

The Robustness of Sequential Implementation under Information Perturbations*

Ian Ball[†] Drew Fudenberg[‡] Stephen Morris[§]

24 September 2026

Abstract

Bayesian monotonicity is a necessary condition for full implementation in Bayes–Nash equilibrium. Under the sequential equilibrium refinement, however, multi-stage mechanisms can expand the scope for implementation. We show that this additional implementation power is fragile. Whenever a multi-stage mechanism sequentially implements a social choice correspondence that is not Bayesian monotone, we construct arbitrarily small “information perturbations” and show that there are sequential equilibria under these perturbations that yield inadmissible decisions with probability bounded away from zero. Under these perturbations, players’ types are not independent conditional on the state and some player is sometimes uncertain of their payoff type. We show by example that the general result fails if perturbations are required to satisfy conditional independence or “known payoff types.”

*We thank Pierpalo Battigalli for helpful comments. For financial support, we thank the National Science Foundation grant SES-2417162 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC-2047/2 – 390685813.

[†]Department of Economics, MIT, ianball@mit.edu.

[‡]Department of Economics, MIT, drewf@mit.edu.

[§]Department of Economics, MIT, semorris@mit.edu.

1 Introduction

The incomplete-information implementation problem asks which social choice correspondences can be fully implemented in pure Bayes–Nash equilibrium when players have privately observed payoff-relevant characteristics, which we call *payoff types*. Multi-stage mechanisms and the sequential equilibrium refinement can expand the scope for implementation relative to static mechanisms and Bayes–Nash equilibrium. We show that this additional implementation power is not robust to arbitrarily small “information perturbations” that change the information available to players. In these perturbations, players observe private signals rather than directly observing their payoff types. The signals can be arbitrarily informative about the payoff types, and the distribution of payoff types is unchanged. We also give an example in which the additional implementation power of sequential equilibrium is robust to all sufficiently small perturbations that satisfy an additional property—either “known payoff types” or conditional independence. (These properties are described below and formally defined in Section 6.)

We consider a social choice environment with a common prior. Each player knows their own payoff type, but they do not know the payoff types of their opponents. We allow for correlated payoff types and interdependent values. Suppose that a multi-stage mechanism sequentially implements a social choice correspondence that is not Bayesian monotone, and so cannot be fully implemented in pure Bayes–Nash equilibrium (Jackson, 1991). Our main result, Theorem 1, constructs arbitrarily small information perturbations and pure sequential equilibria under these perturbations that induce decisions outside the social choice correspondence with probability bounded away from zero. This is an incomplete-information analogue of the complete-information result in Aghion, Fudenberg, Holden, Kunitomo, and Tercieux (2012). The incomplete-information setting introduces new challenges that require a substantially different global construction with richer signal spaces and nondegenerate off-path beliefs.

The intuition for our result is as follows. Consider a multi-stage mechanism that fully implements a social choice correspondence F in pure sequential equilibrium. If F is not Bayesian monotone, the mechanism has a “good” sequential equilibrium that induces a social choice function in F and a “bad” Bayes–Nash equilibrium (BNE) that induces a social choice function outside F . By the definition of full implementation,

this bad BNE cannot be supported as a sequential equilibrium of the mechanism. We construct a family of perturbed information structures that converge to the baseline environment (in a sense that we formalize). In each perturbed information structure, each player observes a private signal rather than their true payoff type. We construct a pure sequential equilibrium in which the players coordinate on the bad BNE with high probability; otherwise, they coordinate on the good sequential equilibrium. From their signal realization, each player can learn their own payoff type with high confidence, but at least one player does not always learn their payoff type perfectly. If a player observes a deviation by one of their opponents, they infer that their opponent observed a particular revealing signal realization. They conclude that the players are actually coordinating on the moves specified by the good sequential equilibrium under the true payoff-type profile. Given this inference, sequential optimality follows.

