Discussion Paper Series – CRC TR 224

Discussion Paper No. 702
Project B 02

Fully Self-Justifiable Outcomes

Francesc Dilmé[1]

October 2025

[1] University of Bonn. Email: fdilme@uni-bonn.de.

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Collaborative Research Center Transregio 224 - www.crctr224.de
Rheinische Friedrich-Wilhelms-Universität Bonn - Universität Mannheim

Fully Self-Justifiable Outcomes

Francesc Dilmé

An equilibrium outcome of a game in extensive form is *fully self-justifiable* if it is supported by justifiable equilibria (McLennan, 1985) regardless of the order in which actions implausible under the given outcome are excluded. We show that the set of fully self-justifiable outcomes is non-empty and contains the set of sequentially stable outcomes (Dilmé, 2024). In signaling games, fully self-justifiable outcomes pass all the selection criteria in Cho and Kreps (1987). Full self-justifiability allows for the systematic use of the logic of selection criteria in signaling games to select equilibria in any finite extensive form game.

Kohlberg and Mertens (1986) introduced (KM-)stable sets of equilibria, a concept that has been pivotal owing to its desirable features and robust selection power across different classes of games. Roughly speaking, a set of Nash equilibria is KM-stable if it is minimal with respect to the property that, for any vanishing sequence of normal-form trembles, there is a sequence of Nash equilibria approaching the set. KM-stable sets of equilibria exist for all games and have desirable properties (they satisfy forward induction, iterated dominance, and invariance).

KM-stability is nevertheless difficult to use in practice for two main reasons: proving or disproving the robustness of a given set of equilibria against all possible perturbations is often difficult, and set-valued concepts are difficult to manipulate and compare across games or parameter values. Numerous equilibrium concepts and selection criteria have since been introduced to both ease the identification of stable equilibria and study the effect of requiring plausibility conditions on off-path behavior without the need to explicitly consider sequences of perturbations and Nash equilibria. While some of these concepts are commonly used in some classes of games, such as signaling games, their relationship to stability is often unclear, and they are often not consistently applied across different types of games.

September 26, 2025. University of Bonn. fdilme@uni-bonn.de. This work was funded by a grant from the European Research Council (ERC 949465). I thank Tilman Börgers, Tommaso Denti, Jon X. Eguia, Amanda Friedenberg, Stephan Lauermann, David Miller, and Georg Nöldeke for their helpful comments. I am also grateful for input from participants in seminars at Michigan State University, New York University, Pennsylvania State University, Stony Brook University, University of Bonn, and the University of Michigan. Support from the German Research Foundation (DFG) through CRC TR 224 (Project B02) and under Germany's Excellence Strategy – EXC 2126/1–390838866 and EXC 2047–390685813 – is gratefully acknowledged.

$$\begin{tikzcd}
& \text{sequentially stable (Dilmé, 2024)} \arrow[d, Rightarrow] & \\
& \textbf{fully self-justifiable} \arrow[r, Rightarrow] \arrow[dl, Rightarrow] \arrow[dr, Rightarrow] & \text{forward induction (Cho, 1987)} \\
\textbf{fully justifiable} \arrow[d, Rightarrow] & & \textbf{self-justifiable} \arrow[d, Rightarrow, "\text{(in signaling games)}"] \\
\text{justifiable (McLennan, 1985)} & & \text{IC, D1, D2, NWBR (Cho and Kreps, 1987)} \\
& & \text{(universal) divinity (Banks and Sobel, 1987)}
\end{tikzcd}$$

Figure 1

We introduce *fully self-justifiable outcomes*, a solution concept obtained by combining the selection procedures proposed by Cho and Kreps (1987) for signaling games with the procedure of McLennan (1985) for obtaining equilibria with justifiable beliefs. We show that fully self-justifiable outcomes satisfy several previously defined selection criteria based on iterated equilibrium domination (see Figure 1). We use full self-justifiability to extend the signaling-games selection criteria to all games in extensive form. In addition, we show that all sequentially stable outcomes (Dilmé, 2024) are fully self-justifiable; thus, fully self-justifiable outcomes exist in all finite games in extensive form, and provide a powerful tool and a new foundation for the study of stable behavior.

We briefly recall the procedure used by McLennan (1985) to obtain justifiable equilibria. In the first step, one identifies all *useless* actions—that is, those which are never a weak best response under any sequential equilibrium (i.e., in each sequential equilibrium there is another action delivering a strictly higher continuation payoff). In the second step, one looks for *second-order useless* actions—that is, those which are never a weak best response under any sequential equilibrium assigning probability zero to histories that include more useless actions than the other histories in the same information set. The procedure continues iteratively, with histories compared lexicographically according to the orders of uselessness of the actions they contain. A sequential equilibrium is then *justifiable* if its belief system satisfies the conditions on the relative likelihoods of histories that result from this procedure. McLennan shows that justifiable equilibria exist in all games. Still, justifiability has limited selection power in many games, such as signaling games, as most actions are weak best responses under some sequential equilibrium.

We define our basic equilibrium selection procedure by combining the procedure of McLennan (1985) with the tests used by Cho and Kreps (1987) to determine the internal consistency of an outcome in signaling games. Fixing an outcome ω, in the first step one identifies and excludes all actions which are *useless under* ω—that is, those which are never a weak best response under any sequential equilibrium *with outcome* ω. In the second step, one excludes actions which are *second-order useless under* ω—that is, those which are never a weak best response under any sequential equilibrium with outcome ω assigning probability zero to histories that include more useless actions than the other histories in the same information set. This process is repeated until no further actions can be excluded. A sequential equilibrium with outcome ω is called *self-justifiable* if its belief system satisfies the conditions on the relative likelihoods of histories that result from this procedure (using ω).

We show that all sequentially stable outcomes are self-justifiable (i.e., they are outcomes of self-justifiable equilibria); hence self-justifiable outcomes exist in all games. We also observe that self-justifiability often has more selection power than justifiability, but that this is not always the case: we give an example of a game with a self-justifiable outcome that is not justifiable.

We then impose an additional robustness property: independence from the order in which implausible actions are excluded.[1] We define *fully justifiable* and *fully self-justifiable* outcomes as those supported by justifiable and self-justifiable equilibria, respectively, independently of the order of exclusion of implausible actions.[2] Of all the criteria defined, we show that full self-justifiability has the greatest selection power: fully self-justifiable outcomes are both fully justifiable and self-justifiable, and they satisfy several other selection criteria that have been proposed in the literature (see below). (In addition, we show that fully justifiable outcomes are justifiable.) Moreover, all sequentially stable outcomes are fully self-justifiable; hence every game has a fully self-justifiable outcome.

Through a series of examples, we illustrate the advantages of studying full justifiability and full self-justifiability, rather than justifiability and self-justifiability, in applications. Because of the added flexibility in the exclusion order, it is often easier to rule out the full justifiability or full self-justifiability of an outcome (and thus its sequential stability) than to rule out its justifiability or self-justifiability. On the other hand, it may be more difficult to prove that an outcome *is* fully justifiable or fully self-justifiable than that it is justifiable or self-justifiable.

As an application of our results, we show how certain forms of reasoning heretofore used in the analysis of signaling games to identify behavior robust to trembles can be generalized to all games in extensive form. We first show that in signaling games, while justifiability is of limited use, self-justifiable and fully self-justifiable outcomes satisfy all of the selection criteria proposed in Cho and Kreps (1987) and Banks and Sobel (1987). We then discuss two natural ways to

[1] In the procedure defining self-justifiability, *all* implausible actions must be excluded at each step. This condition ensures that the procedure is uniquely defined; however, it is not an essential element of the intuitive reasoning behind the iterated exclusion process. Independence from the order of exclusion guarantees independence from the manner in which this reasoning is implemented.

[2] We focus on fully justifiable and fully self-justifiable outcomes, rather than assessments, because the former exist in all games while the latter may not: different orders of exclusion may lead to different sets of sequential equilibria (see Example 3.1), but we show that these sets always contain common on-path behavior.

use the logic behind the exclusion of type–message pairs in the analysis of signaling games to identify fully justifiable outcomes in any finite game in extensive form. The first way is to apply this logic directly in the game, but excluding actions instead of type–message pairs. The second is to transform parts of the game into signaling games, then apply the selection criteria for signaling games to the latter. This second approach simplifies the analysis of the original game by analyzing its subgames separately or by replacing some of its players with *agents*. Such approaches can be used, for example, to select equilibria in signaling games with multiple senders, multiple receivers, hidden investment, or preemptive offers.

The framework developed in this paper presents different equilibrium notions that combine intuitiveness and simplicity. In our view, the most appealing of these is full self-justifiability, because of both its selection power and its usefulness in applications. Our results provide an additional foundation for the study of sequential stability and constitute tools that make it possible either to establish or to rule out the sequential stability of an outcome without needing to analyze sequences of strategy profiles.

Literature review. We see our work as combining the two main approaches used in the literature to select equilibria with plausible belief systems. The first approach, that of McLennan (1985), is to iteratively exclude implausible actions without fixing a particular equilibrium.[3] The second is to fix an equilibrium (or outcome) and assess its internal plausibility by excluding actions that fail a consistency test. For example, Cho (1987) defines forward induction equilibria by excluding "bad deviations" that are available on the path of a given equilibrium (see Section A.2 for a detailed discussion).[4] In signaling games, selection criteria such as the Intuitive Criterion, D1, D2, and Never-a-Weak-Best-Response (introduced in Cho and Kreps, 1987) consist in excluding from the support of beliefs the types that are implausible for a given outcome (where the meaning of "implausible" depends on the criterion; see Dilmé, 2025, for a study of the iterated application of these criteria). Similarly, Banks and Sobel (1987) define divine and universally divine equilibria as those possessing belief systems that survive a certain iterative procedure.[5]

[3] Similar procedures have been proposed in which actions or strategies are removed from the game rather than excluded (i.e., instead of being ruled out from plausible equilibrium behavior). For example, rationalizability (Moulin, 1979; Bernheim, 1984; Pearce, 1984) and interim correlated rationalizability (Battigalli and Siniscalchi, 2003; Dekel et al., 2007) are based on the iterated removal of actions. We discuss the relationship to proper equilibria (Myerson, 1978) in Section A.4.

[4] Govindan and Wilson (2009) consider a different definition of forward induction, based on a one-step procedure that excludes pure strategies that are not relevant for any weakly sequential equilibrium with a given outcome. See Section A.3 for a discussion of the relationship between their concept and that of fully self-justifiable outcomes.

[5] There are other approaches to equilibrium selection. For example, Grossman and Perry (1986) define *perfect sequential equilibria* by first extending the set of strategies to "metastrategies" that specify actions as a function of beliefs. Alternatively, for signaling games, Mailath et al. (1993) define *undefeated equilibria*, which are a refinement of pure-strategy sequential equilibria, and study them in a class of one-dimensional signaling settings.

Our contribution to this literature is to provide a unifying selection procedure—one that is simple and yet has significant selection power—and to relate it to other equilibrium concepts. Notably, we show that full self-justifiability refines the selection criteria in McLennan (1985), Cho (1987), Banks and Sobel (1987), and Cho and Kreps (1987). That is, a fully self-justifiable outcome is the outcome of a justifiable equilibrium, a (universally) divine equilibrium, and a forward induction equilibrium, and also passes (the iterated applications of) the Intuitive Criterion, D1, D2, and Never-a-Weak-Best-Response (NWBR). We also illustrate how the logic used by Cho and Kreps (1987) to select equilibria of signaling games can be applied to any game in extensive form.

In addition, our work strengthens the connection between the aforementioned selection criteria and the property of stability against trembles of the players (Kohlberg and Mertens, 1986; Dilmé, 2024).[6] Specifically, some of these criteria have been seen as tools for identifying behavior that is robust to perturbations of the game (see discussions in McLennan, 1985, Cho, 1987, Banks and Sobel, 1987, and Cho and Kreps, 1987). In this paper we show that sequentially stable outcomes are fully self-justifiable; hence they pass all of the earlier selection criteria. (Figure 1 summarizes the relationships between all of the criteria considered.) Thus, fully self-justifiable outcomes not only exist in all games, but also provide a simple way to investigate whether an outcome is sequentially stable without explicitly using tremble sequences. The fact that sequentially stable outcomes (which are defined by requiring robustness to perturbations of the game) are fully self-justifiable adds plausibility to the notion of sequential stability (in terms of both self-consistency and robustness to any process of iterated exclusion).

The rest of the paper is organized as follows. In Section 1, we introduce the notation. In Section 2, we define self-justifiable outcomes and compare them to the justifiable equilibria of McLennan (1985). In Section 3, we introduce the process of iterated exclusion of never weak best responses, define fully self-justifiable and fully justifiable outcomes, and study their main properties. Section 4 discusses the implications of full justifiability in games with signaling. Section 5 concludes. Appendix A contains a comparison between full self-justifiability and the forward induction criterion in Cho (1987) and Govindan and Wilson (2009), as well as additional examples. Appendix B contains the proofs of all results.

1 Extensive form, sequential equilibria, and sequential stability

1.1 Games in extensive form

We begin by providing the definition and notation for a game in extensive form with perfect recall.

[6] Refinements of sequential equilibria are suited to identify sequentially stable outcomes (as they themselves refine sequential equilibrium outcomes) than outcomes of KM-stable sets (as they may not exist or be sequential equilibrium outcomes).

A (finite) *game* $G := \langle A, H, \mathcal{I}, N, \iota, \pi, u\rangle$ has the following components: (1) A finite set of *actions* A. (2) A finite set of *histories* H. Here, a history is a finite sequence of actions $h \equiv (h_j)_{j=1}^{|h|}$ (note that $|h|$ denotes the length of history h), and the set H has the property that if $h \equiv (h_j)_{j=1}^{|h|} \in H$ with $|h| > 0$, then $(h_j)_{j=1}^{|h|-1} \in H$ as well. (In particular, $\emptyset =: (h_j)_{j=1}^{0} \in H$.) The set of *terminal histories* is denoted by Z. (3) An *information partition* $\mathcal{I}$, that is, a partition of $H \backslash Z$ such that there is a partition $\{A^I | I \in \mathcal{I}\}$ of A with the property that, for each $I \in \mathcal{I}$ and $h \in H$, we have $(h, a) \in H$ for some $a \in A^I$ if and only if $h \in I$. The elements of $\mathcal{I}$ are called *information sets*.[7] (4) A finite set of *players* $N \not\ni 0$. (5) A *player assignment* $\iota : \mathcal{I} \to N \cup \{0\}$, assigning each information set either to a player or to nature (represented by 0), such that there is perfect recall.[8] (6) A *strategy by nature* $\pi : \cup_{I \in \iota^{-1}(\{0\})} A^I \to (0, 1]$ satisfying $\sum_{a \in A^I} \pi(a) = 1$ for each $I \in \iota^{-1}(\{0\})$. (7) For each player $i \in N$, a (von Neumann–Morgenstern) *payoff function* $u_i : Z \to \mathbb{R}$.

A *strategy profile* is a map $\sigma : A \to [0, 1]$ such that $\sum_{a \in A^I} \sigma(a) = 1$ for all $I \in \mathcal{I}$ (i.e., it is a probability distribution for each set of actions available at each information set) and $\sigma(a) = \pi(a)$ for all a played by nature (i.e., nature plays according to π). We let Σ be the set of strategy profiles. An *outcome* ω (of G) is a probability distribution over terminal histories. We use $\Omega := \Delta(Z)$ to denote the set of outcomes. Each strategy profile $\sigma \in \Sigma$ generates a unique outcome ω^σ, where each terminal history $z \in Z$ is assigned probability $\omega^\sigma((a_j)_{j=1}^{J}) := \prod_{j=1}^{|z|} \sigma(z_j) \in [0, 1]$.

1.2 Sequential equilibria and sequentially stable outcomes

Our analysis will be focused on sequential equilibria and sequentially stable outcomes. We now briefly review these two concepts.

Kreps and Wilson (1982) defined a *belief system* as a map μ assigning a probability $\mu(h) \in [0, 1]$ to each non-terminal history $h \in H \backslash Z$, in such a way that $\sum_{h \in I} \mu(h) = 1$ for all $I \in \mathcal{I}$. They defined a *sequential equilibrium* as a pair consisting of a belief system μ and a strategy profile σ that is *consistent* (i.e., μ can be obtained as the limit of the beliefs corresponding to a fully-mixed sequence $(\sigma_n) \to \sigma$) and *sequentially rational* (i.e., an action a receives positive probability under σ only if it maximizes the continuation payoff at I^a given σ and μ). We denote the set of sequential equilibria with outcome ω by SE_ω.

Dilmé (2024) defined sequentially stable outcomes as follows. A *tremble* is a map $\xi : A \to (0, 1]$ such that $\sum_{a \in A^I} \xi(a) \leq 1$ for all $I \in \mathcal{I}$ and $\xi(a) \leq \pi(a)$ for all a such that $\iota(I^a) = 0$. An outcome ω is a *sequentially stable outcome* if for any tremble sequence $(\xi_n) \to 0$ there exists a sequence of payoff perturbations $(u_n) \to u$ and outcomes $(\omega_n) \to \omega$ such that each ω_n is a Nash outcome of the game

[7] Note that we assume, without loss of generality, that each action is available at only one information set; one can always rename actions to ensure this.

[8] Perfect recall means that for all $I, I' \in \mathcal{I}$ with $\iota(I) = \iota(I')$ and all $h, \hat{h} \in I$, if $(h', a) \preceq h$ for some $h' \in I'$ and $a \in A$, then $(\hat{h}', a) \preceq \hat{h}$ for some $\hat{h}' \in I'$. Here $(h', a) \preceq h$ indicates that (h', a) precedes or equals h.

