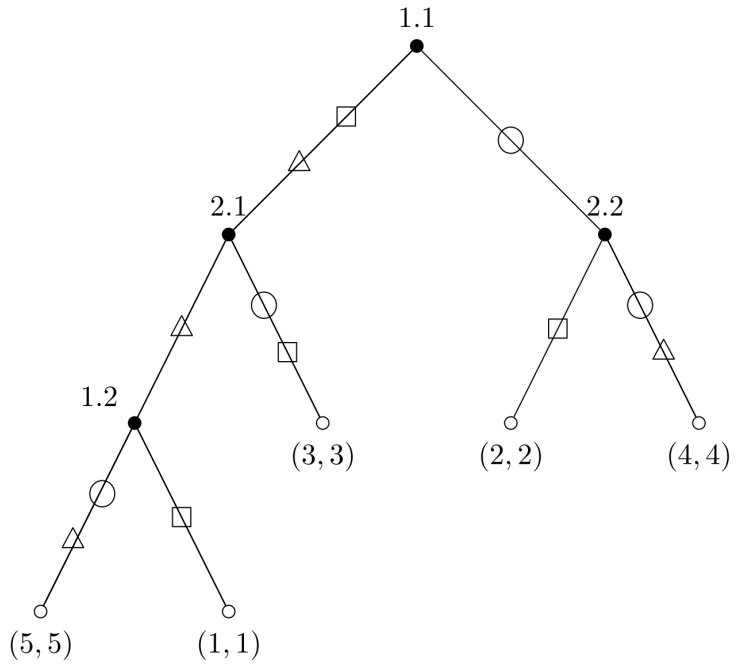
For simple games, like Example 1, FNE coincides with the unique prediction following from subgame perfection – or EFR, see Battigalli (1997) and Perea (2018).

Proposition 2 *If either Γ has a unique NE path, or it is generic and of two decision stages, then the FNE outcome is unique and it coincides with the SPE outcome.*

Proof. Note that in the last stage, on its own path, every NE has to agree with the SPE. Consequently, in the last stage on the SPE path, the FNE action also has to agree with the SPE (since there is no NE that would justify a different move). Suppose, for contradiction, that there exists an FNE path that is not the SPE path. Then, since – by the above – the second-stage actions would be identical (across the hypothetical FNE and the SPE) following both (supposedly different) first-stage equilibrium actions, the first-mover would strictly prefer (by genericity and SPE being a NE) the SPE action, implying that the FNE were not a NE. ■

In general, however, FNE need not lead to a unique outcome – and therefore it will not necessarily lead to the subgame-perfect outcome either. The following counter-example illustrates.

Example 2



This game (in which two players take turns to move) has 5 (pure) NE, with three distinct paths, but we only depict three of them, Δ, $\square$ and O. We now show that Δ and

O, which lead to different outcomes, are both FNE. Take Δ (the SPE) first. At the off-path decision node (2.2) it prescribes the same move as O, whose path this node is on. Next, take O. There are two off-path decision nodes (2.1 and 1.2). At both of these, O prescribes the same move as Δ, whose path both nodes are on. ($\square$ is not an FNE as it takes the "wrong" move at 1.2 and 2.2.)

4 Orthodox faith(s)

A reasonable question to ask is: What would happen if – say, because this paper were widely read – FNE became the new behavioral norm? If the prediction had been unique, the answer would have been obvious: no change. Multiplicity, however, raises the possibility of several different answers. In this section we show that iterating faith in NE makes it progressively strictly more restrictive ("orthodox"), always leading to a unique solution, eventually.