In our constructed information structures, players do not always know their own payoff types, and players' signals are correlated conditional on the true payoff-type profile. We next show by example that these properties are necessary for the general result. To do this formally, we consider two restricted classes of information structures. An information structure satisfies *known payoff types (KPT)* if each player always knows their own payoff type. An information structure satisfies *conditional independence (CI)* if the players' types are statistically independent, conditional on the true payoff-type profile. For each of these restricted perturbation classes, we construct an example of a social choice function that is not Bayesian monotone but can be sequentially implemented by a multi-stage mechanism in a way that is robust to all sufficiently small information perturbations in the restricted class. That is, the probability of an inadmissible outcome vanishes with the size of the perturbation under every pure sequential equilibrium. Formally, these examples show that the general non-robustness conclusion in Theorem 1 no longer holds if the information perturbations are restricted to either the KPT or CI classes.

The rest of the paper is organized as follows. Section 1.1 discusses related literature. Section 2 introduces a simple example that illustrates the main construction. Section 3 introduces the general social choice environment and standard implementation notions. Section 4 defines information perturbations. Section 5 presents the main non-robustness result. In Section 6, we show that the general non-robustness conclusion fails if perturbations are required to satisfy either KPT or CI. Section 7 concludes. Appendix A contains the proofs.

1.1 Related literature

As illustrated in Table 1, our paper contributes to a line of work on implementation under both complete and incomplete information. With complete information, Maskin monotonicity is a necessary condition for Nash implementation (Maskin, 1999).¹ Subsequently, Moore and Repullo (1988) show that multi-stage mechanisms can implement many social choice functions that violate Maskin monotonicity. Aghion et al. (2012) show, however, that such implementation is not robust to small information perturbations.² The perturbations they consider are similar in spirit to those in Fudenberg, Kreps, and Levine (1988).

With incomplete information, there has been a parallel development. Bayesian monotonicity is a necessary condition for full implementation in pure Bayes–Nash equilibrium (Jackson, 1991). Subsequently, it was shown that under extensive-form refinements, multi-stage mechanisms can fully implement some social choice functions that violate Bayesian monotonicity. Brusco (1995) introduces necessary and sufficient conditions for full implementation in perfect Bayesian equilibrium. Bergin and Sen (1998) and Baliga (1999) provide sufficient conditions for full implementation in pure sequential equilibrium.³ Our paper shows that this additional implementation power is not robust to small information perturbations.

	Complete information	Incomplete information
Static implementation	Maskin (1999)	Jackson (1991)
Multi-stage implementation	Moore and Repullo (1988)	Baliga (1999), Bergin and Sen (1998), Brusco (1995)
Fragility under information perturbations	Aghion et al. (2012)	This paper

Table 1. Robustness of implementation under complete and incomplete information

Under incomplete information, there is an important new consideration of whether

¹Maskin (1999) also shows that with three or more players, under a no-veto power condition, Maskin monotonicity is sufficient for Nash implementation.

²In a different part of their paper, they also show that the truthful equilibria in the specific mechanisms in Moore and Repullo (1988) do not remain equilibria under small information perturbations.

³In a quasilinear setting, Duggan (1998) shows that incentive compatibility is sufficient for extensive-form implementation. Brusco (2006) characterizes full implementation in perfect Bayesian equilibrium by a class of two-stage mechanisms with a public signal after the first stage.

information perturbations are required to satisfy the known-payoff-type (KPT) property. In a complete-information setting in which the state is a vector of payoff types, [Eccles and Wegner \(2016\)](#) show that sequential implementation in two-stage mechanisms is robust to KPT perturbations, but sequential implementation in three-stage mechanisms may not be. Also with complete information, [Chen, Holden, Kunimoto, Sun, and Wilkenning \(2023\)](#) provide a two-stage mechanism that implements every social choice function in a way that is robust to “private-value perturbations,” which are slightly more general than KPT perturbations. We do not provide a general positive robustness result for KPT perturbations, but we show that our general non-robustness result does not hold under a restriction to KPT perturbations.

In the complete-information setting, [Palfrey and Srivastava \(1991\)](#) show that the static refinement of undominated Nash equilibrium substantially expands the implementation possibilities. [Chung and Ely \(2003\)](#) show that this additional implementation power is not robust to information perturbations.