$G(\xi_n, u_n)$.[9] Dilmé (2024) also shows that σ is part of a sequential equilibrium if and only if there is some tremble sequence $(\xi_n) \to 0$ there exists a sequence of payoff perturbations $(u_n) \to u$ and strategy profiles $(\sigma_n) \to \sigma$ such that each σ_n is a Nash equilibrium of the game $G(\xi_n, u_n)$.

Both sequential equilibria and sequentially stable outcomes exist in all games and possess numerous desirable properties. For example, sequential equilibria are subgame perfect and have consistent beliefs. Sequentially stable outcomes are outcomes of sequential equilibria, satisfy forward induction, and pass all of the selection criteria in Cho and Kreps (1987) and Banks and Sobel (1987) in signaling games.

2 Justifiable and self-justifiable outcomes

In this section, we revisit the definition of justifiable equilibria and define justifiable outcomes. We then define self-justifiable outcomes and relate them to both justifiable and sequentially stable outcomes.

2.1 Justifiable equilibria and justifiable outcomes

Consider the game in Figure 2, which serves as the opening example in McLennan (1985) (cf. Figure 1 of that paper). McLennan observed that this game has two sequential-equilibrium outcomes, $\omega := \mathrm{T}_1$ and $\hat{\omega} := (\mathrm{B}_1, \mathrm{B}_2)$ (i.e., the outcomes assigning probability one to terminal histories T_1 and $(\mathrm{B}_1, \mathrm{B}_2)$, respectively). He noted, however, that in all sequential equilibria supporting ω, player 2 assigns probability no lower than $3/4$ to M_1 (see Example 2.1 below). McLennan argued that such beliefs are implausible, since player 1's payoff from playing M_1 is strictly lower than her payoff under ω, regardless of player 2's response.

To rule out equilibria with such implausible beliefs, McLennan introduced the concept of justifiable equilibria, defined using the following iterative procedure. We say $a \in A$ is *(first-order) useless* if it is not a weak best response in any sequential equilibrium. Let A^1_Ω denote the set of useless actions. Let SE^1_Ω be the set of sequential equilibria in which, whenever two histories h and h' in an information set are such that h has more useless actions than h', we have $\mu(h) = 0$. We say $a \in A \backslash A^1_\Omega$ is *second-order useless* if it is not a weak best response for any sequential equilibrium in SE^1_Ω. Next, let A^2_Ω denote the set of actions that are second-order useless. Let SE^2_Ω be the set of sequential equilibria in which, whenever two histories h and h' in an information set are such that h has either more useless actions than h' or the same number of useless actions but more actions which are second-order useless, we have $\mu(h) = 0$. We iterate this process until we arrive at a step

[9] σ_n is a Nash equilibrium of the $G(\xi_n, u_n)$ if, for all $a \in A$, (i) $\sigma_n(a) \geq \xi_n(a)$, and (ii) $\sigma_n(a) > \xi_n(a)$ only if a is a best response given σ_n and payoffs u_n. A Nash outcome of $G(\xi_n, u_n)$ is the outcome of a Nash equilibrium of the $G(\xi_n, u_n)$.

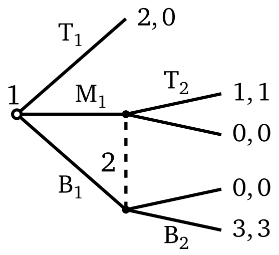

Figure 2

s at which there are no actions that are sth-order useless. We denote by S_Ω the number of steps needed to reach this point. Note that $S_\Omega < |A|$.

A sequential equilibrium is a *justifiable equilibrium* (or has *justifiable beliefs*) if it belongs to $SE_\Omega^{S_\Omega}$. We refer to outcomes of justifiable equilibria as justifiable outcomes.

Definition 2.1. An outcome ω is *justifiable* if $SE_\Omega^{S_\Omega} \cap SE_\omega \neq \emptyset$.

Example 2.1. Let us return to the game in Figure 2. The set of sequential equilibria of this game is[10]

$$\{(\overbrace{\mathrm{B}_1}^{\sigma_1}, \overbrace{\mathrm{B}_2}^{\sigma_2}, \overbrace{\mathrm{B}_1}^{\mu_2})\} \cup \left\{(\mathrm{T}_1, \mathrm{T}_2, x\,\mathrm{M}_1 + (1-x)\,\mathrm{B}_1)) \middle| x \in (3/4, 1]\right\}$$
$$\cup \left\{(\mathrm{T}_1, y\,\mathrm{T}_2 + (1-y)\,\mathrm{B}_2, \tfrac{3}{4}\,\mathrm{M}_1 + \tfrac{1}{4}\,\mathrm{B}_1) \middle| y \in [1/3, 1]\right\}\,.$$

As noted earlier, the corresponding sequential-equilibrium outcomes are $\omega = \mathrm{T}_1$ and $\hat{\omega} = (\mathrm{B}_1, \mathrm{B}_2)$.

As McLennan (1985) explains, action M_1 is the only useless action; hence $A_\Omega^1 = \{\mathrm{M}_1\}$. Actions T_1 and T_2 are second-order useless because they are never weak best responses under the (unique) sequential equilibrium in which player 2 assigns probability zero to M_1; moreover, they are the only second-order useless actions. Finally, it is clear that there are no third-order useless actions: both B_1 and B_2 are weak best responses under the sequential equilibrium in which player 2 assigns probability zero to M_1. Therefore, the unique justifiable equilibrium is $(\mathrm{B}_1, \mathrm{B}_2, \mathrm{B}_1)$, and the unique justifiable outcome is $(\mathrm{B}_1, \mathrm{B}_2)$.

2.2 Self-justifiable equilibria and self-justifiable outcomes

Cho and Kreps (1987) took a different approach to selecting equilibria with plausible beliefs. Their selection criteria, applicable to signaling games, were formulated as "tests" to evaluate the internal consistency of a given outcome. To verify whether a given outcome passes a criterion, one iteratively excludes type–message pairs that are implausible under that outcome, and finally

[10] As usual, the name of an action here denotes the distribution assigning probability one to it. For example, $(\sigma_1, \sigma_2, \mu_2) = (\mathrm{T}_1, \frac{1}{2}\,\mathrm{T}_2 + \frac{1}{2}\,\mathrm{B}_2, \mathrm{M}_1)$ is the assessment where player 1 plays T_1 for sure, player 2 plays T_2 and B_2 each with probability 1/2, and player 2 assigns probability one to history M_1 in her information set.

verifies the existence of sequential equilibria assigning probability zero to excluded types. The criteria of Cho and Kreps have been used extensively in both theoretical and applied work.

In this section, we combine the approaches of Cho and Kreps (1987) and McLennan (1985) to test the internal consistency of equilibria of general games in extensive form. Our procedure is to iteratively exclude actions that are implausible under a given outcome, then identify equilibria such that the associated beliefs are justifiable given the order of action exclusion. As we shall see, requiring justifiability conditional on an outcome makes for greater selection power in many games.

Fix an outcome $\omega \in \Omega$. We define a procedure analogous to the one described in Section 2.1, except that we consider only sequential equilibria with outcome ω (rather than arbitrary sequential equilibria). So, for example, the first step of the procedure is as follows. We say that $a \in A$ is *(first-order) useless under* ω if it is not a weak best response in any sequential equilibrium with outcome ω. By an abuse of notation, we let A^1_ω denote the set of useless actions under ω.[11] Similarly, we let SE^1_ω be the set of sequential equilibria with outcome ω in which, whenever two histories h and h' in an information set are such that h has more useless actions under ω than h', we have $\mu(h)=0$. The other steps of the process are analogous. As before, we proceed iteratively until we reach a step S_ω with no actions that are $(S_\omega+1)$th-order useless under ω. Note that $S_\omega \leq |A|$.

We say that a sequential equilibrium (σ, μ) is *self-justifiable* if it belongs to $SE^{S_{\omega^\sigma}}_{\omega^\sigma}$ (recall that ω^σ is the outcome generated by σ). In words, a sequential equilibrium is self-justifiable if, given its implied behavior (i.e., its outcome), it survives an iterative process in which, at each step, all actions that are currently implausible for the given outcome are excluded. Self-justifiable outcomes are the outcomes of self-justifiable equilibria.

Definition 2.2. An outcome ω is *self-justifiable* if $SE^{S_\omega}_\omega \neq \emptyset$.

Example 2.2. Recall that the two sequential-equilibrium outcomes of the game in Figure 2 are $\omega = \mathrm{T}_1$ and $\hat{\omega} = (\mathrm{B}_1, \mathrm{B}_2)$ and that only $\hat{\omega}$ is justifiable. We now show that $\hat{\omega}$ is also the only self-justifiable outcome. First, observe that all actions except B_1 and B_2 (which are on path under $\hat{\omega}$) are first-order useless under $\hat{\omega}$. Since $\hat{\omega}$ is the outcome of a sequential equilibrium in which history M_1 is assigned probability zero, $\hat{\omega}$ is self-justifiable.

Now consider ω. It is easy to see that M_1 is the only useless action under ω. Since there is no sequential equilibrium with outcome ω in which player 2 assigns probability zero to M_1, ω is not self-justifiable.

[11] In this and similar expressions, we replace Ω by ω to indicate that we are considering equilibria with outcome ω, not all equilibria.

2.3 Comparison between justifiable and self-justifiable outcomes

We now compare the concepts of justifiable and self-justifiable outcomes. As explained above, the main difference between them is that justifiable outcomes are obtained by iteratively excluding actions that are not plausible under any sequential equilibrium, while self-justifiable outcomes are obtained by fixing an outcome ω and then iteratively excluding actions that are not plausible given ω.

In the examples in McLennan (1985), all self-justifiable outcomes are justifiable, while the converse is not true (see Example 2.3). The reason why self-justifiability tends to be a stronger condition than justifiability is that, for a fixed outcome ω, we have $A^1_\omega \supset A^1_\Omega$; that is, the first step of the procedure to obtain self-justifiable outcomes reduces the set of plausible actions no less than the first step of the procedure to obtain justifiable outcomes. For example, in many signaling games, there are no useless actions (because every action is optimal under some sequential equilibrium); hence all sequential-equilibrium outcomes are justifiable. On the other hand, as we explain in Section 4.1, the actions excluded by the selection criteria of Cho and Kreps (1987) for a given outcome are all useless under that outcome; hence self-justifiable outcomes pass all of these criteria.

Still, one can construct examples of games having self-justifiable outcomes that are not justifiable. This is because, in the second step of the respective procedures, there are no restrictions on the relative probability of useless actions. In Appendix A, we present a game with a self-justifiable outcome ω that is not justifiable (see Example A.3). The game features two actions that are useless under an outcome ω. In the first step of the procedure for verifying the self-justifiability of ω, both of these actions are excluded; thus, in the second step, no additional restrictions are introduced, which is shown to imply that ω is self-justifiable. However, only one of the two actions is useless (the other is weakly optimal in a sequential equilibrium with an outcome different from ω). This leads to an additional restriction in the second step of the procedure for obtaining justifiable outcomes, which leads to ω failing to be justifiable. Providing this example, as well as the others below, to show that all implications in Figure 1 are strict, is a contribution of this paper.

In Section 3, we will define generalizations of justifiability and self-justifiability that can be ordered in terms of selection power.

Example 2.3. Consider the game in Figure 3, which corresponds to the game in Figure 6 of McLennan (1985). As McLennan explains, this game has two sequential-equilibrium outcomes,

$$\omega := \tfrac{1}{2}(\mathrm{T}_0, \mathrm{B}_1, \mathrm{T}_3) + \tfrac{1}{2}(\mathrm{B}_0, \mathrm{T}_2, \mathrm{T}_3) \quad \text{and} \quad \hat{\omega} := \tfrac{1}{2}(\mathrm{T}_0, \mathrm{T}_1) + \tfrac{1}{2}(\mathrm{B}_0, \mathrm{B}_1)\ . \tag{2.1}$$

Both outcomes are justifiable: every action is a weak best response under some sequential equilibrium, so none of the actions are excluded by the procedure for obtaining justifiable equilibria.

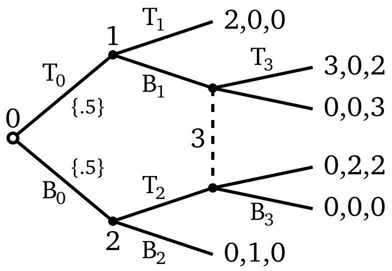

Figure 3

McLennan uses this example to illustrate that justifiability is a weaker condition than strong stability (Kohlberg and Mertens, 1986), as the outcome $\hat{\omega}$ is not strongly stable. We will now prove that ω is self-justifiable, and that $\hat{\omega}$ is not.

We first show that $\hat{\omega}$ is *not* self-justifiable. To see this, note that the set of sequential equilibria with outcome $\hat{\omega}$ (i.e., $SE_{\hat{\omega}}$) is

$$\Big\{\Big(\overbrace{\tfrac{1}{2}\,\mathrm{T}_0+\tfrac{1}{2}\,\mathrm{B}_0}^{=:\hat{\sigma}_0^x},\overbrace{\mathrm{T}_1}^{=:\hat{\sigma}_1^x},\overbrace{\mathrm{B}_2}^{=:\hat{\sigma}_2^x},\overbrace{x\,\mathrm{T}_3+(1-x)\,\mathrm{B}_3}^{=:\hat{\sigma}_3^x},\overbrace{\tfrac{2}{3}\,(\mathrm{T}_0,\mathrm{B}_1)+\tfrac{1}{3}\,(\mathrm{B}_0,\mathrm{T}_1)}^{=:\hat{\mu}_3^x}\Big)\,\Big|\,x\in[0,\tfrac{1}{2}]\Big\}\ .$$

We use $(\hat{\sigma}^x,\hat{\mu}^x)$ to denote the sequential equilibrium with outcome $\hat{\omega}$ in which player 3 plays T_3 with probability x. When $x>2/3$, player 1 strictly prefers B_1 to T_1. When $x>1/2$, player 2 strictly prefers T_2 to B_2. Hence $x\leq 1/2$ for all equilibria $(\hat{\sigma}^x,\hat{\mu}^x)$ in $SE_{\hat{\omega}}$.

Note that under $(\hat{\sigma}^x,\hat{\mu}^x)$, player 1's payoff from deviating to B_1 is $3x$, which is strictly smaller than her equilibrium payoff of 2 for all $x\in[0,1/2]$. Thus B_1 is useless under $\hat{\omega}$ (i.e., it is never a weak best response under any sequential equilibrium with outcome $\hat{\omega}$). Because T_2 is a weak best response when $x=1/2$, it is not useless under $\hat{\omega}$, and so we have that $A^1_{\hat{\omega}}=\{\mathrm{B}_1\}$. In any equilibrium in $SE^1_{\hat{\omega}}$, player 3 assigns probability zero to $(\mathrm{T}_0,\mathrm{B}_1)$, so she optimally chooses T_3, but this makes it optimal for players 1 and 2 to deviate to B_1 and T_2, respectively. Hence there is no sequential equilibrium with outcome $\hat{\omega}$ in which player 3 assigns probability zero to $(\mathrm{T}_0,\mathrm{B}_1)$ in her information set. This implies that $A^2_{\hat{\omega}}=A\backslash\{\mathrm{B}_1\}$ and $SE^2_{\hat{\omega}}=\emptyset$. Therefore $\hat{\omega}$ is *not* self-justifiable.

We now show that ω is self-justifiable. Observe that there is only one sequential equilibrium with outcome ω, which is given by

$$\Big(\overbrace{\tfrac{1}{2}\,\mathrm{T}_0+\tfrac{1}{2}\,\mathrm{B}_0}^{=:\sigma_0},\overbrace{\mathrm{B}_1}^{=:\sigma_1},\overbrace{\mathrm{T}_2}^{=:\sigma_2},\overbrace{\mathrm{T}_3}^{=:\sigma_3},\overbrace{\tfrac{1}{2}\,(\mathrm{T}_0,\mathrm{B}_1)+\tfrac{1}{2}\,(\mathrm{B}_0,\mathrm{T}_1)}^{=:\mu_3}\Big)\ .$$

Now, the actions T_1, B_2, and B_3 are useless under ω, and there are no second-order useless actions under ω. Hence the unique sequential equilibrium with outcome ω is self-justifiable, and so ω is self-justifiable.

We conclude that the game in Figure 3 has two justifiable outcomes, but only one of them is self-justifiable. Hence, in this game, self-justifiability has greater selection power than justifiability.

2.4 Justifiability, self-justifiability, and sequential stability

Like McLennan (1985) and Cho and Kreps (1987), we are interested in the relationship between self-justifiability and stability.[12] In this section, we show that the use of self-justifiability can help in identifying stable behavior. We focus on relating self-justifiable outcomes to sequentially stable outcomes (as introduced in Dilmé, 2024, and defined in Section 1.2). The reason is that sequentially stable outcomes exist in all games (without the requirement of "generic payoffs") and are defined in terms of behavioral trembles, which makes them easier to compare with concepts based on the iteration exclusion of actions instead of pure strategies. Furthermore, as shown in Dilmé (2024), sequential stability is closely related to Kohlberg–Mertens stability.

Our first result is the following.

Proposition 2.1. *Every sequentially stable outcome is both justifiable and self-justifiable.*

Because all finite games have sequentially stable outcomes, Proposition 2.1 implies that all finite games have outcomes that are both justifiable and self-justifiable (recall that McLennan, 1985, already proved the existence of justifiable outcomes in all finite games). Another implication of Proposition 2.1 is that if ω is a sequentially stable outcome, then there is a sequential equilibrium with (self-)justifiable beliefs with outcome ω, a fact that strengthens the foundation for sequential stability. Finally, Proposition 2.1 establishes both justifiability and self-justifiability as powerful tools for obtaining sequentially stable outcomes without using sequences of strategy profiles: if an outcome is shown not to be justifiable (or not to be self-justifiable), then it is not sequentially stable, while if it is the unique justifiable (or self-justifiable) outcome, it is also the unique sequentially stable outcome.