Let $r \in \mathbb{N}^+$ and let F^0NE denote a NE and $\mathcal{E}^0$ denote $\mathcal{E}$. We can now define iteratively stronger and stronger versions of faith in the result of faith.[16]

Definition 2 *A F^rNE is a $F^{r-1}NE$ that at every node in the set of all nodes on some $F^{r-1}NE$ path ($\mathcal{E}^{r-1}$) prescribes play according to (any)one of the $F^{r-1}NE$ paths that reach that node. That is, a strategy profile, s, is a F^rNE if and only if it is a $F^{r-1}NE$ profile and for all $i \in \{1, 2, ..., I\}$ and $n \in \mathcal{E}^{r-1} \cap H^i$, there exists some $F^{r-1}NE$ profile, $\widehat{s}$, that has n on its path and satisfies $s^i(n) = \widehat{s}^i(n)$.*

Let us start with a useful preliminary result, showing that, if two NE have different paths, one of them – on its own path but off the other's path – must prescribe a different action than the other, thus jeopardizing the other.[17]

Lemma 1 *Take any two NE with distinct paths of a generic, finite game of perfect information. In any subgame starting with a node on both of their paths there must exist a node that is on exactly one of their paths and where they disagree on the action chosen.*

[16] Note that for $r = 1$ this definition is literally the definition of FNE.

[17] Of course, there can be a third NE that protects that NE from elimination.

Proof. Take the first node in the subgame where they differ, n, and consider the two nodes where the differing action take the play. Both these nodes are on exactly one path. If their continuation paths agree from both nodes on, then, by genericity, the player controlling n would behave sub-optimally in one of the equilibria. Otherwise, there is a node in the subgame on exactly one of the paths where the actions chosen differ. ■

We can now prove the result that restricting faith to its consequences, progressively restricts accepted behavior.

Lemma 2 *Suppose a generic, finite game of perfect information has $d > 1$ F^{r-1}NE paths. Then it has at most $d - 1$ F^rNE paths.*

Proof. We will show that at least one F^{r-1}NE takes an action at a node off its path, but in $\mathcal{E}^{r-1}$, that disagrees with all F^{r-1}NE that reach that node. Take any two F^{r-1}NE with differing paths, call them #1 and #2. By Lemma 1, there is a node, n_k, that is on the path of #2 and is not on the path of #1, where they disagree. Then either #1 is not a F^rNE and the claim is true, or there exists another F^{r-1}NE, call it #3, that protects #1. Note that the path of #3 must agree with #2 till n_k, but, by Lemma 1, in the remaining subgame there is a node, n_{k+j}, that is on the path of exactly one of #2 or #3 and they disagree. Then either one of these equilibria is not a F^rNE and the claim is true, or there exists another F^{r-1}NE (it cannot be #1 as its path does not reach n_{k+j}), call it #4, that protects it. In the remaining subgame... eventually, we either run out of F^{r-1}NE or of stages in the game. ■

With this powerful result in hand we can show that the finite iteration of faith in F^rNE leads to a unique prediction.

Proposition 3 *Let Q denote the number of NE paths of Γ. There exists $R \in \{0, 1, ..., Q-1\}$ such that, for all $r \geq R$, there is a unique F^rNE path of Γ.*

Proof. Lemma 2 proves that for high enough r there is at most one F^rNE path. So, all that is left to show is existence. This follows from the proof of Proposition 1: we are progressively trimming the game by eliminating paths that are not NE paths of the

current version of the game restricted to $\mathcal{E}^r$. As each of these is a generic finite game of perfect information, it always has a (pure-strategy) NE. ■

As an illustration, note that the unique F^2NE path of the game depicted in Example 2 – that has two FNE paths – is $>$, the SPE path. As we argue in the next section this observation is not true in general.

5 Discrepancy with backward induction

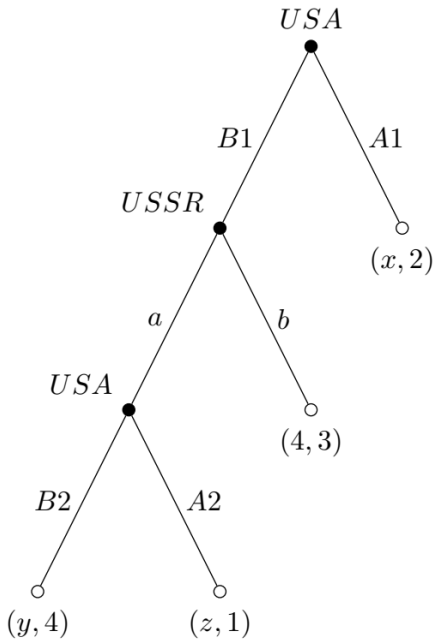
Since Γ can have multiple FNE, some of them clearly do not satisfy backward induction (since, generically, there is a unique SPE). But the discrepancy is even stronger. As Example 3 will illustrate, FNE need not be consistent with the backward induction outcome at all.