Our paper also relates to the literature on robustness of equilibrium refinements under elaborations of a benchmark game following [Fudenberg et al. \(1988\)](#). In complete-information normal-form games, [Dekel and Fudenberg \(1990\)](#) characterize $S^\infty W$ —one round of deletion of weakly dominated strategies followed by iterated deletion of strongly dominated strategies—using elaborations with vanishing payoff uncertainty, while [Börger \(1994\)](#) gives a related epistemic characterization in terms of approximate common certainty of admissibility. [Frick and Romm \(2015\)](#) show that when extending these results to incomplete information, the conclusion depends on whether the baseline types retain exactly their original beliefs about payoff-relevant states or are instead allowed arbitrarily small perturbations of those beliefs. In the latter case, the set of predicted behavior is generally strictly larger. Our result identifies a related sensitivity for sequential implementation: preserving the ex-ante distribution of payoff types does not by itself preserve the baseline sequential-equilibrium prediction, because allowing players to be uncertain about their own payoff-type components can change the off-path inferences that support sequential behavior.

Another strand of the literature focuses on refinements of rationalizability rather than Nash equilibrium. By the structure theorem in [Weinstein and Yildiz \(2007\)](#), under a richness condition, refinements of rationalizability cannot yield robust predictions sharper than the predictions of rationalizability itself. Under incomplete information, [Müller \(2016\)](#) shows that many social choice functions that are not

virtually implementable under rationalizability can nevertheless be virtually implemented under strong rationalizability (Battigalli and Siniscalchi, 1999, 2003), which imposes a form of forward induction. Battigalli and Siniscalchi (2007) show that strong rationalizability is not robust to some forms of belief perturbations, while Battigalli and Catonini (2026) show that full implementation with respect to strong rationalizability is robust to the planner lacking knowledge of the type space that represented the agents' possible belief hierarchies. Piermont and Zuazo-Garin (2026) and Wang (2026) analyze the robustness of strong rationalizability to particular classes of perturbations.

2 Motivating example

We begin with a simple example that illustrates the additional power of implementation in pure sequential equilibrium and the non-robustness of such implementation. There are two players, denoted $i = 1, 2$. Each player i observes their payoff type $\theta_i \in \Theta_i = \{0, 1\}$, which is drawn uniformly and independently. The set of decisions is

$$A = \left\{ x^{\theta', c, d} : \theta' \in \Theta \text{ and } c, d \in \{0, 1\}^2 \right\}.$$

Utility functions will be specified below. Consider the following three-stage mechanism. The structure of reporting, challenging, and defending is inspired by an example in Moore and Repullo (1988).

1. Each player i reports $\theta'_i \in \Theta_i$. Play proceeds to stage 2.
2. After observing the reports in stage 1, each player i chooses whether to challenge ($c_i = 1$) or not ($c_i = 0$). If neither player challenges, the game ends and the decision is $x^{\theta', 0, 0}$. Otherwise, play proceeds to stage 3.
3. After observing the challenge decisions in stage 2, each player who was challenged chooses whether to defend ($d_i = 1$) or not ($d_i = 0$). The game ends and the decision is $x^{\theta', c, d}$.⁴

The players have private values:

$$u_i(x^{\theta', c, d}, \theta_i) = c_{-i} \cdot v_i(\theta'_i, d_i, \theta_i) + c_i \cdot w_i(d_{-i}), \quad (1)$$

⁴We adopt the convention that $d_i = 0$ whenever $c_{-i} = 0$.

where

$$v_i(\theta'_i, d_i, \theta_i) = \begin{cases} -1 + d_i & \text{if } \theta_i = \theta'_i, \\ -3 - d_i & \text{if } \theta_i \neq \theta'_i, \end{cases}$$

and

$$w_i(d_{-i}) = 2 - 3d_{-i}.$$

Player i 's payoff in (1) has two components: v_i captures the effect of being challenged, and w_i captures the effect of challenging. Upon being challenged, it is strictly optimal for a player to defend if they reported truthfully and not to defend if they misreported. Challenging their opponent changes the player's utility by 2 if the opponent does not defend but by -1 if the opponent defends.

2.1 Equilibria

First we consider pure sequential equilibria of the mechanism. We claim that the unique sequential equilibrium strategy profile is as follows. In stage 1, each player reports truthfully. In stage 2, neither player challenges. In stage 3, each player who was challenged in stage 2 defends if and only if they reported truthfully in stage 1. To verify this claim, we work backwards. The sequentially optimal behavior in stage 3 is pinned down, no matter what each player believes about their opponent's payoff type. In stage 2, Bayes' rule implies that after each report profile, each player is certain that their opponent reported truthfully and hence will defend against a challenge, so challenging is strictly suboptimal. Finally, in stage 1, players are indifferent over all report profiles, so in particular, truthtelling is optimal. Consistency can be established from a suitable sequence of totally mixed strategies.