The proof of Proposition 2.1 proceeds as follows. Take some sequences (ξ_n), (σ_n), and (ε_n) satisfying $\xi_n \to 0$, $\varepsilon_n \to 0$, $\omega^{\sigma_n} \to \omega$, and each σ_n is a sequential ε_n-equilibrium given ξ_n. We first observe that if $\hat{a}$ is the only useless action under ω, then it must be that $\sigma_n(\hat{a})=\xi_n(\hat{a})$ for n large enough. Intuitively, in any sequential equilibrium (σ,μ) supported by a subsequence (σ_{k_n}), $\hat{a}$ becomes increasingly suboptimal as n increases, which implies that $\sigma_{k_n}(\hat{a})=\xi_{k_n}(\hat{a})$ for n large enough. Consequently, if $\xi_n(\hat{a})$ tends to zero much faster than for any other action, the probability of any history containing $\hat{a}$ will tend to zero faster than any history not containing it. As a result, $(\sigma,\mu)\in SE^1_\omega$, that is, ω is robust to (ξ_n) only if $SE^1_\omega \neq \emptyset$.

[12] McLennan (1985) discusses the usefulness of justifiability in identifying strongly stable sets of Nash equilibria (as defined by Kohlberg and Mertens, 1986), when these exist. He says, "Unfortunately, it is usually not easy to verify that a component of Nash equilibria is strongly stable, but the notion of justifiability may prove helpful in this respect. Specifically, if an isolated equilibrium path is supported by a strongly stable set of Nash equilibria, it must also be supported by a sequential equilibrium with justifiable beliefs" (p. 895). Similarly, Cho and Kreps (1987) say, "Besides posing 'intuitive tests' of equilibrium outcomes in signaling games, we follow the program above in order to relate our tests to Kohlberg–Mertens stability. We seek, in general, to show that any equilibrium outcome that fails any of the tests we construct fails as well to be a stable equilibrium outcome" (p. 198).

This argument can be applied iteratively. Now, assume that ω is *not* self-justifiable. Consider the tremble sequence defined by

$$\xi_n(a) := \begin{cases} e^{-n^{S_\omega - s + 2}} & \text{if } a \in A_\omega^s \text{ for some } s \in \{1, ..., S_\omega\}, \\ e^{-n} & \text{otherwise.} \end{cases} \tag{2.2}$$

This tremble sequence has the property that the relative probability of two histories is lexicographically determined by the uselessness orders of their actions under ω, as required to obtain self-justifiable beliefs. Hence, as n increases, the probability that ξ_n assigns to actions of lower orders of uselessness under ω decreases to 0 much faster than the probability it assigns to actions of higher orders of uselessness under ω. Applying the previous argument iteratively, the proof argues that because $SE_\omega^{S_\omega} = \emptyset$ (since ω is not self-justifiable), there are no sequences $(u_n) \rightarrow u$ and $(\omega_n) \rightarrow \omega$ where each ω_n is a Nash equilibrium outcome of $G(\xi_n, u_n)$. This implies that ω is not sequentially stable.

Example 2.4. McLennan (1985) shows that in the game in Figure 3, the outcome ω defined in Example 2.3 is *not* the outcome of a stable set of equilibria. We obtain an analogous result: because ω is not self-justifiable (as shown in Example 2.3), it is not sequentially stable. Since self-justifiable outcomes always exist (by Proposition 2.1), it follows that the outcome $\hat{\omega}$ is self-justifiable (as explicitly shown in Example 2.3), and therefore it is also the unique sequentially stable outcome of the game.

3 Fully justifiable and fully self-justifiable outcomes

In this section, we introduce full justifiability and full self-justifiability. These concepts fulfill three key objectives. First, they ensure the robustness of equilibrium behavior under any process of iterated exclusion, not just under the sequences specified in the procedures defining justifiability and self-justifiability. Second, they can be clearly ranked in terms of selection power. Third, as we will demonstrate, they provide stronger and more flexible tools for verifying or ruling out the sequential stability of an outcome.

3.1 Generalizing justifiability

In each step of the procedures for obtaining justifiable and self-justifiable outcomes (described in Sections 2.1 and 2.2, respectively), the set of remaining plausible actions is reduced by excluding *all* actions that are not weak best responses under any sequential equilibrium that is plausible given the previous exclusions. While such a requirement is convenient in that it uniquely determines the procedure, it may seem ad hoc: the internal logic of the iterated exclusion of implausible actions seems independent of the fact that *all* currently implausible actions have to be excluded in each step. Also, in practice, to exclude all actions that are never weak best responses under a certain collection of sequential equilibria, one often needs to compute the whole set of such equilibria, and this may be difficult in applications.

The equilibrium selection criteria that we introduce in this section, full justifiability and full self-justifiability, are defined similarly to justifiability and self-justifiability, but without the requirement that all useless actions be excluded in each step. As we shall see, these criteria are stronger and more flexible to use than the earlier ones. However, to define them, we need to deal with the different possible orders in which actions may be excluded.

We begin by generalizing the definition of justifiability.

Definition 3.1. Fix some non-intersecting sequence $(A_s)_{s=1}^S$.[13] A sequential equilibrium (σ,μ) is $(A_s)_{s=1}^S$*-justifiable* if, for each pair of histories h and h' belonging to the same information set, we have $\mu(h)=0$ whenever there is some $\hat{s}$ such that (i) h and h' have the same number of actions in A_s for all $s<\hat{s}$, and (ii) h has more actions in $A_{\hat{s}}$ than h' does.

For a given outcome $\omega\in\Omega$ and non-intersecting $(A_s)_{s=1}^S$, we use $SE_\omega((A_s)_{s=1}^S)$ to denote the set of $(A_s)_{s=1}^S$-justifiable equilibria with outcome ω (note that $SE_\omega((A_s)_{s=1}^S)$ may be empty). Consistently with our previous notation, we let $SE_\Omega((A_s)_{s=1}^S)$ indicate the set of all $(A_s)_{s=1}^S$-justifiable equilibria; that is, $SE_\Omega((A_s)_{s=1}^S):=\cup_{\omega\in\Omega}SE_\omega((A_s)_{s=1}^S)$.

Definition 3.1 generalizes the definitions of justifiability and self-justifiability in Sections 2.1 and 2.2, respectively, by allowing the sequence of sets of excluded actions to be exogenously given. Indeed, note that SE_Ω^s is the set of $(A_\Omega^{\hat{s}})_{\hat{s}=1}^s$-justifiable equilibria, and SE_ω^s is the set of $(A_\omega^{\hat{s}})_{\hat{s}=1}^s$-justifiable equilibria with outcome ω (where $(SE_\Omega^s, A_\Omega^s)$ and $(SE_\omega^s, A_\omega^s)$ are as defined in Sections 2.1 and 2.2, respectively). Visual inspection of the procedure defining self-justifiability confirms that actions in A_s in Definition 3.1 play the same role as sth-order useless actions in the earlier definitions: $\mu(h)=0$ whenever h has more actions in A_1 than h', or when h has the same number of actions in A_1 as h' but more actions in A_2, or when h has the same number of actions in A_1 and in A_2 as h' but more actions in A_3, etc.

For a given $\omega\in\Omega\cup\{\Omega\}$, we use $NWBR_\omega((A_s)_{s=1}^S)$ to denote the set of actions that are never a weak best response for any sequential equilibrium in $SE_\omega((A_s)_{s=1}^S)$. Note that if $SE_\omega((A_s)_{s=1}^S)=\emptyset$ (i.e., if there is no $(A_s)_{s=1}^S$-justifiable sequential equilibrium with outcome ω), then $NWBR_\omega((A_s)_{s=1}^S)=A$ (i.e., there is no action which is a weak best response for some $(A_s)_{s=1}^S$-justifiable sequential equilibrium with outcome ω).

[13] A sequence $(A_s)_{s=1}^S$ is *non-intersecting* if $A_s\neq\emptyset$ and $A_s\cap A_{s'}=\emptyset$ for all $s,s'\in\{1,...,S\}$ with $s\neq s'$. We take $(A_s)_{s=1}^0$ to be $\emptyset$.

3.2 Iterated exclusion through NWBR

We now introduce a procedure we call "iterated exclusion through NWBR". This is a generalization of the procedures described in Sections 2.1 and 2.2.

Definition of IENWBR_ω

For a fixed $\omega \in \Omega \cup \{\Omega\}$, a process of *iterated exclusion through NWBR under* ω (denoted by IENWBR_ω) consists of the iterated exclusion of actions that are deemed implausible according to $NWBR_\omega$. More formally, in the first step of the process, a non-empty set $A_1 \subset NWBR_\omega((A_s)_{s=1}^0)$ of implausible actions is excluded, if such A_1 exists (otherwise the process ends), where $(A_s)_{s=1}^0 \equiv \emptyset$. In the second step of the process, a non-empty set $A_2 \subset NWBR_\omega((A_s)_{s=1}^1) \setminus A_1$ of implausible actions is excluded, if such A_2 exists (otherwise the process ends). And so on.

Note that, unlike the processes described in Sections 2.1 and 2.2, a process of IENWBR_ω may not be unique, because the set of actions excluded in each step s need not be maximal: it may be a strict subset of $NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^{s-1}) \setminus \cup_{\hat{s}=1}^{s-1} A_{\hat{s}}$. Because IENWBR_ω is not unique, we now define the concept of implementations of IENWBR_ω.

Definition 3.2. Fix $\omega \in \Omega \cup \{\Omega\}$. A non-intersecting sequence $(A_s)_{s=1}^S$ is an *implementation of* $IENWBR_\omega$ if $A_s \subset NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^{s-1})$ for all s. It is *complete* if $NWBR_\omega((A_s)_{s=1}^S) = \cup_{s=1}^S A_s$.

That is, an implementation of IENWBR_ω is a non-intersecting sequence of sets that is consistent with the iterated exclusion procedure described above. Note that if $(A_s)_{s=1}^S$ is an implementation of IENWBR_ω, then $S \leq |A|$; hence all implementations are finite (and the set of implementations is finite). Note also that, because the number of constraints on the equilibria in $SE_\omega((A_{\hat{s}})_{\hat{s}=1}^s)$ increases with s, we have

$$SE_\omega((A_{\hat{s}})_{\hat{s}=1}^s) \subset SE_\omega((A_{\hat{s}})_{\hat{s}=1}^{s-1}) \tag{3.1}$$

for all $s = 1, \ldots, S$; that is, the set of sequential equilibria which remain plausible shrinks after each step. This implies that

$$NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^s) \supset NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^{s-1}) \; ; \tag{3.2}$$

that is, the set of implausible actions grows with each step. In particular, the set of implausible actions at a given step includes all of the actions excluded at previous steps; that is, $NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^s) \supset \cup_{\hat{s}=1}^s A_{\hat{s}}$ for all s. An implementation is thus complete if no actions remain to be excluded in step S. Note that if $SE_\omega((A_{\hat{s}})_{\hat{s}=1}^s) = \emptyset$ for some s, then $NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^s) = A$.

3.3 Full justifiability and full self-justifiability

We now define the concepts of fully justifiable and fully self-justifiable outcomes. These definitions are analogous to the definitions of justifiable outcomes (Definition 2.1) and self-justifiable outcomes (Definition 2.2) but require the corresponding conditions for all implementations.

Definition 3.3. Fix an outcome $\omega \in \Omega$.

(i) ω is *fully justifiable* if $SE_\Omega((A_s)_{s=1}^S) \cap SE_\omega \neq \emptyset$ for all implementations $(A_s)_{s=1}^S$ of IENWBR_Ω.

(ii) ω is *fully self-justifiable* if $SE_\omega((A_s)_{s=1}^S) \neq \emptyset$ for all implementations $(A_s)_{s=1}^S$ of IENWBR_ω.

Note that, as explained above, full justifiability and full self-justifiability require an additional degree of plausibility relative to justifiability and self-justifiability, respectively, since they require an outcome to be supported by a plausible equilibrium independently of the order of exclusion of implausible actions. Consequently it is often easier to prove, for example, that an outcome is not fully self-justifiable than that it is not self-justifiable, since, for the former, it suffices to find a single implementation $(A_s)_{s=1}^S$ of IENWBR_ω for which $SE_\omega((A_s)_{s=1}^S) = \emptyset$. This fact is particularly useful because the full self-justifiability of an outcome can often be proven by showing that no other sequential-equilibrium outcome is fully self-justifiable.

More concretely, in a large game, it may be difficult to compute the complete set of sequential equilibria, but it is often easy to show that a given action is never a weak best response under any sequential equilibrium. For example, if an action that is strictly dominated by another action is never a weak best response. Similarly, by identifying the set of sequential equilibria in a subgame, one can identify the actions in this subgame that are never a weak best response under any sequential equilibrium of the big game. The procedure defining full self-justifiability allows one to exclude easily-identifiable actions first, which in turn makes it easier to exclude further actions, and so on (see Examples (3.2) below).

Proposition 3.1. *If an outcome is fully (self-)justifiable, then it is (self-)justifiable.*

Proposition 3.1 establishes that, in addition to being more flexible in their application, full justifiability and full self-justifiability are respectively stronger than justifiability and self-justifiability. This result follows from the observation that the sequence of sets used to determine justifiability or self-justifiability (namely, $(A_\omega^s)_{s=1}^S$, for $\omega \in \Omega \cup \{\Omega\}$) is an implementation of IENWBR_ω. Hence, if, for example, ω is fully justifiable, it is also justifiable. Example A.1 in the appendix shows that in some games, full self-justifiability has strictly more selection power than self-justifiability; that is, there are self-justifiable outcomes that are not fully self-justifiable. (For this reason, full self-justifiability will turn out to be a more effective criterion to use in characterizing sequentially stable outcomes.) Similar examples show that full justifiability is strictly stronger than justifiability.

Although, by definition, $\omega \in \Omega$ is fully self-justifiable if $SE_\omega((A_s)_{s=1}^S) \neq \emptyset$ for all implementations $(A_s)_{s=1}^S$ of IENWBR$_\omega$, one may wonder whether it suffices to check this condition for a single complete implementation. In fact, as we explain below, the desired property is independent of the choice of complete implementation of IENWBR$_\omega$ in signaling games. However, this is not the case in general. Example A.2 in the appendix presents a game with an outcome ω and complete implementations $(A_s)_{s=1}^S$ and $(A'_s)_{s=1}^{S'}$ of IENWBR$_\omega$ such that $SE_\omega((A_s)_{s=1}^S) \neq \emptyset$ and $SE_\omega((A'_s)_{s=1}^{S'}) = \emptyset$.

Remark 3.1. As previously noted, the procedures for obtaining fully justifiable and fully self-justifiable outcomes are more flexible than those for obtaining justifiable and self-justifiable outcomes, because they do not require the exclusion of *all* implausible actions at each step—a task for which, in many cases, one would need to identify all surviving sequential equilibria. However, establishing full justifiability or full self-justifiability still requires showing that a supporting sequential equilibrium exists for all implementations, which may be difficult. This difficulty can often be circumvented as follows. Because fully self-justifiable outcomes always exist (see Proposition 3.3), if one can use the procedure to rule out the full self-justifiability of all sequential-equilibrium outcomes but one, the remaining outcome is guaranteed to be fully self-justifiable. In many games, the set of sequential-equilibrium outcomes is small and easy to characterize (unlike the set of sequential equilibria), and ruling out the full self-justifiability of all outcomes but one is also not difficult. Similarly, if there is a unique outcome passing the procedure defining full self-justifiability for a given implementation, then this is guaranteed to be the unique fully self-justifiable (and sequentially stable) outcome.[14]

Full self-justifiability is stronger than full justifiability

In Section 2.1, we observed that while self-justifiability tends to be stronger than justifiability (see Example 2.3), it is not stronger in all games. As explained above, the reason is that although the procedure defining self-justifiability may exclude more actions in the initial step, this procedure does not impose any condition on the relative likelihoods of actions which are excluded in the same step. By contrast, full self-justifiability is stronger than full justifiability.

Proposition 3.2. *Every fully self-justifiable outcome is fully justifiable.*

Proposition 3.2 is proved by showing that if $(A_s)_{s=1}^S$ is an implementation of IENWBR$_\Omega$, then

[14] A further advantage of full justifiability is that the procedure defining it, unlike the procedure defining justifiability, does not require one to know the set of sequential equilibria of the game. For example, in a given step, one can often show that an action is dominated independently of the continuation play, given the current restrictions on the beliefs. In the game in Figure 2 (discussed in Example 2.1), action M_1 is strictly dominated by T_1, so because (B_1, B_2) is the unique sequential-equilibrium outcome in which player 2 assigns probability one to B_1, this outcome is the unique fully justifiable outcome. See also Example 3.2.

it is also an implementation of IENWBR$_\omega$. Intuitively, in each step s, we have

$$SE_\omega((A_{\hat{s}})_{\hat{s}=1}^{s}) \subset SE_\Omega((A_{\hat{s}})_{\hat{s}=1}^{s}) ,$$

as $SE_\omega((A_{\hat{s}})_{\hat{s}=1}^{s})$ is equal to the intersection between $SE_\Omega((A_{\hat{s}})_{\hat{s}=1}^{s})$ and SE_ω. As a result, $NWBR_\omega((A_{\hat{s}})_{\hat{s}=1}^{s}) \supset NWBR_\Omega((A_{\hat{s}})_{\hat{s}=1}^{s})$. Hence, if $SE_\omega((A_s)_{s=1}^{S}) \neq \emptyset$ for all implementations of IENWBR$_\omega$, then $SE_\Omega((A_s)_{s=1}^{S}) \neq \emptyset$ for all implementations of IENWBR$_\Omega$.