In order to understand the reason for this, note that when we restrict attention to TG, some NE paths of Γ may cease to be NE paths in TG (in Example 1, GL is no longer a NE in TG). When the eliminated path is the SPE path, the predictions must differ. Note that in order for the SPE path to be eliminated, it must be that there is a node on a NE path of Γ, where the (off-path) SPE action takes play outwith TG. In other words, the cause for the differing prediction lies *exactly* with the weakness of SPE: requiring Nash behavior following a deviation from it.[18]

Next, we present a stylized analysis – adapted from Harrington (2015) – of the Cuban missile crisis, where we argue that FNE may be a better explanation of what happened than SPE is.

[18] Nonetheless, FNE does not always eliminate SPE when the latter is questionable. It simply provides a better rationale for it.

Example 3



Given the USSR's military expansion in Cuba, the US can perform an immediate air strike (A1) or blockade (B1); in the latter case, the USSR can maintain (a) or withdraw (b) the missiles; if they maintain, the US has again to decide whether to perform an air strike (A2) or just stick to the blockade (B2). The USSR's preferences are relatively straightforward and are represented in the figure.The preferences of the US are less clear, except perhaps that their favorite outcome is no air strike and USSR's withdrawal (utility of 4). There are six possible preference orderings of the other three outcomes. In five of them, FNE and SPE agree.[19] *However, when $x > y > z$, SPE "incorrectly" predicts an immediate air strike (A1aB2), while FNE predicts the observed history: B1b(A2).*

It is easy to make an argument that the preferences that lead to the discrepancy are not unreasonable. Some US generals were convinced that any air strike should take place

[19] When z is on top – there is a unique NE and – the prediction agrees with history: B1b(A2), this is the showcased version in Harrington (2015). When y is on top the common prediction is B1b(B2), but that is not surprising as this would be a case when A2 is not a credible threat. When $x > z > y$, the agreed prediction is again B1b(A2), out of three NE.

by surprise. In addition, not withdrawing could be interpreted as an expectation that air strike would not be chosen by US, signalling that USSR would retaliate (probably in Turkey).

6 Discussion

We have put forward an unusual combination of collective rationality, embodied as equilibrium behavior, and (selective) individual rationality – maximization of expected payoffs given "allowable" beliefs. We are also combining strategic-form reasoning (NE), together with extensive-form reasoning (forward induction). The connecting element we take from "empirical" observation: players in a game often do – and even more often are supposed to – behave according to NE. As a result of this, we have proposed a novel, behavioral refinement of NE, based on the assumption that indeed NE is the behavioral norm: the players are known to strive to play according to a NE whenever possible. Note that the players do not need to "update" their faith given our results: their faith in NE will not be contradicted if play always results in an FNE outcome. At the same time, if players did become more and more orthodox in their faith, it would lead to a unique prediction (in generic games).

6.1 More on the relationship with SPE

The most powerful observation about FNE play is that it may be inconsistent with SPE. Importantly, this happens always when SPE is (most) vulnerable to the Rosenthal critique: off all NE paths. This observation opens the possibility for considering a hybrid version of FNE, where we additionally impose subgame perfection (only) within $\mathcal{E}$. That is, we could require that a FNE restricted to the Trimmed Game be the SPE of it. From Proposition 1 it is clear that this is feasible, moreover – under genericity – it would predict a unique outcome.[20]

[20] Unsurprisingly, examples show that this solution need not coincide with the unique solution that we arrive at by iterating faith as in Proposition 3.