It remains to show that in every sequential equilibrium, each type of each player reports truthfully in stage 1; then sequentially optimal play is pinned down in stages 2 and 3, by the argument above. Suppose for a contradiction that there is a sequential equilibrium in which some type θ_i of some player i reports $\theta'_i = 1 - \theta_i$. Upon seeing this report θ'_i , player $-i$ will challenge in stage 2 because they infer that with probability at least $1/2$ player i misreported and hence will not defend (this probability is $1/2$ if type $1 - \theta_i$ reports truthfully and 1 otherwise). Therefore, type θ_i 's payoff from reporting $\theta'_i = 1 - \theta_i$ is at most $-3 + 2 = -1$. But type θ_i can guarantee a payoff of 0 by instead reporting truthfully, not challenging, and then defending if challenged. This is a contradiction.

On the other hand, the mechanism admits additional pure Bayes–Nash equilibria. Here is one example. In stage 1, each player misreports their type. In stage 2, neither player challenges. In stage 3, each player who was challenged in stage 2 defends. Defending in stage 3 after having misreported is not sequentially optimal (no matter the player’s beliefs about their opponent’s payoff type). But anticipating this defense, it is optimal for each player not to challenge in stage 2. And anticipating no challenge, each player is indifferent over all reporting strategies in stage 1.

Consider the social choice function $f: \Theta \rightarrow A$ defined by $f(\theta) = x^\theta$. Our equilibrium analysis shows that (a) every pure sequential equilibrium of the mechanism induces f , but (b) there exists a pure Bayes–Nash equilibrium that induces a social choice function different from f . In the terminology of implementation theory, this mechanism fully implements f in pure sequential equilibrium, but it does not fully implement f in pure Bayes–Nash equilibrium. In fact, no mechanism fully implements f in pure Bayes–Nash equilibrium; this follows from [Jackson \(1991\)](#) because f violates the condition of Bayesian monotonicity (as we verify below in the proof of [Proposition 1](#)). Thus, moving from Bayes–Nash to sequential equilibrium makes more social choice functions implementable. We show next, however, that these undesirable Bayes–Nash outcomes can be approximately recovered as sequential equilibria if the information structure is slightly perturbed.

2.2 Perturbations

To perturb this setting, we suppose that each player does not directly observe their payoff type but instead observes a private signal, which we call their *type*. For each player i , define the space T_i of types by

$$T_i = \left\{ (t_i^{\theta'_i}, z_i) : \theta'_i \in \Theta_i \text{ and } z_i \in \{0, 1\} \right\}.$$

Let $T = T_1 \times T_2$. Fix $\varepsilon \in (0, 1)$. Suppose that the payoff-type profile θ and the type profile t are drawn from the following distribution over $\Theta \times T$. The distribution over payoff-type profiles is uniform, as in the unperturbed setting. For each payoff-type

profile $\theta \in \Theta$, the conditional distribution over type profiles $t \in T$ is given by⁵

$\pi^\varepsilon(t \theta)$	$(t_2^{\theta_2}, 0)$	$(t_2^{\theta_2}, 1)$	$(t_2^{1-\theta_2}, 0)$
$(t_1^{\theta_1}, 0)$	$1 - \varepsilon$	0	0
$(t_1^{\theta_1}, 1)$	0	0	$\varepsilon/2$
$(t_1^{1-\theta_1}, 0)$	0	$\varepsilon/2$	0

This conditional distribution can be interpreted as follows. Given the payoff-type profile $\theta = (\theta_1, \theta_2)$, the “default” is for each player i to receive signal $(t_i^{\theta_i}, 0)$. However, with probability ε , the default signal profile is changed as follows. A player $i \in \{1, 2\}$ is selected uniformly. The first component of player i ’s signal is flipped from $t_i^{\theta_i}$ to $t_i^{1-\theta_i}$, and the second component of player $-i$ ’s signal is flipped from $z_{-i} = 0$ to $z_{-i} = 1$. Under this interpretation, each player observes only the final, possibly modified, signal. Note that player i is uncertain of their own payoff type after observing $z_i = 0$, and that the players’ signals are correlated conditional on the payoff-type profile.