An important implication of Proposition 3.2 is that, when studying a game, one can first check which outcomes are fully justifiable (or justifiable). Then, if more than one outcome is fully justifiable, one can check whether each of them is fully self-justifiable.

The fact that full self-justifiability can have strictly more selection power than full justifiability can be seen from Example 2.3 above. In that example, because every action is a weak best response under some sequential equilibrium, no action is excluded under IENWBR$_\Omega$, and hence both sequential-equilibrium outcomes are fully justifiable. Furthermore, because one of the justifiable outcomes is not self-justifiable, Proposition 3.1 implies that it is also not fully self-justifiable.

3.4 Full self-justifiability and sequential stability

In this section, we show that sequentially stable outcomes are fully self-justifiable (and therefore fully justifiable).

Proposition 3.3. *Every sequentially stable outcome is fully self-justifiable.*

Proposition 3.3 generalizes Proposition 2.1. It implies that fully self-justifiable outcomes exist in all finite games. It also establishes full self-justifiability as a useful tool for identifying sequentially stable outcomes without using sequences of strategy profiles: if an outcome ω is not fully self-justifiable, then it is not sequentially stable, while if ω is the unique fully self-justifiable outcome, then it is the unique sequentially stable outcome. For example, each of the games studied in Examples 2.3, A.1, and A.2 has a unique fully self-justifiable outcome, which is therefore its unique sequentially stable outcome. Note that Propositions 3.1–3.3 are jointly summarized in Figure 1. The proof of the result uses the same logic described after Proposition 2.1.

While full self-justifiability is a strong condition, it is weaker than sequential stability. The intuitive reason is that when an outcome ω fails to be fully self-justifiable, the corresponding iterated exclusion procedure can be seen as identifying a tremble sequence that destabilizes ω. (For details, see the argument concerning self-justifiability following the statement of Proposition 2.1.) Such tremble sequences assign extreme values to the relative asymptotic likelihoods of excluded actions: the tremble probability of an action excluded earlier in the process vanishes much faster

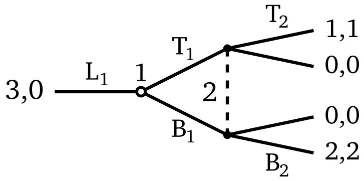

Figure 4

than the tremble probability of an action excluded later. However, some games feature outcomes that are stable against tremble sequences with extreme relative asymptotic likelihoods (and so are fully self-justifiable), but that are fragile to some tremble sequences with comparable relative tremble probabilities (and so are not sequentially stable). Example A.4 below presents one such game.

3.5 Outcomes versus strategy profiles

Our analysis of full justifiability and full self-justifiability has been focused on outcomes rather than on equilibria (i.e., equilibrium strategy profiles). The reason is that, unlike justifiable and self-justifiable equilibria, fully justifiable and fully self-justifiable equilibria may not exist in all games.[15] That is, as shown in Example 3.1 below, there are games with no sequential equilibrium that belongs to $SE_\Omega((A_s)_{s=1}^S)$ (resp. $SE_\omega((A_s)_{s=1}^S)$) for all implementations of IENWBR$_\Omega$ (resp. IENWBR$_\omega$).

The lack of existence of sequential equilibria surviving all implementations is expected: a similar lack of existence occurs in the study of stable behavior, where robustness against all trembles is required.[16] However, if an outcome is fully self-justifiable, then not only is it supported by a self-justifiable equilibrium, but also, independently of the specific iterative process by which actions are excluded, it is also supported by a sequential equilibrium with beliefs that are justifiable according to this process. We believe that this additional form of robustness strengthens the concept of fully self-justifiable outcomes.

Example 3.1. This example shows that $SE_\omega((A_s)_{s=1}^S)$ may depend on the implementation $(A_s)_{s=1}^S$.

[15] Here we are considering the natural definitions: (σ,μ) is a fully justifiable (resp. fully self-justifiable) equilibrium if $(\sigma,\mu)\in SE_\Omega((A_s)_{s=1}^S)$ (resp. $(\sigma,\mu)\in SE_{\omega^\sigma}((A_s)_{s=1}^S)$) for all implementations of IENWBR$_\Omega$ (resp. IENWBR$_{\omega^\sigma}$). Note that Proposition 2.1 trivially implies the existence of self-justifiable equilibria, while McLennan (1985) proved the existence of justifiable equilibria.

[16] Equilibria robust to trembles, called *strictly perfect equilibria* (Okada, 1981), do not exist in many games of interest. This fact has motivated the development of set-valued equilibrium concepts (e.g., stable sets of equilibria, introduced by Kohlberg and Mertens, 1986) and outcome-valued equilibrium concepts (e.g., the sequentially stable outcomes of Dilmé, 2024).

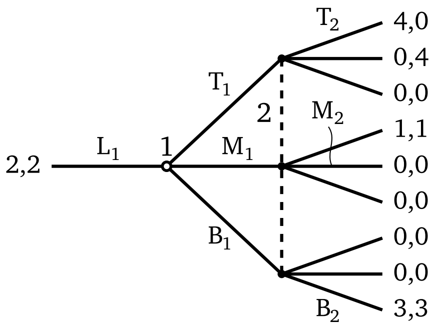

Figure 5

Consider the game in Figure 4. The set of sequential equilibria $(\sigma_1, \sigma_2, \mu_2)$ of this game is

$$
\begin{aligned}
&\{(\mathrm{L}_1, x\,\mathrm{T}_2 + (1-x)\,\mathrm{B}_2, 1/3\,\mathrm{T}_1 + 2/3\,\mathrm{B}_1) | x \in [0,1]\} \\
&\cup \{(\mathrm{L}_1, \mathrm{T}_2, y\,\mathrm{T}_1 + (1-y)\,\mathrm{B}_1) | y \in [0, 1/3]\} \\
&\cup \{(\mathrm{L}_1, \mathrm{B}_2, y\,\mathrm{T}_1 + (1-y)\,\mathrm{B}_1) | y \in [2/3, 1]\} \ .
\end{aligned}
$$

Since all of these equilibria have the same outcome (namely, L_1), this outcome is both fully justifiable and fully self-justifiable. Note that both T_1 and B_1 are useless. In an implementation of IENWBR_Ω where B_1 is excluded first, the only remaining equilibrium is $(\mathrm{L}_1, \mathrm{T}_2, \mathrm{T}_1)$. In an implementation where T_1 is excluded first, the only remaining equilibrium is $(\mathrm{L}_1, \mathrm{B}_2, \mathrm{B}_1)$. Hence this game has no fully justifiable or fully self-justifiable equilibrium. (However, all of its sequential equilibria are both justifiable and self-justifiable.)

3.6 Further examples

Example 3.2. This example illustrates that fully justifiable outcomes are often easier to obtain than justifiable outcomes. Consider the game in Figure 5 (which corresponds to the game in Figure 3 in McLennan, 1985). We can find the justifiable outcomes of this game as follows. First, it is clear that M_1 is a useless action: by playing L_1, player 1 obtains a strictly higher payoff regardless of what player 2 does. However, even in this simple example, it is not obvious whether any other action is useless.[17] McLennan (1985) shows that none of the other actions is useless and that the second-order useless actions are T_1 and T_2. From this he concludes that the only justifiable equilibrium has outcome $(\mathrm{B}_1, \mathrm{B}_2)$.

It is considerably easier to find the fully justifiable outcomes: one can first exclude M_1 (without needing to compute any sequential equilibria), then T_2, and finally L_1 and M_2. Since $(\mathrm{B}_1, \mathrm{B}_2)$ is

[17] It is often possible to show that an action is not useless “by example”, that is, by exhibiting a sequential equilibrium in which the action is a weak best response. To show that an action is useless, it is typically necessary either to find another action that strictly dominates it (for example, in the game in Figure 5, L_1 dominates M_1), to use arguments like those mentioned at the end of Section 4.1, or to characterize the whole set of sequential equilibria.

the unique sequential-equilibrium outcome consistent with these exclusions, it is fully justifiable (and hence justifiable as well).

Example 3.3. This example further illustrates the advantages of full justifiability over justifiability, especially in big games. We now consider a T-repetition of the game in Figure 5, for some $T \in \mathbb{N}$, where payoffs are aggregated using discount factors $\delta_1, \delta_2 \in (0,1]$. As the stage game has multiple sequential equilibria, for high values of δ_i it follows that the repeated game has a very large number of sequential equilibria. This makes it difficult to determine whether a given action in a given period is useless (or to find its order of uselessness), because doing so would require one to compute the full set of sequential equilibria of the repeated game. Consequently, it is difficult to identify justifiable outcomes, as the procedure defining these requires the elimination of *all* implausible actions at each step. We now argue that fully justifiable outcomes are much easier to identify, because one can proceed backward in an intuitive way.

We use H^t to indicate the set of histories h^t containing the actions played in periods $1, ..., t-1$ (with $h^0 \equiv \emptyset$). For a given history $h^t \in H^t$, we let $a(h^t)$ denote the action corresponding to $a \in \{\mathrm{L}_1, \mathrm{T}_1, \mathrm{M}_1, \mathrm{B}_1, \mathrm{T}_2, \mathrm{M}_2, \mathrm{B}_2\}$ at time t given the previous history.[18] In the last period, for each given h^T, we can exclude action $\mathrm{M}_1(h^T)$ and then $\mathrm{M}_2(h^T)$, which leaves $(\mathrm{B}_1(h^T), \mathrm{B}_2(h^T))$ as the only justifiable outcome. Defining

$$A_1 := \cup_{h^T \in H^T} \{\mathrm{M}_1(h^T)\} \text{ and } A_2 := \cup_{h^T \in H^T} \{\mathrm{M}_2(h^T)\},$$

we have that all sequential equilibria (σ, μ) that are $(A_s)_{s=1}^2$-justifiable satisfy $\sigma(\mathrm{B}_1(h^T)) = \sigma(\mathrm{B}_2(h^T)) = 1$ for all h^T. Now, in the period before the last, we can use the same argument to rule out $\mathrm{M}_1(h^{T-1})$ and then $\mathrm{M}_2(h^{T-1})$ for all $h^{T-1} \in H^{T-1}$. Again, defining

$$A_3 := \cup_{h^{T-1} \in H^{T-1}} \{\mathrm{M}_1(h^{T-1})\} \text{ and } A_4 := \cup_{h^{T-1} \in H^{T-1}} \{\mathrm{M}_2(h^{T-1})\},$$

we have that all sequential equilibria (σ, μ) that are $(A_s)_{s=1}^4$-justifiable satisfy $\sigma(\mathrm{B}_1(h^{T-1})) = \sigma(\mathrm{B}_2(h^{T-1})) = 1$ for all h^{T-1}. Iterating this argument, we conclude that the only fully justifiable outcome (and hence the unique sequentially stable outcome) of the repeated game is the repetition of the unique justifiable outcome of the stage game.

4 Full self-justifiability in games with signaling

Since the work of Spence (1973), signaling games have played a central role in the analysis of games with private information. Because signaling games often exhibit high equilibrium multiplicity, a number of equilibrium selection criteria have been proposed for them, often accompanied

[18] Note that our definition of a game in Section 1 requires each action to be played in a single information set. Hence it is necessary to distinguish between analogous actions played in different information sets depending on the histories that give rise to them.

by intuitive arguments to facilitate analysis. However, the usefulness of these selection criteria is limited by the narrow class of games to which they apply: while many games of interest include some form of signaling, they rarely adhere fully to the definition of a signaling game (e.g., games featuring prior investment, multiple receivers).

In applications involving such non-signaling games, researchers have sometimes used intuitive reasoning to select outcomes "in the spirit of" a given selection criterion. These arguments tend to be ad hoc and are not applied consistently across different models. For example, the version of the Spence model used in Cho and Kreps (1987) to illustrate the selection power of D1 is not a signaling game under their definition, as two receivers act after the receiver's signal choice.[19] In this section, we give a natural, systematic way to extend the intuitive reasoning behind Cho and Kreps's selection criteria to general games. In particular, we argue that this intuitive reasoning can be naturally extended and systematically applied to obtain fully self-justifiable outcomes in general games in extensive form.

4.1 Full self-justifiability in signaling games

We begin by investigating the implications of full self-justifiability in signaling games.

The players in a signaling game are a sender and a receiver. The game proceeds as follows: first nature decides the sender's type, then the sender sends a message, and finally the receiver takes an action. Justifiability is of limited use as a selection criterion in signaling games, because it only allows for the exclusion of type–message pairs where the message is not a weak best response for the type in any sequential equilibrium. In most of the examples discussed in Cho and Kreps (1987), all sequential-equilibrium outcomes are justifiable. For instance, in their beer–quiche game, every action is a weak best response under some sequential equilibrium, and so both the beer outcome and the quiche outcome are justifiable.

Cho and Kreps proposed four selection criteria for signaling games: the Intuitive Criterion (IC), D1, D2, and Never a Weak Best Response (NWBR). Like self-justifiability, these criteria are based on testing whether a given outcome ω is internally consistent by excluding all actions that are implausible under all sequential equilibria with outcome ω. The following proposition establishes that self-justifiable outcomes pass all these criteria, as well as divinity and universally divinity (Banks and Sobel, 1987).

Proposition 4.1. *In signaling games, all self-justifiable outcomes are fully self-justifiable, pass IC, D1, D2, and NWBR, and are divine and universally divine.*

[19] Note also that the game they study is infinite, but it can be easily discretized. See, for example, variations of NWBR in Noldeke and van Damme (1990), Swinkels (1999), and Ekmekci and Kos (2023). Belief monotonicity requirements in Daley and Green (2012) and Dilmé (2019), or divinity in Kremer and Skrzypacz (2007) play a similar role.

The proof of Proposition 4.1 is as follows. Fix some $\omega \in \Omega$ and an off-path message m. Assume that ω is self-justifiable. Recall that NWBR is the strongest among the criteria of Cho and Kreps (1987); that is, a type–message pair is excluded by one of the other criteria only if it is excluded under NWBR. Recall also that ω passes (the one application of) NWBR only if there is a sequential equilibrium in which, upon receiving m, the receiver assigns probability zero to the sender types for which sending m is never a weak best response for any sequential equilibrium with outcome ω—that is, the types for which m is a first-order useless action. The exclusion of these types is equivalent to the first step in the iterative procedure defining self-justifiability (as described in Section 2.2).

Note that self-justifiability requires further exclusions of actions in later steps. It is easy to see that, in a signaling game, an outcome is self-justifiable if and only if it passes the iterated exclusion of implausible types through NWBR (i.e., IENWBR), as defined in Dilmé (2025). Since the assessment of IENWBR is independent of the order of exclusion, self-justifiable outcomes are fully self-justifiable in signaling games. In particular, because outcomes passing the iterated application of NWBR also pass the criteria of Banks and Sobel (1987) for divinity and universal divinity (as shown in Dilmé, 2025), self-justifiable outcomes are both divine and universally divine.

4.2 Extending the logic of Cho and Kreps (1987) to games in extensive form

We now illustrate how the logic behind the selection criteria of Cho and Kreps (1987) can be used in obtaining self-justifiable outcomes in games in extensive form. We first present some definitions and a result, and then relate them to the selection criteria of Cho and Kreps.

Fix an outcome $\omega \in \Omega$. Let AS_ω be the set of assessments with outcome ω, and let $A^0(\omega)$ be the set of actions available on the path of ω but played with probability zero under ω. For each $I \in \mathcal{I}$ reached with positive probability under ω, let $u(I|\omega)$ be the payoff of player $\iota(I)$ under ω conditional on I being reached. Similarly, for each action $a \in A$ and assessment (σ, μ), let $u(a|\sigma,\mu)$ be player $\iota(I^a)$'s payoff from playing a.[20]

Proposition 4.2. *Fix some $\omega \in \Omega$ and $a \in A^0(\omega)$. Let $SE' \subset SE_\omega$ and $AS' \subset AS_\omega$ be sets of sequential equilibria and assessments, respectively, satisfying $SE' \subset AS'$. Assume that for each $(\sigma,\mu) \in AS'$ with $u(a|\sigma,\mu) = u(I^a|\omega)$, there is some $a' \in A^0(\omega)$ with $u(a'|\sigma,\mu) > u(I^{a'}|\omega)$. Then a is not a weak best response under any $(\sigma,\mu) \in SE'$.*

[20] Note that $u(I|\omega)$ is uniquely defined as $\sum_{z \in Z^I} \omega(z)\, u_{\iota(I)}(z) / \sum_{z \in Z^I} \omega(z)$, where Z^I is the set of terminal histories succeeding some history in I. Similarly, $u(a|\sigma,\mu)$ is uniquely defined as

$$u(a|\sigma,\mu) = \sum_{h \in I^a} \sum_{z \in Z^{(h,a)}} \mu(h)\, \sigma(z|(h,a))\, u_{\iota(I^a)}(z) \ ,$$

where $Z^{(h,a)}$ is the set of terminal histories that succeed or are equal to (h,a), and for each $z \in Z^{(h,a)}$ we set $\sigma(z|(h,a)) := \prod_{j=J+1}^{|z|} \sigma(z_j)$, with J denoting the index such that $(z_j)_{j=1}^{J} = (h,a)$.