6.2 Relationship with EFR

Battigalli and Friedenberg (2012) soften the stark result of Reny (1992) and Battigalli (1997) that in generic games of perfect information dynamic rationality leads to the same outcome as backward induction. They explain that EFR – as in Pearce (1984) and Battigalli (1997) – corresponds to one among many possible "extensive form best responses sets" (EFBRS), which is appropriate when the analyst has absolutely no idea of the environment within which the game is played. In practice however, there may be a context to the strategic situation at hand, so that there may be an interest to study a game relative to different type structures. Which EFBRS obtains depends then on the given type structure. They show that in generic games of perfect information, EFBRS's outcomes are always NE outcomes. They do not manage in establishing the converse, but still identify a class of NE (which contains the SPE) that induce an EFBRS. This gap leaves room for the conjecture that FNE are always part of some EFBRS. While this would clarify the relationship between FNE and EFR-like concepts, it would not affect the specificity of the FNE concept emphasized above, in particular, that a SPE may not be a FNE.

6.3 Locally undominated strategies

Since FNE imposes no restrictions outside TG, players may play strategies that are strictly dominated at the beginning of a subgame.[21] While this potential behavior occurs when the solution imposes no restrictions on the (NE) strategies – that is, we do not run into the paradox that SPE does – one might argue that FNE ensures an insufficient amount of sequential rationality. This complaint could be resolved at the price of "reforming" faith in NE: we could restrict attention to (faith in) NE with locally undominated strategies (LUNE), and define Reformed FNE as the subset of LUNE that, when on a LUNE path, always continues along one of them. It is easy to see that all our results would continue to hold,[22] including the discrepancy with SPE. In other words, if we believe that the solution

[21] And/or may expect other players to do so.

[22] Interestingly, examples show that the sets of FNE and of RFNE are not nested.

concept should not include NE that are not LUNE then we can model the norm that way and the analysis still applies. In the end, the choice of norm is an empirical question, we are not perorming a normative analysis here.

6.4 Pre-play communication and commitment

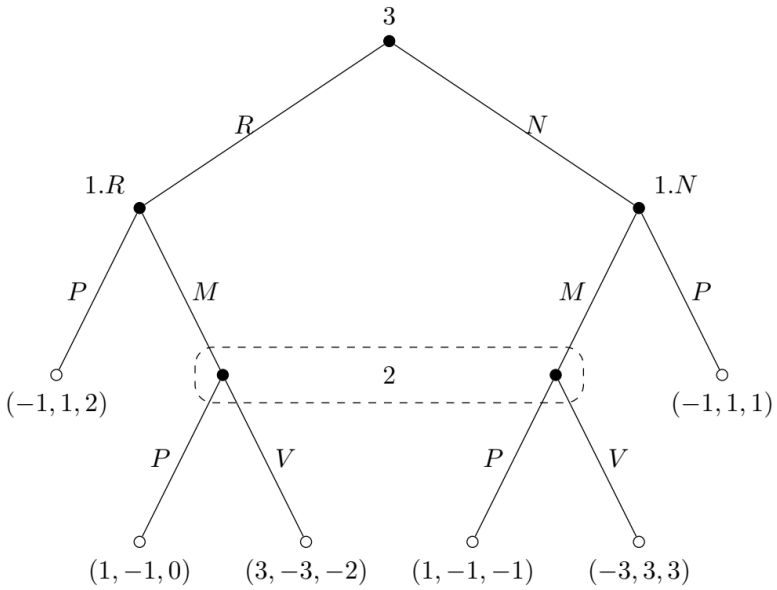
We do not consider the possibility of pre-play communication: we accept the game Γ as the best possible description of the strategic situation. Catonini (2021) does look at this question – not within the framework of NE refinement – incorporating updating beliefs about the compliance with a pre-play agreement. He also obtains that SPE may not be the best prediction (the non-empty set of self-enforcing agreement might not include the SPE outcome). However, a straightforward extension of FNE to games with simultaneous moves would prescribe SPE in examples where Catonini does not.