Note that the NWBR condition of Cho and Kreps is based on applying Proposition 4.2 with $SE'=SE_\omega$ and letting AS' be the set of assessments with outcome ω in which the receiver best-responds to some belief after message m.[21]

The Intuitive Criterion, D1, and D2 are defined using weaker conditions than NWBR. For each of these conditions, one can state a result analogous to Proposition 4.2 to extend the exclusion logic to all games. For example, the extended Intuitive Criterion (applied to a set AS') can be defined as requiring an action a to be excluded if $a\in A^0(\omega)$ and, for any $(\sigma,\mu)\in AS'$, we have $u(a|\sigma,\mu)< u(I^a|\omega)$. Similarly, a can be excluded "in the spirit of D1" applied to AS' if $a\in A^0(\omega)$ and there is some $a'\in A^0(\omega)$ such that, for each $(\sigma,\mu)\in AS'$ with $u(a|\sigma,\mu)\geq u(I^a|\omega)$, we have $u(a'|\sigma,\mu)> u(I^{a'}|\omega)$.[22] For example, our approach can be used to formalize Cho and Kreps (1987)'s analysis of the Spence model: one can exclude type-message pairs (θ,m) satisfying that there is some type θ' such that, when the competing price offers are such that θ prefers deviating to m, θ' strictly benefits from that. Since only the Riley outcome has this property, it is the only fully self-justifiable outcome.

Proposition 4.2 is useful in obtaining self-justifiable outcomes because, in each step s of the procedure, one can exclude actions by choosing $SE'=SE^s_\omega$ and some $AS'\supset SE^s_\omega$ and verifying the property in the statement of Proposition 4.2 (or any of the weaker versions described above). For example, one can let AS' be the set of assessments respecting the relative order of uselessness of histories (as well as satisfying the necessary conditions for them to be sequential equilibria).[23]

Example 4.1. Let us again consider the game in Figure 3. Note that while this game is *not* a signaling game, it is similar to one in structure. In Example 2.3 we showed that the outcome $\hat{\omega}:=\frac{1}{2}(\mathrm{T}_0,\mathrm{T}_1)+\frac{1}{2}(\mathrm{B}_0,\mathrm{B}_1)$ is not self-justifiable. We now use the extension of the D1 criterion to games in extensive form to prove the same thing. Observe that under $\hat{\omega}$, regardless of player 3's strategy, if player 1 weakly prefers to choose B_1, then player 2 strictly prefers to choose T_2. Hence, from the extended D1 criterion, there is no sequential equilibrium with outcome $\hat{\omega}$ in which B_1 is a weak best response; that is, B_1 is useless under $\hat{\omega}$. Now note that (i) T_2 is a weak best response under a sequential equilibrium with outcome $\hat{\omega}$, (ii) the unique best response to

[21] Recall that a type–message pair (θ,m) is excluded according to NWBR (for some outcome ω and off-path message m) if, whenever the receiver's response to m makes θ indifferent between the on-path message and m, there is a type that strictly prefers to send m. Because in signaling games all types occur with positive probability, all pairs (θ,m) that do not occur with positive probability under ω belong to $A^0(\omega)$.

[22] D2 can be extended as well by excluding an action a if $a\in A^0(\omega)$ and, for each $(\sigma,\mu)\in AS'$ with $u(a|\sigma,\mu)\geq u(I^a|\omega)$, there is some $a'\in A^0(\omega)\backslash\{a\}$ with $u(a'|\sigma,\mu)>u(I^{a'}|\omega)$.

[23] Choosing AS' strictly bigger than SE^s_ω makes it easier to check the condition in Proposition 4.2 (as in that case we need not compute SE^s_ω), but it reduces the set of actions that can be excluded. In Cho and Kreps (1987), for example, the set of assessments used for the Intuitive Criterion is bigger than that used for NWBR (the latter coincides with SE_ω), which makes the Intuitive Criterion less powerful but easier to apply. See Section A.2 in the appendix for a discussion of forward induction equilibria (Cho, 1987), which are obtained through a procedure resembling the one we use to extend the Intuitive Criterion to games in extensive form.

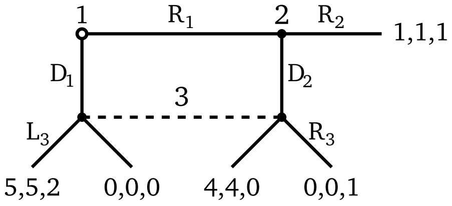

Figure 6

assigning probability zero to $(\mathrm{T}_0, \mathrm{B}_1)$ is T_3, and (iii) if player 3 plays T_3, player 1 strictly prefers to choose B_1. Therefore $\hat{\omega}$ is not self-justifiable.[24]

Example 4.2 (Selten's horse). Consider the game in Figure 6, which is a version of *Selten's horse* (see Figure 1 in Selten, 1975, where the two fives are replaced by two threes). This game has two sequential-equilibrium outcomes: $\omega := (\mathrm{D}_1, \mathrm{L}_3)$ and $\hat{\omega} := (\mathrm{R}_1, \mathrm{R}_2)$. We now show that $\hat{\omega}$ is not self-justifiable. Note that, under $\hat{\omega}$, whenever a strategy by player 3 is such that D_2 is a weak best response for player 2, D_1 is a strict best response for player 1. Thus D_2 can be excluded "in the spirit of D1". Since L_3 is a best response by player 3 to the belief assigning probability zero to $(\mathrm{R}_1, \mathrm{D}_2)$, there is no sequential equilibrium with outcome $\hat{\omega}$ with this belief, so $\hat{\omega}$ is not self-justifiable. As a result, ω is the unique fully self-justifiable (and sequentially stable) outcome.

4.3 Transforming games with signaling into signaling games

In this section we present another way to study non-signaling games using the equilibrium selection criteria defined for signaling games. Broadly, the technique is to transform part of a non-signaling game into an "equivalent" signaling game, then apply the signaling game criteria for this part. If an outcome fails to pass a given selection criterion in the signaling game, we conclude that it is not fully self-justifiable in the main game.

On-path subforms

We first provide a result that permits checking the full self-justifiability of an outcome by checking its full self-justifiability in continuation plays. To provide a more useful result, we will use the concept of subforms, introduced in Kreps and Wilson (1982), which is a generalization of the concept of subgames.

A *subform* is a collection of histories $H' \subset H$ that is closed under succession and preserves information sets; that is, for all $I' \in \mathcal{I}$ and $h' \in I'$, if $h' \in H'$, then (i) $h'' \in H'$ whenever $h'' \succ h'$, and (ii) $I' \subset H'$. Given a subform H', we let H'_0 denote the set of its minimal histories, that is, the set

[24] Note that the argument is even simpler when we analyze full self-justifiability, because then one does not need to verify that there is a sequential equilibrium with outcome $\hat{\omega}$ in which T_2 is a weak best response.

of all $h' \in H'$ such that there is no $h'' \in H'$ satisfying $h'' \prec h'$. For example, a subgame is a subform H' with $|H'_0| = 1$. Given a subform H' and a distribution $\pi_0 \in \Delta(H'_0)$ with full support, $G(H', \pi_0)$ denotes the game constructed by letting nature initially choose a minimal history in H' according to π_0, and then proceeding as in G. We say that H' has full support under ω if all $h' \in H'_0$ occur with positive probability under ω.

Our first result is that the restriction of a fully self-justifiable outcome to an on-path subform is fully self-justifiable. This is important in games with private information, as subforms often have the structure of a signaling game.

Proposition 4.3. *An outcome ω is fully self-justifiable if and only if its continuation in $G(H', \omega|_{H'_0})$ is fully self-justifiable for any subform H' with full support under ω.*

Proposition 4.3 provides a systematic procedure to rule out the full self-justifiability of an outcome by proving it is not fully self-justifiable in an on-path subform. For example, if a player observes the history $h'_0 \in H'_0$ and plays afterward, then h'_0 can be interpreted as its type in the continuation game. If such a continuation game is a signaling game (or can be transformed into a signaling game using the results below), then the continuation outcome must pass all selection criteria in Banks and Sobel (1987) and Cho and Kreps (1987). The proof takes advantage of the flexibility in the implementations' construction in the definition of full self-justifiability (with respect to self-justifiability): any implementation of $\text{IENWBR}_{\omega'}$ in $G(H', \omega|_{H'_0})$ (where ω' is the continuation outcome in H') is an implementation of IENWBR_{ω} in G.

Subgame replacement

Subgame perfection (Selten, 1965) lends plausibility to equilibrium behavior by ensuring that players continue playing mutual best responses even off the path of play. Fully self-justifiable outcomes inherit this plausibility, as they are supported by sequential equilibria, which are subgame perfect. Here we exploit another advantage of subgame perfection: it enables backward induction arguments to simplify game analysis.

Recall that for a given history $\hat{h} \in H$, $Z^{\hat{h}}$ denotes the set of terminal histories succeeding $\hat{h}$. For any outcome ω, we let $\omega^{\hat{h}}$ assign $\omega(z)$ to any $z \notin Z^{\hat{h}}$ and assign $\omega(Z^{\hat{h}})$ to $\hat{h}$. Our second result gives conditions under which one can analyze outcomes in a simplified game, where certain subgames are replaced by payoff profiles.

Proposition 4.4. *Let $\hat{G}$ be a subgame of G starting at history $\hat{h}$ with a unique sequential-equilibrium outcome $\hat{\omega}$. Let G' be the game obtained by removing $\hat{G}$ from G and setting $u'(\hat{h}) := u(\hat{\omega})$. Then ω is fully self-justifiable if and only if $\omega^{\hat{h}}$ is fully self-justifiable in G' and, if $\hat{h}$ is on the path of ω, the continuation outcome after $\hat{h}$ coincides with $\hat{\omega}$.*

Proposition 4.4 allows simplifying the game by replacing subgames with the corresponding sequential equilibrium payoffs. We focus on subgames instead of subforms because subforms may contain degenerate distributions over their initial nodes, which our definition of a game does not allow.

Agent-equivalence

Signaling games feature exactly two players (the sender and the receiver), who have multiple information sets. This structure is appropriate for applications in which a sender's type is private information and the sender interacts with the same receiver regardless of this information. In other applications, however, the sender's type may indicate an intrinsic characteristic (e.g., ability), so that it is more natural for different types to correspond to different players. Similarly, while in some applications the receiver is the same player regardless of what message is sent, in other applications the message may be more naturally interpreted as the choice of the sender to interact with one receiver rather than another. We now establish that replacing a player by multiple agents having the same payoff (each playing in a different information set) does not affect the set of fully self-justifiable outcomes.

Definition 4.1. We say that G and $\hat{G}$ are *agent-equivalent* if $H=\hat{H}$, $\iota^{-1}(0)=\hat{\iota}^{-1}(0)$, $\pi=\hat{\pi}$, and $u_{\iota(a)}(z)=\hat{u}_{\hat{\iota}(a)}(z)$ for all $a\in A$ and $z\in Z^a\equiv\{z\in Z|\exists j, z_j=a\}$.

Proposition 4.5. *If $\hat{G}$ and G are agent-equivalent, they have the same set of fully self-justifiable outcomes.*

Proposition 4.5 is useful because it allows us to treat similar games consistently. For example, two players who never play along the same history can always be interpreted as agents of the same player. As a result, the set of fully self-justifiable outcomes of a signaling game is independent of whether each sender type represents a different player, or whether the receiver is the same after each message or not.[25]

Example

Example 4.3. We now apply the previous results to construct a signaling game that is equivalent to the game in panel (a) of Figure 7, which coincides with Figure 3. First, we use Proposition 4.5 to merge players 1 and 2 into a single player, denoted by player $1'$, who acts as the sender. Second, using Proposition 4.4, we can add a move for player 3 to play after T_1 and B_2; player 3 can thus be viewed as the receiver. Panel (b) of Figure 7 depicts the resulting signaling game. In

[25] For example, Kohlberg and Mertens (1986) use the beer–quiche game of Cho and Kreps (1987) but assume that the two sender types represent different players (see Figure 14 in their paper).

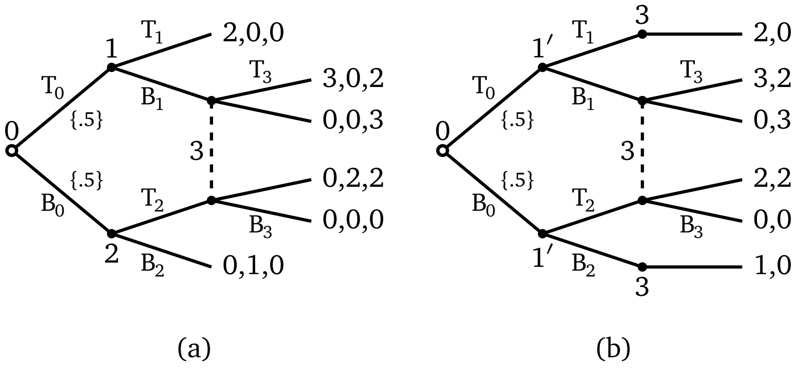

Figure 7

this game, the outcome $\hat{\omega}$ defined in (2.1) fails D1: for any best response of player 3 where B_1 is weakly optimal, T_2 is strictly optimal, so B_1 can be excluded according to the D1 criterion. Since no other action can be excluded, and there is no sequential equilibrium with outcome $\hat{\omega}$ in which player 3 assigns probability zero to $(\mathrm{B}_0, \mathrm{T}_2)$, $\hat{\omega}$ is not self-justifiable.

4.4 Signaling games with prior investment

In this section, we apply our analysis to signaling games with prior investment. These are games where prior to sending the message, a player chooses an investment, which endogenously determines the type distribution. We will show that while these are not signaling games, Cho and Kreps (1987)'s criteria are useful to identify fully self-justifiable and sequentially stable outcomes.

We consider the following signaling game with prior investment. First, an investor chooses an investment $k \in \{1, ..., K\}$. From the chosen investment k, nature draws a type $\theta \in \Theta$ using a full-support distribution $\pi_k \in \Delta(\Theta)$. Then, a sender observes the investment and the type, and chooses a message $m \in M_\theta$. Finally, the receiver, having only observed m, chooses a response $r \in R_m$. We assume that the sender and receiver's payoffs depend additively on a term that depends on (θ, k) and another that depends on (θ, m, r).

Note that numerous economic settings can be studied using signaling games with prior investment. For example, in education signaling, parents typically make the initial educational decisions regarding their children, while the children make their own decisions when they become adults. Additionally, in education signaling, adults may undertake non-observable "investment" decisions (such as time spent studying or engaging in healthy activities) and observable signaling actions (such as pursuing education); our results also apply to the case where the investor and the sender are the same player.

Given an outcome ω of a signaling game with prior investment, we refer to the corresponding *signaling outcome* ω^{sig} as the implied joint distribution over types, messages, and responses. We use G^{sig} to denote the signaling game obtained by removing the investment and where each type receives the same probability as it receives under ω.[26]

Proposition 4.6. *If ω is fully self-justifiable in G, then ω^{sig} is fully self-justifiable in G^{sig}.*

Proposition 4.6 allows the use of standard selection criteria in signaling games with prior investments. In particular, if a message-type pair is excluded for some criterion, then ω is fully self-justifiable (or sequentially stable) only if it is supported by a sequential equilibrium where the receiver assigns a probability zero to it.[27] Note that Proposition 4.6 does not follow from Proposition 4.3 because, except if the investor fully mixes, there is no subform in the signaling game with full support. Still, the logic is similar: if $(A_s^{\text{sig}})_{s=1}^S$ is an implementation of $\text{IENWBR}_{\omega^{\text{sig}}}$ in G^{sig}, then it is also an implementation of IENWBR_ω in G. The same logic can also be generalized to other games where the sender does not observe the investment or where the receiver observes the investment.

5 Conclusions

We have introduced several new equilibrium concepts based on the idea of justifiable beliefs (McLennan, 1985)—that is, on the selection of behavior that is robust to the iterated exclusion of actions that are never weak best responses.

The first concept, self-justifiability, requires that behavior be internally consistent. More concretely, a self-justifiable outcome is supported by behavior consistent with the exclusion of actions that are never weak best responses in sequential equilibria compatible with the outcome. The second concept, full justifiability, requires that behavior be consistent independently of the exclusion procedure used. That is, a fully justifiable outcome is robust to changes in the order of exclusion of actions that are never weak best responses.

The combination of these plausibility requirements (internal consistency and independence from the choice of exclusion procedure) yields the concept of full self-justifiability. Full self-justifiability is a relatively strong selection criterion, in that it implies many previously defined selection criteria. It is also highly flexible (since it can be verified by excluding implausible actions in any order), making it a useful tool in applications. In particular, as we show, our results concerning full self-justifiability make it possible to extend certain intuitive arguments used for signaling games to general games in extensive form. Furthermore, full self-justifiability is implied by sequential stability; hence it provides a way to identify sequentially stable outcomes without considering

[26] More formally, G^{sig} is the signaling game where nature chooses each θ with probability $\omega(Z^\theta)$, where Z^θ is the set of terminal histories containing θ.

[27] Recall that type-message pairs excluded by the selection criteria of Cho and Kreps (1987) and Banks and Sobel (1987) are independent of the distribution of types, as long as it has full support.

sequences of strategy profiles, and thus creates an additional foundation for arguments involving stability in applications.

Overall, our work illuminates the relationships between a number of different equilibrium concepts used in the literature. It provides a unifying framework for these concepts and simplifies many aspects of their analysis, enabling more effective and consistent equilibrium selection across various applications.

In future research, it would be interesting to apply similar plausibility requirements to other base equilibrium concepts. For example, in the iterative procedure defining fully self-justifiable outcomes, one might exclude actions that are never weak best responses under any perfect Bayesian equilibrium (Fudenberg and Tirole, 1991) or any weakly sequential equilibrium (Reny, 1992). While a criterion derived from a weaker base equilibrium concept may have less selection power than ours, it may also be easier to apply and capture richer behavior.

A Additional discussion

A.1 Refinements of implementations

There are many ways to implement IENWBR$_\omega$ for a given $\omega \in \Omega \cup \{\Omega\}$. As explained in Section 3.2, the procedure is to iteratively exclude actions until only plausible actions remain; however, it is not specified which actions, or how many, should be excluded in each step. In this section, we argue that implementations in which smaller sets of actions are excluded in each step are more powerful in selecting outcomes.

Given two non-intersecting sequences $(A'_s)_{s=1}^{S'}$ and $(A_s)_{s=1}^{S}$, we say that $(A'_s)_{s=1}^{S'}$ is a *refinement* of $(A_s)_{s=1}^{S}$ if for every $\hat{s} \in \{1,...,S\}$ there is some $\hat{s}'$ such that $\cup_{s=1}^{\hat{s}'} A'_s = \cup_{s=1}^{\hat{s}} A_s$. Such a sequence $(A'_s)_{s=1}^{S'}$ refines $(A_s)_{s=1}^{S}$ in the sense that in the corresponding iterated exclusion procedure, the set of actions excluded in each step is smaller. Note that if an action a is excluded earlier than a' under $(A_s)_{s=1}^{S}$, then a is excluded earlier than a' under $(A'_s)_{s=1}^{S'}$ as well.

Proposition A.1. *Let $\omega \in \Omega \cup \{\Omega\}$. If $(A'_s)_{s=1}^{S'}$ is a refinement of an implementation $(A_s)_{s=1}^{S}$ of* IENWBR$_\omega$*, then $(A'_s)_{s=1}^{S'}$ is also an implementation of* IENWBR$_\omega$ *and $SE_\omega((A'_s)_{s=1}^{S'}) \subset SE_\omega((A_s)_{s=1}^{S})$.*

The intuition for Proposition A.1 is the following. Let $(A_s)_{s=1}^{S}$ be an implementation of IENWBR$_\omega$. Fix some step $\hat{s} \in \{1,...,S\}$ and some action $\hat{a} \in A$ that is excluded in the $\hat{s}$th step; that is, $\hat{a} \in A_{\hat{s}}$ (and hence $\hat{a} \in NWBR_\omega((A_s)_{s=1}^{\hat{s}-1})$). Let $(A'_s)_{s=1}^{S'}$ be a refinement of $(A_s)_{s=1}^{S}$, let $\hat{s}'$ be such that $\cup_{s=1}^{\hat{s}'-1} A'_s = \cup_{s=1}^{\hat{s}-1} A_s$, and let $\hat{s}'' \geq \hat{s}'$ be such that $\hat{a} \in A'_{\hat{s}''}$. It then follows that

$$NWBR_\omega((A'_s)_{s=1}^{\hat{s}''-1}) \supset NWBR_\omega((A'_s)_{s=1}^{\hat{s}'-1}) \supset NWBR_\omega((A_s)_{s=1}^{\hat{s}-1}) \ .$$

The first containment ($\supset$) holds because, as explained in Section 3.2, the set of never weak best responses increases along any non-intersecting sequence. The second containment follows from Definition 3.1, which implies that if $(A'_s)_{s=1}^{\hat{s}'-1}$ is finer than $(A_s)_{s=1}^{\hat{s}-1}$, then $SE_\omega((A'_s)_{s=1}^{\hat{s}'-1}) \subset SE_\omega((A_s)_{s=1}^{\hat{s}-1})$. Therefore, we have that $\hat{a} \in NWBR_\omega((A'_s)_{s=1}^{\hat{s}''-1})$. It then follows that $A'_{\hat{s}'} \subset NWBR_\omega((A'_s)_{s=1}^{\hat{s}'-1})$ for all $\hat{s}'$, and so $(A'_s)_{s=1}^{S'}$ is an implementation of IENWBR$_\omega$.

Maximal and finest implementations

As we explained in Section 3.3, full self-justifiability is stronger than self-justifiability. This is because the procedure for checking the self-justifiability of an outcome ω corresponds to the *maximal* implementation of IENWBR$_\omega$, where all excludable actions are excluded in each step. (That is, the self-justifiability of ω is verified using the implementation $(A_s)_{s=1}^{S}$ where $A_s = NWBR_\omega((A_{s'})_{s'=1}^{s-1}) \setminus \cup_{s'=1}^{s-1} A_{s'}$ for all s.)

Opposite to the maximal implementation are implementations in which only one action is excluded in each step. Formally, we say an implementation $(A_s)_{s=1}^{S}$ of IENWBR$_\omega$ is *finest* if $|A_s| = 1$ for all s. The following corollary of Proposition A.1 says that to show that ω is fully self-justifiable, it suffices to verify that $SE_\omega((A_s)_{s=1}^{S}) \neq \emptyset$ for all complete and finest implementations.

Corollary A.1. *An outcome ω is fully self-justifiable if and only if $SE_\omega((A_s)_{s=1}^{S}) \neq \emptyset$ for all complete and finest implementations of* IENWBR$_\omega$.

Example A.2 illustrates that, even for complete and finest implementations, the choice of implementation $(A_s)_{s=1}^{S}$ (i.e., the order of exclusion of actions) may affect whether $SE_\omega((A_s)_{s=1}^{S})$ is empty or not.

Remark A.1. We have defined full (self-)justifiability by requiring (self-)justifiability for *all* exclusion orders. As we have argued, this increases the selection power and eases ruling out outcomes, but demonstrating that a given outcome is fully (self-)justifiable becomes more difficult. One could define *partial (self-)justifiability* by requiring (self-)justifiability for *some* exclusion order. This concept would be weaker than (self-)justifiability, easier to demonstrate for a given outcome, but more difficult to disprove. A further alternative would be to require (self-)justifiability by some finest implementation. By Corollary A.1, this alternative would be a compromise: it would be stronger than (self-)justifiability, but weaker than full (self-)justifiability. We leave it to future research to analyze the usefulness of these approaches.

A.2 Forward induction equilibria (Cho, 1987)

In this section, we briefly review the concepts of introspective consistency and forward induction equilibria, introduced by Cho (1987). We also show that any fully self-justifiable outcome is supported by a forward induction equilibrium.

We recall Cho's definitions using our notation. Let $I \in \mathcal{I}$ be an information set. We define $BR_I \subset \Delta(A^I)$ as follows: $\sigma_I \in BR_I$ if there exist a belief system μ and a strategy profile $\tilde{\sigma} \in \Sigma$ such that

$$\mathbb{E}^{(\tilde{\sigma}_{-I},\sigma_I)}[u_{\iota(I)}(z)|\mu,I] \geq \mathbb{E}^{(\tilde{\sigma}_{-I},\sigma'_I)}[u_{\iota(I)}(z)|\mu,I] \text{ for all } \sigma'_I \in \Delta(A^I) .$$

For each $a \in A$,

$$\mathcal{I}^a := \left\{ I' \in \mathcal{I} \,\middle|\, \text{there are } h \in I^a \text{ and } h' \in I' \text{ such that } h \preceq h' \right\} ,$$

where I^a is the (unique) information set where a is available. That is, $\mathcal{I}^a$ is the set of information sets that can be reached if a is played. Given $\sigma \in \Sigma$, we say a is a *bad deviation from* σ if $\sigma(a) = 0$ and, for all $(\tilde{\sigma}_{I'} \in BR_{I'})_{I' \in \mathcal{I}^a}$, we have

$$\mathbb{E}^{((\sigma_I)_{I \notin \mathcal{I}^a}, a, (\tilde{\sigma}_I)_{I \in \mathcal{I}^a \setminus \{I^a\}})}[u_{\iota(I)}(z)] < \mathbb{E}^{\sigma}[u_{\iota(I)}(z)] .$$

We denote by $BAD(\sigma)$ the set of all bad deviations from σ. Finally, for each $I \in \mathcal{I}$ and $\sigma \in \Sigma$, we define

$$J(I|\sigma) := \left\{ h \in I \,\middle|\, h_j \in BAD(\sigma) \text{ for some } j \in \{1, ..., |h|\} \right\} .$$

Cho (1987) defines an assessment (σ, μ) as *introspectively consistent* if there is some sequence of fully mixed strategy profiles (σ_n), with a corresponding sequence of belief systems (μ_n), such that $(\sigma_n, \mu_n) \to (\sigma, \mu)$ as $n \to \infty$, and

$$\mu_n(J(I|\sigma)|I) \to 0 \quad \text{as } n \to \infty \tag{A.1}$$

whenever $J(I|\sigma)$ is a proper subset of I. An assessment (σ, μ) is a *forward induction equilibrium* if it is sequentially rational and introspectively consistent.

Proposition A.2. *If an outcome is fully self-justifiable, then it is the outcome of a forward induction equilibrium.*

A.3 Outcomes that satisfy forward induction (Govindan and Wilson, 2009)

Govindan and Wilson (2009) introduce the concept of *outcomes that satisfy forward induction*. Their main result is that the outcome of a two-player game with perfect recall and generic payoffs satisfies forward induction if it is invariant (in a sense that they define) for the solution concept of sequential equilibrium. Because their approach differs significantly from ours, we only briefly review their definition here, then observe that it resembles the first iteration of our process for obtaining self-justifiable outcomes.

To define forward induction, Govindan and Wilson use the concept of *weakly sequential equilibrium* (Reny, 1992), a weakening of sequential equilibrium in which sequential rationality is imposed only in on-path information sets, and which includes a belief system over the other players' pure strategies at every information set. They define *relevant pure strategies* with respect to an outcome ω.[28] Next they define a *relevant information set* for ω as one that is not excluded by every profile of strategies that are relevant for ω. Finally, they say that ω *satisfies forward induction* if it results from a weakly sequential equilibrium in which, at every information set that is relevant for ω, the support of the belief of the player acting there is confined to profiles of nature's strategies and other players' strategies that are relevant for ω.

We see weakly sequential equilibrium outcomes as the natural analogue to outcomes ω satisfying that $SE^1_\omega \neq \emptyset$, but excluding pure strategies instead of actions. Indeed, the one-step procedure defining weakly sequential equilibrium outcomes prescribes to identify pure strategies that are implausible given ω (non-relevant pure strategies, in their language), and look for weakly sequential equilibria with outcome ω assigning probability zero to these actions. The first step of the definition of self-justifiable outcomes does the same for implausible actions given ω and sequential equilibria with outcome ω. In fact, it is not difficult to see that if G coincides with its agent-extensive form, then any self-justifiable outcome satisfies forward induction.

We believe that our approach offers several advantages. First, our definitions use actions instead of pure strategies, which makes them significantly easier to work with. Second, by allowing actions to be excluded iteratively, we increase the robustness and selection power of our solution concepts. Third, we are able to show that self-justifiable outcomes exist in all games, and to generalize self-justifiability to full self-justifiability, which satisfies additional plausibility properties. Finally, we are able to provide a clear connection between our equilibrium concepts and several others, such as (sequential) stability.

A.4 Proper equilibria (Myerson, 1978)

In this section, we relate fully self-justifiable outcomes with the concept of proper equilibria, which was proposed by Myerson (1978) for normal form games.

For each $\varepsilon > 0$, a fully mixed strategy profile σ of a normal-form game is an *ε-proper equilibrium* if it satisfies that if a pure strategy s_i gives a player a strictly lower payoff than another strategy s'_i, then $\sigma_i(s_i) \leq \varepsilon\, \sigma_i(s'_i)$. Then, σ is a *proper equilibrium* if it is the limit of ε-proper equilibria as

[28] Govindan and Wilson (2009) define a pure strategy as *relevant* with respect to ω if there is a weakly sequential equilibrium with outcome ω for which the strategy is weakly optimal, in the sense that, at every information set it does not exclude, it prescribes an optimal continuation given the player's equilibrium belief over the other players' pure strategies there. Note that if one replaces "weakly sequential equilibrium" by "sequential equilibrium" and "pure strategy" by "action", then the definition resembles the opposite of the definition of a useless action given ω.

$\varepsilon \to 0$. Myerson shows that for any proper equilibrium outcome of the normal form game G, there is an equivalent sequential equilibrium outcome in any extensive form game with normal form G.

Because of the different approaches in their definitions, it is difficult to compare the concepts of proper equilibria and fully self-justifiable outcomes. Still, we believe that characterizing fully self-justifiable outcomes is in general simpler, which is especially helpful in characterizing stability. Indeed, showing properness requires guessing a sequence of fully mixed strategy profiles (of the normal form game) converging to the candidate strategy profile and verifying that it satisfies the corresponding conditions for an appropriate sequence $(\varepsilon_n) \to 0$.[29] Instead, our procedure is set in the limit and defined in terms of actions (instead of contingent plans), hence the reasoning in excluding actions is often intuitive and does not require using sequences of strategy profiles.

The requirement of properness can be used to identify useless actions in the process of obtaining fully self-justifiable outcomes. For example, if there are two actions $a, a' \in A^I$ satisfying that a gives a strictly lower payoff than a' for all sequential equilibria with outcome ω, then a is never a weak best response, hence a is useless (this argument can be applied at any iteration of the IENWBR_ω). As a result, if ω is fully self-justifiable, it satisfies a version of properness: it is supported by a sequence of Nash equilibria of perturbed games where the asymptotic probability of a is infinitely smaller than the asymptotic probability of a' (e.g., of the form (2.2)).

A.5 Further examples

Example A.1. This example shows that full self-justifiability is stronger than self-justifiability. We study the game in Figure 8, nicknamed the "big fish." Consider the outcome $\omega := \frac{1}{2}(\mathrm{T}_0, \mathrm{T}_1) + \frac{1}{2}(\mathrm{B}_0, \mathrm{B}_1')$. This outcome is sequential: for example, there is a sequential equilibrium with outcome ω in which player 2 plays T_2, player 3 plays T_3, player 4 plays the middle action, and player 2 believes that each of the histories in her information set is equally likely (the other beliefs are obtained through Bayes' rule). We will show that ω is self-justifiable but not fully self-justifiable.

Our argument runs as follows: we prove that (i) the off-path actions B_1 and T_1' are both useless under ω, (ii) when B_1 is ruled out first, there is a surviving sequential equilibrium with outcome ω, and (iii) when T_1' is ruled out first, there is no such sequential equilibrium. The difficulty in constructing such an example is that B_1 and T_1' will be profitable deviations under any continuation play consistent with beliefs where T_1' is ruled out first, while are not weak best responses in any sequential equilibrium with outcome ω when B_1 is ruled out first. In many games, this would imply that there is a belief system (assigning comparable probabilities to histories after B_1 and

[29] Note that the requirement of ε_n-properness of each element σ_n approaching the candidate σ is self-referential: pure strategies giving a low payoff under σ_n must receive a low probability under σ_n. Such a requirement is imposed even for pure strategies that give the same payoff in the limit, because their payoff may be slightly different along the sequence.

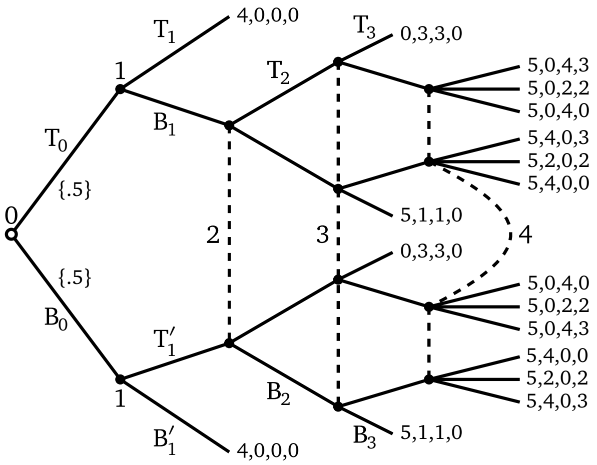

Figure 8

histories after T$_1'$) in which one of the actions is suboptimal and the other is a weak best response, but in such cases it would be impossible for both B$_1$ and T$_1'$ to be useless under ω. In the game in Figure 8, however, the equilibrium payoff correspondence of the continuation game is not lower-hemicontinuous in the relative probability of histories after B$_1$ and histories after T$_1'$.

To begin, we fix a sequential equilibrium (σ, μ) and let x denote the probability with which player 4 plays the middle action. The continuation payoffs of players 2 and 3 at their information sets, given in Table 1, depend only on x (and not on their beliefs regarding whether player 1 has chosen B$_1$ or T$_1'$). For $x < 1/2$, the game in Table 1 has a unique Nash equilibrium, which we denote by E$_1$. In this equilibrium, players 2 and 3 play B$_2$ and B$_3$, respectively, and both obtain 1, so that player 1 obtains 5. For $x = 1/2$, there are two Nash equilibria: E$_1$ (as just described), and another, denoted by E$_2$, in which players 2 and 3 play T$_2$ and T$_3$, respectively, so that player 1 obtains 0. For $x > 1/2$, there are three Nash equilibria: E$_1$, E$_2$, and another one, denoted by E$_3$, in which players 2 and 3 randomize according to

$$\left(\tfrac{1}{2x}\,\mathrm{T}_2 + \tfrac{2x-1}{2x}\,\mathrm{B}_2, \tfrac{1}{2x}\,\mathrm{T}_3 + \tfrac{2x-1}{2x}\,\mathrm{B}_3\right).$$

The latter gives players 2 and 3 each a payoff of $3/(2x) \in [3/2, 3)$, while player 1 obtains $5(1 - 1/(4x^2)) \in (0, 15/4]$. Panel (a) of Figure 9 depicts, for the game in Table 1, the equilibrium probabilities with which T$_2$ is played (which are the same as the probabilities with which T$_3$ is played), as functions of x. Panel (b) depicts player 1's payoffs in the equilibria of the subform beginning at player 2's information set as functions of x. Panel (c) depicts player 1's payoffs in the equilibria of the subform in relation to the probability player 2 assigns to (T$_0$, B$_1$).