6.5 Imperfect information

While it is beyond the scope of this paper, let us briefly consider games of imperfect information. Putting aside the need to consider mixed strategies, there are additional complications. There are two obvious ways to extend our analysis, both of them leading to difficulties. The simplistic method could be to stick to our literal definition and to ignore non-singleton information sets when identifying the set of FNE. This would be similar, but not identical, to what happens in the case of subgame perfection, where only proper subgames are checked. The practical advantage of this method would be that it does not affect existence, but at the cost of not taking into account relevant decisions. The superior method is based on the observation that, unlike SPE[23], FNE can be straightforwardly extended to imperfect-information games by replacing "decision node" by "information set" in the definition. While this approach seems more satisfactory, unfortunately it would lead to the loss of existence, even when a pure-strategy NE exists and there are no chance nodes, as our last example illustrates. Note that, following Harsányi, this may imply lack of existence for certain games of incomplete information as well.

[23] In the case of SPE we need to define Perfect Bayesian Equilibrium, what is a major qualitative step.

Example 4



This example is inspired by a simple game of poker, with player 3 instead of chance. There are two pure NE: (MM,P,R) and (MP,V,N). The first one does not reach player 1's right node (1.N). The only NE reaching this node is (MP,V,N), which prescribes player 1 to choose P instead of M. Hence, (MM,P,R) is not a FNE. The second NE, (MP,V,N), does not reach player 2's information set. The only NE reaching it is (MM,P,R), which prescribes player 2 to choose P instead of V. Hence, again (MP,V,N) is not a FNE.

References

[1] Aumann, R., & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 1161-1180.

[2] Balkenborg, D. & Winter, E. (1997). A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics*, 27, 325-345.

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

[4] Battigalli, P. & Friedenberg, A. (2012). Forward induction reasoning revisited. *Theoretical Economics* 7, 57-98.

[5] Battigalli, P., & Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2), 356-391.

[6] Ben-Porath, E. (1997). Rationality, Nash equilibrium and backwards induction in perfect-information games. *Review of Economic Studies*, 64(1), 23-46.

[7] Blume, L., Brandenburger, A., & Dekel, E. (1991a). Lexicographic probabilities and choice under uncertainty. *Econometrica*, 61-79.

[8] Blume, L., Brandenburger, A., & Dekel, E. (1991b). Lexicographic probabilities and equilibrium refinements. *Econometrica*, 81-98.

[9] Catonini, E. (2021). Self-enforcing agreements and forward induction reasoning. *Review of Economic Studies*, 88, 610-642.

[10] van Damme, E. (2002). Strategic equilibrium. Chapter 41 in (Aumann and Hart eds.) *Handbook of game theory with economic applications*, 3, 1521-1596.

[11] van Damme, E. (1989). Stable equilibria and forward induction. *Journal of Economic Theory*, 48(2), 476-496.

[12] Govindan, S., & Wilson, R. (2009). On forward induction. *Econometrica*, 77(1), 1-28.

[13] Greenberg, J. (1990). *The theory of social situations: an alternative game-theoretic approach*. Cambridge University Press.

[14] Gul, F. (1996). Rationality and coherent theories of strategic behavior. *Journal of Economic Theory*, 70(1), 1-31.

[15] Harrington, J. (2015). *Games, strategies and decision making*. Worth Publishers, NY.

[16] Hillas, J., & Kohlberg, E. (2002). Foundations of strategic equilibrium. *Handbook of Game Theory with Economic Applications*, 3, 1597-1663.

[17] Kalai, E. (2020). Viable Nash equilibria: Formation and defection. Preprint, Northwestern University, February.

[18] Kohlberg, E., & Mertens, J-F. (1986). On the strategic stability of equilibria. *Econometrica*, 1003-1037.

[19] Ortoleva, P. (2012). Modeling the change of paradigm: Non-Bayesian reactions to unexpected news. *American Economic Review*, 102(6), 2410-2436.

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

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

[22] Reny, P.J. (1992). Backward induction, normal form perfection and explicable equilibria. *Econometrica,* 60(3), 627-649.

[23] Rosenthal, R. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, 25, 92-100.

[24] Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit: Teil i: Bestimmung des dynamischen preisgleichgewichts. *Journal of Institutional and Theoretical Economics*, (H. 2), 301-324.

[25] Stalnaker, R. (1998). Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36(1), 31-56.