We see that, for (σ, μ) to be a sequential equilibrium with outcome ω, either E$_2$ or E$_3$ must be played in the continuation play when player 1 chooses B$_1$ or T$_1'$ (since player 1 obtains $5 > 4 = u_1(\omega)$ under E$_1$). This implies that $x \geq 1/2$. Hence player 1's payoff from choosing B$_1$ after T$_0$ or choosing

	T_3	B_3
T_2	$3, 3$	$0, x2+(1-x)4$
B_2	$x2+(1-x)4, 0$	$1, 1$

Table 1

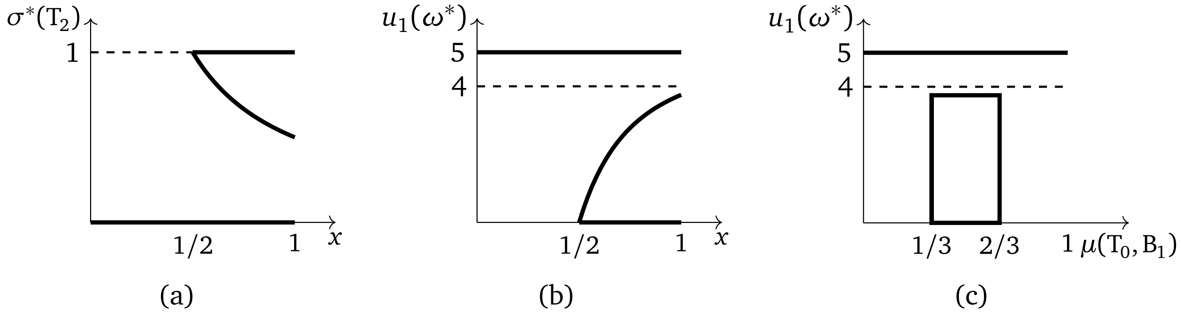

Figure 9

T_1' after B_0 is either 0 or $5(1-1/(4x^2)) \leq 15/4 < 4$, which makes it strictly optimal for her to choose either T_1 or B_1'.

These observations imply that B_1 and T_1' are never weak best responses of sequential equilibria with outcome ω. Hence they are excluded in the first step of the procedure for verifying self-justifiability. Since no other actions can be excluded in later steps, and since the sequential equilibrium with outcome ω described above has justifiable beliefs, we conclude that ω is self-justifiable.

To see that ω is not fully self-justifiable, suppose we exclude B_1 first. Then, in every sequential equilibrium in $SE_\omega(\{B_1\})$, player 4 must assign probability 0 to (T_0, B_1, T_2, B_3) and (T_0, B_1, B_2, T_3), which implies that $x = 0$. But, as we argued above, there is no such sequential equilibrium with outcome ω. (Note that the same argument applies if T_1' is excluded first.) It is easy to see that the unique fully self-justifiable outcome (and hence the unique sequentially stable outcome) is one in which player 1 chooses B_1 and T_1', player 2 chooses B_2, and player 3 chooses B_3.

Observe that every action of player 1, 2, or 3 is a best response under some sequential equilibrium, which implies that ω is fully justifiable. This shows that a game may have outcomes that are both fully justifiable and self-justifiable, but are not fully self-justifiable.

Example A.2. This example shows that, even for complete and finest implementations of IENWBR_ω, the choice of implementation (i.e., the order of exclusion of actions) may affect whether the set of justifiable equilibria resulting from the exclusion procedure is empty or not. Consider the game in Figure 10(a) with $y = 5$. Let ω be the outcome assigning probability one to L_1. Note that there is a sequential equilibrium with such an outcome—for example, one in which player 2 chooses T_2, player 3 chooses T_3, player 4 chooses T_4, and player 2 assigns probability one to T_1 (the other

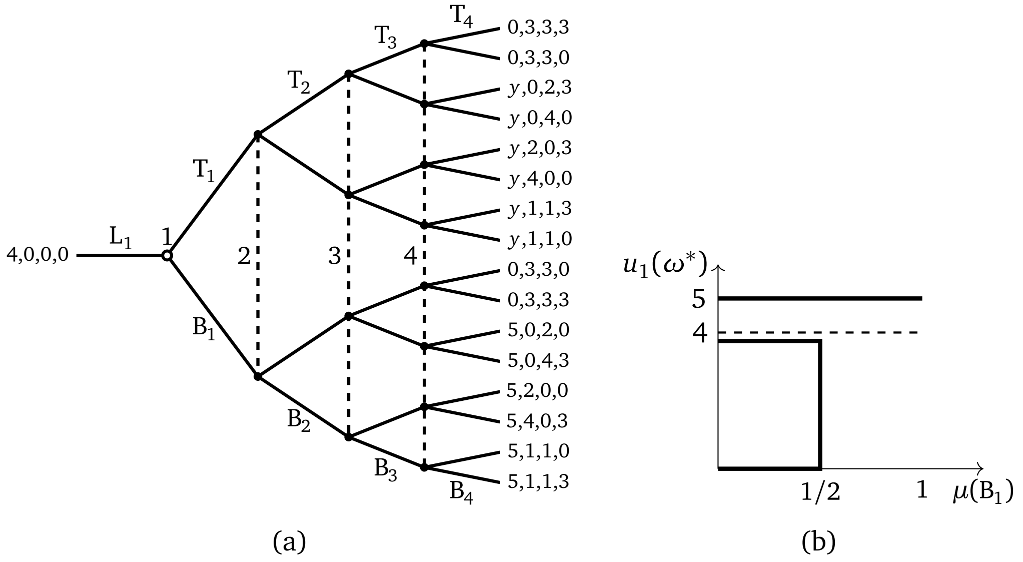

Figure 10

beliefs are pinned down by Bayes' rule).

Fix a sequential equilibrium with outcome $\omega = \mathrm{L}_1$. Let x denote the probability with which player 4 chooses T_4 in this equilibrium. The expected payoffs of players 2 and 3 conditional on player 1 not playing L_1 are given by Table 1; these payoffs are independent of the beliefs of players 2 and 3 concerning the action played by player 1. As discussed in Example A.1, if $x < 1/2$, then player 1's payoff from choosing T_1 or B_1 is 5; but then player 1 has a strict incentive not to choose L_1. So it must be that $x \geq 1/2$. If $x \geq 1/2$ and if player 1 chooses T_1 or B_1, then under any continuation play she obtains either 0 or $5\,(1-1/(4x^2)) \in (0, 15/4]$. Thus, neither T_1 nor B_1 is a weak best response in any sequential equilibrium with outcome ω.

Now consider the following two implementations of IENWBR$_\omega$. In the first implementation, T_1 is excluded in the first step, which corresponds to $x = 0$. Since there is no sequential equilibrium with outcome ω in which $x = 0$, we have $SE_\omega(\{\mathrm{T}_1\}) = \emptyset$; that is, ω fails this implementation. In the second implementation, B_1 is excluded first, which corresponds to $x = 1$; that is, player 4 chooses T_4 for sure. Next we exclude B_4, and finally T_1 (no other action can be excluded). Note that

$$SE_\omega((\{\mathrm{B}_1\}, \{\mathrm{B}_4\}, \{\mathrm{T}_1\}))$$

is non-empty: it contains the sequential equilibrium described at the beginning of this example. Hence ω passes this implementation of IENWBR$_\omega$.

Example A.3. This example shows that a game may have outcomes that are self-justifiable but not justifiable. Consider the game in Figure 10(a) with $y = 0$. Note that $(\mathrm{B}_1, \mathrm{B}_2, \mathrm{B}_3, \mathrm{B}_4)$ is a sequential-equilibrium outcome, as is L_1; the latter is sustained by a belief system assigning probability one to T_1, with players 2, 3, and 4 playing T_2, T_3, and T_4, respectively, with probability one.

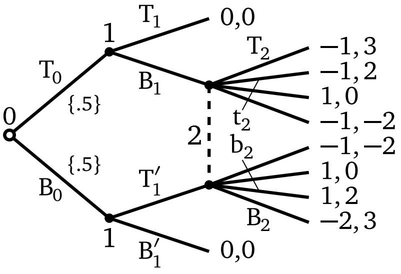

Figure 11

We first prove that (B_1, B_2, B_3, B_4) is the unique justifiable outcome. To see this, note that T_1 is strictly dominated by L_1 (since $y = 0$). Because each of the other actions is played under some sequential equilibrium, T_1 is the unique useless action. Now note that player 4 plays B_4 with probability one in all equilibria with $\mu(T_1) = 0$. Hence, as we argued in Example A.2, (B_2, B_3, B_4) is the unique continuation outcome if player 1 does not choose L_1, which proves that (B_1, B_2, B_3, B_4) is the unique justifiable outcome. Now, the continuation payoff of player 1 upon entering in any equilibrium of the continuation game as a function of the belief on B_1 is depicted in Figure 10(b).

We now show that L_1 is self-justifiable. Indeed, the arguments in Example A.2 imply that both T_1 and B_1 are useless under L_1, and there are no other (higher-order) useless actions. Therefore, because $SE^1_{L_1} = SE_{L_1} \neq \emptyset$, we have that L_1 is self-justifiable.

Example A.4. We now provide an example of a fully self-justifiable outcome that is not a sequentially stable outcome. Consider the game in Figure 11, which is based on Figure 3 in Banks and Sobel (1987). Fix the outcome $\omega := \frac{1}{2}(T_0, T_1) + \frac{1}{2}(B_0, B_1')$. We first observe that there are no useless actions under ω. Indeed, the following are two equilibria with outcome ω:

$$(\sigma,\mu) := \Big(\overbrace{\tfrac{1}{2}T_0 + \tfrac{1}{2}B_0}^{\sigma_0}, \overbrace{T_1, B_1'}^{\sigma_1}, \overbrace{\tfrac{1}{2}T_2 + \tfrac{1}{2}t_2}^{\sigma_2}, \overbrace{\tfrac{2}{3}(T_0, B_1) + \tfrac{1}{3}(B_1, T_2)}^{\mu_2}\Big) \text{ and}$$

$$(\hat{\sigma},\hat{\mu}) := \Big(\tfrac{1}{2}T_0 + \tfrac{1}{2}B_0, T_1, B_1', \tfrac{1}{2}b_2 + \tfrac{1}{2}B_2, \tfrac{1}{3}(T_0, B_1) + \tfrac{2}{3}(B_1, T_1')\Big) .$$

Note that T_1' is a weak best response under (σ, μ) and B_1 is a weak best response under $(\hat{\sigma}, \hat{\mu})$. Hence ω is fully self-justifiable. However, as Banks and Sobel show, ω is fragile to a tremble sequence that induces a posterior $(0.49(T_0, B_1) + 0.51(B_0, T_1'))$ in player 2's information set.

This example illustrates the fact that outcomes that can only be destabilized with tremble sequences inducing positive posteriors over more than one history are not sequentially stable but may be fully self-justifiable.

B Proofs

Proof of Proposition 2.1

Proof. The proof follows from Propositions 3.1 and 3.3 (note that results in Section 3 do not build on the results in Section 2). □

Proof of Proposition 3.1

Proof. The proof follows from the arguments in the main text. □

Proof of Proposition 3.2

Proof. Assume ω is fully self-justifiable. Let $(A_s)_{s=1}^{S}$ be an implementation of IENWBR$_\Omega$. We note that, for all s, we have

$$SE_\omega((A_{s'})_{s'=1}^{s-1}) \subset SE_\Omega((A_{s'})_{s'=1}^{s-1}) \ .$$

This implies that

$$NWBR_\omega((A_{s'})_{s'=1}^{s-1}) \supset NWBR_\Omega((A_{s'})_{s'=1}^{s-1}) \ .$$

Hence, we have that $A_s \subset NWBR_\omega((A_{s'})_{s'=1}^{s-1})$ and $A_s \neq \emptyset$ for all $s < \hat{s}$. Therefore, $(A_s)_{s=1}^{S}$ is an implementation of IENWBR$_\omega$. Since ω is fully self-justifiable, we have that $SE_\omega((A_s)_{s=1}^{S}) \neq \emptyset$, which implies that $SE_\Omega((A_s)_{s=1}^{S}) \neq \emptyset$. Since this applies to all implementations of IENWBR$_\Omega$, we have that ω is fully justifiable. □

Proof of Proposition 3.3

Proof. Assume that ω is sequentially stable. We will prove that it is fully self-justifiable; hence, by Propositions 3.1 and 3.2, it is also justifiable, self-justifiable, and fully justifiable. Assume, for the sake of contradiction, that ω is not fully self-justifiable, hence it does not pass IENWBR$_\omega$. Let $(A_s)_{s=1}^{S}$ be an implementation of IENWBR$_\omega$ such that $SE_\omega((A_s)_{s=1}^{S}) = \emptyset$. Consider a tremble sequence (ξ_n) defined as follows

$$\xi_n(a) := \begin{cases} e^{-n^{S-s+2}} & \text{if } a \in A_s \text{ for some } s \in \{1, \ldots, S\}, \\ e^{-n} & \text{otherwise,} \end{cases}$$

for all $a \in A$ and $n \in \mathbb{N}$. Let $(u_n) \to u$ and (σ_n) be two sequences satisfying that each ω^{σ_n} is a Nash equilibrium outcome of $G(\xi_n, u_n)$ and $\omega^{\sigma_n} \to \omega$ (which exist because ω is sequentially stable).

Let $(\hat{n}_n)\to\infty$ satisfy that (i) for each a, either $\sigma_{\hat{n}_n}(a)>\xi_{\hat{n}_n}(a)$ for all k or $\sigma_{\hat{n}_n}(a)=\xi_{\hat{n}_n}(a)$ for all n, and (ii) $(\sigma_{\hat{n}_n})$ supports some assessment (σ,μ) (which has outcome ω). By Proposition 2.1 in Dilmé (2024), (σ,μ) is a sequential equilibrium. There are two possibilities.

1. The first possibility is that all actions in $\cup_{s=1}^{S}A_s$ are not weak best responses under (σ,μ). In this case it must be that $\sigma_{\hat{n}_n}(a)=\xi_{\hat{n}_n}(a)$ for all $n\in\mathbb{N}$ and $a\in\cup_{s=1}^{S}A_s$. We argue that necessarily $(\sigma,\mu)\in SE_\omega((A_s)_{s=1}^{S})$, which will contradict the assumption that $SE_\omega((A_s)_{s=1}^{S})=\emptyset$. To see that, let $\tilde{s}(a)$ be the value s such that $a\in A_s$ if such value exists, and $S+1$ if no such value exists. For each history $h\in H$, let $\#\tilde{s}_s(h)$ be the number of actions in h that have an order of uselessness under ω equal to s. Take two histories $h,h\in H'$, and assume that there is some $\hat{s}\leq S$ such that $\#\tilde{s}_s(h)=\#\tilde{s}_s(h')$ for all $s>\hat{s}$ and $\#\tilde{s}_{\hat{s}}(h)>\#\tilde{s}_{\hat{s}}(h')$ (i.e., they are ordered in the sense of Definition 3.1). Then,
$$\log\left(\frac{\sigma_n(h)}{\sigma_n(h')}\right)=-\sum_{s=1}^{S}n^{S-s+2}\left(\#\tilde{s}_s(h)-\#\tilde{s}_s(h')\right)+\log\left(\frac{\prod_{j|\tilde{s}(h_j)=S+1}\sigma_n(h_j)}{\prod_{j|\tilde{s}(h'_j)=S+1}\sigma_n(h'_j)}\right).$$
The first term on the right side tends to $-\infty$ at rate $n^{S-\hat{s}+2}$, while the second term grows at most linearly in n. Because $S-\hat{s}+2>1$, it follows that $\mu(h)=0$. This implies that $(\sigma,\mu)\in SE_\omega((A_s)_{s=1}^{S})$, a contradiction.

2. The second possibility is that some actions in $\cup_{s=1}^{S}A_s$ are best responses under (σ,μ). In this case, let $\hat{s}\in\{1,...,S\}$ be the smallest value such that there is some $a_{\hat{s}}\in A_{\hat{s}}$ that is a weak best response under (σ,μ). This implies that $(\sigma,\mu)\in SE_\omega((A_s)_{s=1}^{\hat{s}})$. Still, that $a_{\hat{s}}\in NWBR_\omega((A_s)_{s=1}^{\hat{s}})$ implies that there is no sequential equilibrium $SE_\omega((A_s)_{s=1}^{\hat{s}})$ where $a_{\hat{s}}$ is a weak best response. Again, we have a contradiction.

□

Proof of Proposition 4.1

Proof. The proof follows from the argument in the main text that follows the proposition. □

Proof of Proposition 4.2

Proof. Fix some $\omega\in\Omega$ and $a\in A^0_\omega$, and let $SE'\subset SE_\omega$ and $AS'\supset SE'$. Assume that, for each $(\sigma,\mu)\in AS'$ where $u(a|\sigma,\mu)=u(I^a|\omega)$, there is some $a'\in A^0_\omega$ such that $u(a'|\sigma,\mu)>u(I^{a'}|\omega)$. Assume, on the way to a contradiction, that a is a weak best response under some $(\sigma,\mu)\in SE'$, that is, $u(a|\sigma,\mu)=u(I^a|\omega)$. Then, because $(\sigma,\mu)\in AS'$, there is some $a'\in A^0_\omega$ such that $u(a'|\sigma,\mu)>u(I^{a'}|\omega)$, but this contradicts that (σ,μ) is sequentially rational (because $\omega^\sigma=\omega$ and so sequential rationality implies that $u(a'|\sigma,\mu)\leq u(I^{a'}|\omega)$ for all $a'\in A^0_\omega$). □

Proof of Proposition 4.3

Proof. The "if" direction is trivial when the condition is applied to the subform initiated at the empty history. To prove the "only if" direction, let ω be fully self-justifiable. Let H' be a subform such that all $h' \in H'_0$ occur with positive probability under ω. Let ω' denote the continuation outcome of ω in $G(H', \omega|_{H'_0})$. Assume for a contradiction that ω' is not fully self-justifiable in $G(H', \omega|_{H'_0})$. Then, there is some implementation $(A'_s)_{s=1}^S$ of IENWBR$_{\omega'}$ such that $SE_{\omega'}((A'_s)_{s=1}^S) = \emptyset$. Nevertheless, this implies that $SE_{\omega}((A'_s)_{s=1}^S) = \emptyset$ because for any $(\sigma, \mu) \in SE_{\omega}((A'_s)_{s=1}^S)$ it must be that the continuation outcome in $G(H', \omega|_{H'_0})$ is a sequential-equilibrium outcome of $G(H', \omega|_{H'_0})$ which is $(A'_s)_{s=1}^S$-justifiable. This contradicts that ω is fully self-justifiable. □

Proof of Proposition 4.4

Proof. Let ω be a fully self-justifiable outcome of G. Note that, in all sequential equilibria, the continuation play in $\hat{G}$ coincides with $\hat{\omega}$. It is then easy to see that (σ, μ) is a sequential equilibrium of G if and only if $(\sigma|_{A'}, \mu|_{H'})$ is a sequential equilibrium of $\hat{G}$ and the continuation outcome of (σ, μ) in G' is equal to ω'. The result then follows immediately from this observation. □

Proof of Proposition 4.5

Proof. The statement follows from the fact that the sets of sequential equilibria satisfying the same belief restrictions are identical in agent-equivalent games. Hence, $\hat{G}$ and G are agent-equivalent. They have the same set of implementations of IENWBR$_{\omega}$ and, for any implementation $(A_s)_{s=1}^S$ of IENWBR$_{\omega}$, ω is either $(A_s)_{s=1}^S$-justifiable in both games or none. □

Proof of Proposition 4.6

Proof. In this proof, we fix some fully self-justifiable outcome ω of the signaling game with prior investment G. To ensure that each action is played in a single information set, we use the following notation. We use $a_{k,\theta}$ to denote the action of nature choosing θ after k (note that nature's information set is different after each investment level k). Also, we use $a_{k,\theta,m}$ to denote the sender's choice of message $m \in M_{\theta}$ after investment k and type θ. Finally, we use $a_{m,r}$ to denote the receiver's choice of response $r \in R_m$ after investment k, type θ, and message m.

We use a similar notation for the corresponding signaling game G^{sig}, that is, we use a^{sig}_{θ}, $a^{\text{sig}}_{\theta,m}$, and $a^{\text{sig}}_{m,r}$ denote actions in this game. For each $a^{\text{sig}} \in A^{\text{sig}}$, we let $A_{a^{\text{sig}}} := \{(k,a) | k = 1, ..., K\}$; that is, $A_{a^{\text{sig}}} := \{(k,a) | k = 1, ..., K\}$ contains the actions in the signaling game with prior investments with the same type, message, and response as a^{sig}. Note that in G^{sig}, nature chooses each type θ with probability $\omega(Z^{\theta})$.

Assume for a contradiction that ω^{sig} is not fully self-justifiable in G^{sig}. This implies that there is some $(A_s^{\text{sig}})_{s=1}^S$ of $\text{IENWBR}^{\text{sig}}_{\omega^{\text{sig}}}$ with $SE^{\text{sig}}_{\omega^{\text{sig}}}((A_s^{\text{sig}})_{s=1}^S)=\emptyset$. For each $s=1,...,S$, define

$$A_s := \cup_{a^{\text{sig}} \in A_s^{\text{sig}}} A_{a^{\text{sig}}} \ .$$

We will now show that $(A_s)_{s=1}^S$ is an implementation of IENWBR_ω (in G). Note that $(A_s)_{s=1}^S$ is non-intersecting. Assume for a contradiction that there is some s, $\hat{a} \in A_s$, and an equilibrium $(\sigma, \mu) \in SE_\omega((A_{\hat{s}})_{\hat{s}=1}^{s-1})$ where $\hat{a}$ is a weak best response. Let $\hat{a}^{\text{sig}} \in A_s^{\text{sig}}$ be such that $\hat{a} \in A_{\hat{a}^{\text{sig}}}$, and note that it must be that $\hat{a}^{\text{sig}} \in A_s^{\text{sig}}$. Define for all $\theta \in \Theta$, $m \in M_\theta$, and $r \in R_m$,

$$\sigma^{\text{sig}}(a_\theta^{\text{sig}}) := \frac{\sum_{k \in K} \omega(Z^k)\,\sigma(a_{k,\theta})}{\sum_{k \in K} \omega(Z^k)} \ , \qquad \sigma^{\text{sig}}(a_{\theta,m}^{\text{sig}}) := \frac{\sum_{k \in K} \omega(Z^k)\,\sigma(a_{k,\theta})\,\sigma(a_{k,\theta,m})}{\sum_{k \in K} \omega(Z^k)\,\sigma(a_{k,\theta})} \ ,$$

and $\sigma^{\text{sig}}(a_{m,r}^{\text{sig}}) := \sigma(a_{m,r})$, where Z^k is the set of terminal histories that contain k. Define also, for all $\theta \in \Theta$ and $m \in M$,

$$\mu^{\text{sig}}(a_\theta^{\text{sig}}, a_{\theta,m}^{\text{sig}}) = \sum_{k=1}^{K} \mu(k, a_{k,\theta}, a_{k,\theta,m}) \ .$$

It is not difficult to see that $(\sigma^{\text{sig}}, \mu^{\text{sig}}) \in SE^{\text{sig}}_{\omega^{\text{sig}}}((A_{\hat{s}}^{\text{sig}})_{\hat{s}=1}^{s-1})$, but this contradicts that $\hat{a}^{\text{sig}} \in A_s^{\text{sig}}$. Hence, $(A_s)_{s=1}^S$ is an implementation of IENWBR_ω.

Finally, because $(A_s)_{s=1}^S$ is an implementation of IENWBR_ω, there is some sequential equilibrium (σ, μ) in $SE_\omega((A_s)_{s=1}^S)$. Using the previous procedure, we can obtain a sequential equilibrium $(\sigma^{\text{sig}}, \mu^{\text{sig}})$ in $SE^{\text{sig}}_{\omega^{\text{sig}}}((A_s^{\text{sig}})_{s=1}^S)$. This contradicts that $SE^{\text{sig}}_{\omega^{\text{sig}}}((A_s^{\text{sig}})_{s=1}^S)=\emptyset$. Hence, ω^{sig} is fully self-justifiable in G^{sig}. □

Proof of Proposition A.1

Proof. Fix $\omega \in \Omega \in \{\Omega\}$ and let $(A'_{s'})_{s'=1}^{S'}$ be a refinement of an implementation $(A_s)_{s=1}^S$ of IENWBR_ω. Assume for a contradiction that $(A'_{s'})_{s'=1}^{S'}$ is *not* an implementation of IENWBR_ω. Let $\check{a}$ and $\check{s}$ be such that (i) $\check{a} \in A'_{\check{s}}$ and $\check{a} \notin NWBR_\omega((A'_s)_{s=1}^{\check{s}-1})$, and (ii) there is no $\check{a}'$ and $\check{s}'$ with $\check{s}' < \check{s}$ such that $\check{a}' \in A'_{\check{s}'}$ and $\check{a}' \notin NWBR_\omega((A'_s)_{s=1}^{\check{s}'-1})$. Let $\hat{s}$ be such that $\hat{a}' \in A_{\hat{s}}$, and let $\hat{s}' \geq \check{s}$ be such that $A'_{\hat{s}'-1} = A_{\hat{s}-1}$ (which exists because $(A'_{s'})_{s'=1}^{S'}$ is a refinement of $(A_s)_{s=1}^S$). Note that

$$NWBR_\omega((A_s)_{s=1}^{\hat{s}-1}) \subset NWBR((A'_s)_{s=1}^{\hat{s}'-1})$$

because $(A'_s)_{s=1}^{\hat{s}'-1}$-justifiability imposes more conditions than $(A_s)_{s=1}^{\hat{s}-1}$)-justifiability. Hence, we have that

$$\check{a} \in NWBR_\omega((A_s)_{s=1}^{\hat{s}-1}) \subset NWBR((A'_s)_{s=1}^{\hat{s}'-1}) \subset NWBR((A'_s)_{s=1}^{\check{s}-1}) \ ,$$

where the last inclusion holds because $\check{s} \geq \hat{s}$. This clearly contradicts our earlier assumption that $\check{a} \notin NWBR_\omega((A'_s)_{s=1}^{\check{s}-1})$. □

Proof of Corollary A.1

Proof. The "only if" implication is trivial. To prove the "if" implication, assume ω is such that $SE_{\omega}((\tilde{A}_s)_{s=1}^{\tilde{S}}) \neq \emptyset$ for all complete and finest implementations $(\tilde{A}_s)_{s=1}^{\tilde{S}}$. Let $(A_s)_{s=1}^{S}$ be an implementation of IENWBR$_{\omega}$, not necessarily complete or finest. Let $\tilde{s}$ be the number of actions in $\cup_{s=1}^{S} A_s$ (that is, the total number of actions that the implementation excludes). Let $(a_s)_{s=1}^{\tilde{s}}$ be a sequence satisfying that (i) $a_s \neq a_{s'}$ for all s, s', (ii) $a_s \in \cup_{s''=1}^{S} A_{s''}$ for all s, and (iii) if $s < s'$, $a_s \in A_{s_1}$, and $a_{s'} \in A_{s_2}$, then $s_1 \leq s_2$. That is, $(a_s)_{s=1}^{\tilde{s}}$ is a sequence of actions satisfying that, if $s < s'$, then a_s is excluded at a weakly earlier step than $a_{s'}$ under $(A_s)_{s=1}^{S}$. Define, for all $s = 1, ..., \tilde{s}$, $A'_s := \{a_s\}$. We now note that

$$A'_s = \{a_s\} \subset NWBR_{\omega}((A_{s'})_{s'=1}^{s-1})$$

for all $s = 1, ..., \hat{s}$, so $(A'_s)_{s=1}^{S}$ is an implementation of IENWBR$_{\omega}$. Because ω is such that $SE_{\omega}((\tilde{A}_s)_{s=1}^{\tilde{S}}) \neq \emptyset$ for all complete and finest implementations $(\tilde{A}_s)_{s=1}^{\tilde{S}}$, we have that $SE_{\omega}((A'_s)_{s=1}^{\tilde{s}}) \neq \emptyset$.

We finally argue that

$$SE_{\omega}((A'_s)_{s=1}^{\tilde{s}}) \subset SE_{\omega}((A_s)_{s=1}^{S}) \ .$$

This follows from the fact that the restriction of belonging to $SE_{\omega}((A'_s)_{s=1}^{\tilde{s}})$ is more restrictive than that of belonging to $SE_{\omega}((A_s)_{s=1}^{S})$. □

Proof of Proposition A.2

Proof. Let ω be fully self-justifiable. We let A_1 be the set of actions $a \in A$ satisfying that (i) I^a is on the path of ω, and (ii) a is never a weak best response for any sequential equilibrium in with outcome ω. Note that A_1 is an implementation of IENWBR$_{\omega}$. Because ω is fully self-justifiable, there is some sequential equilibrium (σ, μ) in $SE_{\omega}((A_s)_{s=1}^{1})$.[30]

Note that if $BAD(\sigma) = \emptyset$, then (σ, μ) is a forward induction equilibrium with outcome ω, so the result is proven. Hence, assume that $BAD(\sigma) \neq \emptyset$ then (σ, μ). We first argue that all actions in $BAD(\sigma)$ are in A_1. Take some $a \in BAD(\sigma)$ and assume, for the sake of contradiction, that there is some (σ', μ') with outcome ω under which a is a best response. Note that, because σ' and σ have the same outcome, we have $\mathbb{E}^{\sigma'}[u_{\iota(I)}(z)] = \mathbb{E}^{\sigma}[u_{\iota(I)}(z)]$ and

$$\mathbb{E}^{((\sigma'_I)_{I \notin \mathcal{I}^a}, a, (\tilde{\sigma}_I)_{I \in \mathcal{I}^a \setminus \{I^a\}})}[u_{\iota(I^a)}(z)] = \mathbb{E}^{((\sigma_I)_{I \notin \mathcal{I}^a}, a, (\tilde{\sigma}_I)_{I \in \mathcal{I}^a \setminus \{I^a\}})}[u_{\iota(I^a)}(z)]$$

for all $(\tilde{\sigma}_{I'} \in BR_{I'})_{I' \in \mathcal{I}^a}$. Furthermore, we have $\sigma'(a) = 0$ and, because $\sigma'_I \in BR_I$ for all $I \in \mathcal{I}$, we also have

$$\mathbb{E}^{((\sigma_I)_{I \notin \mathcal{I}^a}, a, (\sigma'_I)_{I \in \mathcal{I}^a \setminus \{I^a\}})}[u_{\iota(I)}(z)] < \mathbb{E}^{\sigma}[u_{\iota(I)}(z)] \ .$$

[30] Note that if $SE_{\omega}((A_s)_{s=1}^{1})$ was empty, then $SE_{\omega}((A_{\omega}^{s})_{s=1}^{1})$ would be empty as well, and so would be $SE_{\omega}((A_{\omega}^{s})_{s=1}^{S_{\omega}})$ (recall that $SE_{\omega}((A_{\omega}^{s'})_{s'=1}^{s}) \subset SE_{\omega}((A_{\omega}^{s'})_{s'=1}^{s-1}))$ for all s), but that would contradict that ω be fully self-justifiable.

Hence, a is not a weak best response in any equilibrium in $SE_\omega((A_s)_{s=1}^1)$.

Now, consider the non-intersecting sequence $(BAD(\sigma), A_1 \backslash BAD(\sigma))$. Because this is a refinement of $(A_s)_{s=1}^1$, Proposition A.1 ensures that $(BAD(\sigma), A_1 \backslash BAD(\sigma))$ is an implementation of IENWBR$_\omega$. Take some $(\hat{\sigma}, \hat{\mu}) \in SE_\omega(BAD(\sigma), A_1 \backslash BAD(\sigma))$. It is clear that $BAD(\hat{\sigma}) = BAD(\sigma)$ and that condition (A.1) is satisfied. Hence, $(\hat{\sigma}, \hat{\mu})$ is a forward induction equilibrium and its outcome is ω. □

References

Banks, J. S. and J. Sobel (1987): "Equilibrium selection in signaling games," *Econometrica*, 55, 647–661.

Battigalli, P. and M. Siniscalchi (2003): "Rationalization and incomplete information," *Advances in Theoretical Economics*, 3.

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

Cho, I.-K. (1987): "A refinement of sequential equilibrium," *Econometrica*, 55, 1367–1389.

Cho, I.-K. and D. M. Kreps (1987): "Signaling games and stable equilibria," *The Quarterly Journal of Economics*, 102, 179–221.

Daley, B. and B. Green (2012): "Waiting for news in the market for lemons," *Econometrica*, 80, 1433–1504.

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

Dilmé, F. (2019): "Dynamic quality signaling with hidden actions," *Games and Economic Behavior*, 113, 116–136.

——— (2024): "Sequentially stable outcomes," *Econometrica*, 92, 1097–1134.

——— (2025): "Iterated exclusion of implausible types in signaling games," *Games and Economic Behavior*, 152, 293–312.

Ekmekci, M. and N. Kos (2023): "Signaling covertly acquired information," *Journal of Economic Theory*, 214, 105746.

Fudenberg, D. and J. Tirole (1991): "Perfect Bayesian equilibrium and sequential equilibrium," *Journal of Economic Theory*, 53, 236–260.

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

Grossman, S. J. and M. Perry (1986): "Perfect sequential equilibrium," *Journal of Economic Theory*, 39, 97–119.

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

KREMER, I. AND A. SKRZYPACZ (2007): "Dynamic signaling and market breakdown," *Journal of Economic Theory*, 133, 58–82.

KREPS, D. M. AND R. WILSON (1982): "Sequential equilibria," *Econometrica*, 50, 863–894.

MAILATH, G. J., M. OKUNO-FUJIWARA, AND A. POSTLEWAITE (1993): "Belief-based refinements in signalling games," *Journal of Economic Theory*, 60, 241–276.

MCLENNAN, A. (1985): "Justifiable beliefs in sequential equilibrium," *Econometrica*, 53, 889–904.

MOULIN, H. (1979): "Dominance solvable voting schemes," *Econometrica*, 1337–1351.

MYERSON, R. B. (1978): "Refinements of the Nash equilibrium concept," *International Journal of Game Theory*, 7, 73–80.

NOLDEKE, G. AND E. VAN DAMME (1990): "Signalling in a dynamic labour market," *The Review of Economic Studies*, 57, 1–23.

OKADA, A. (1981): "On stability of perfect equilibrium points," *International Journal of Game Theory*, 10, 67–73.

PEARCE, D. G. (1984): "Rationalizable strategic behavior and the problem of perfection," *Econometrica*, 1029–1050.

RENY, P. J. (1992): "Backward induction, normal form perfection and explicable equilibria," *Econometrica*, 627–649.

SELTEN, R. (1965): "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit: Teil I: Bestimmung des dynamischen Preisgleichgewichts," *Zeitschrift für die gesamte Staatswissenschaft / Journal of Institutional and Theoretical Economics*, 121, 301–324.

——— (1975): "Reexamination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory*, 4, 25–55.

SPENCE, M. (1973): "Job market signaling," *The Quarterly Journal of Economics*, 87, 355–374.

SWINKELS, J. M. (1999): "Education signalling with preemptive offers," *The Review of Economic Studies*, 66, 949–970.