Under this perturbed information structure, we claim that the following strategy profile can be supported as a sequential equilibrium. If player i receives signal $(t_i^{\theta'_i}, z_i)$, then they play as follows. In stage 1, they report $1 - \theta'_i$ if $z_i = 0$ and θ'_i if $z_i = 1$. In stage 2, they do not challenge. In stage 3, if they were challenged in stage 2, they defend if and only if they followed their equilibrium strategy in stage 1. Below, we construct stage-3 beliefs that make the stage-3 strategy sequentially optimal. In stage 2, every report is on path, so each player anticipates that their opponent will defend if challenged. Hence, it is optimal not to challenge. And anticipating no challenge, each player is indifferent over all reporting strategies in stage 1.

To construct consistent beliefs in stage 3 so that, after being challenged, player i is certain that their true payoff type coincides with the report specified by their equilibrium strategy in stage 1, we consider two cases.

1. If player i ’s signal is $(t_i^{\theta'_i}, 0)$ then after being challenged, player i infers that their opponent received a signal with $z_{-i} = 1$; this belief can be justified by a sequence of trembles in which their opponent’s probability of challenging after observing $z_{-i} = 0$ becomes arbitrarily small relative to the probability of challenging after observing $z_{-i} = 1$. Player i concludes that their own signal was modified from

⁵For each player i , we omit the realization $(t_i^{1-\theta_i}, 1)$ since its conditional probability is 0.

the default, so their true payoff type is $1 - \theta'_i$, which is the stage-1 report specified by their equilibrium strategy.

2. If player i 's signal is $(t_i^{\theta'_i}, 1)$, they are certain that their own payoff type is θ'_i , which is the stage-1 report specified by their equilibrium strategy.

Under this sequential equilibrium, if the true payoff-type profile is θ , then with probability $1 - \varepsilon$, each player receives the “default” signal and hence the induced decision is $x^{1-\theta_1, 1-\theta_2}$, which is different from $f(\theta)$. Taking ε arbitrarily small, this sequential equilibrium recovers the undesirable BNE outcome with arbitrarily high probability.

Our construction here is tailored to the details of the two-player, binary-type social choice environment and the particular three-stage mechanism. In our main analysis, we provide a general construction that applies to any social choice environment (even with correlated payoff types and interdependent values), any finite multi-stage mechanism with observed actions, and any social choice correspondence that violates Bayesian monotonicity.

3 Implementation setting

Environment Consider a Bayesian social choice environment $(\Theta, \pi, A, (u_i)_{i=1}^n)$, where $\Theta = \prod_{i=1}^n \Theta_i$ is a finite space of payoff-type profiles, $\pi \in \Delta(\Theta)$ is a common prior with full support, A is a finite set of collective decisions, and $u_i: A \times \Theta \rightarrow \mathbf{R}$ is player i 's utility function. We allow for interdependent values: player i 's utility can depend on others' payoff types. We also allow players' payoff types to be correlated under the prior π .

Social choices A *social choice function* (SCF) is a function $f: \Theta \rightarrow A$, which assigns to each profile θ a decision $f(\theta)$. A *social choice correspondence* (SCC) is a nonempty-valued correspondence $F: \Theta \rightrightarrows A$, which assigns to each profile θ a nonempty set $F(\theta)$ of decisions.

For each player i and payoff type $\theta_i \in \Theta_i$, define the induced interim preference relation $\succsim_{\theta_i}$ on the space of social choice functions as follows:

$$f \succsim_{\theta_i} f' \iff \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f(\theta), \theta) \pi_{-i}(\theta_{-i} | \theta_i) \geq \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f'(\theta), \theta) \pi_{-i}(\theta_{-i} | \theta_i).$$

Define the strict relation $\succ_{\theta_i}$ analogously with a strict inequality in place of the weak inequality.

BNE implementation We largely follow the framework of [Jackson \(1991\)](#). A mechanism is a pair (M, g) consisting of a set $M = \prod_{i=1}^n M_i$ of message profiles and an outcome function $g: M \rightarrow A$. Given a mechanism (M, g) , a strategy for player i is a map $\sigma_i: \Theta_i \rightarrow M_i$. We view a strategy profile $\sigma = (\sigma_i)_{i=1}^n$ as a map from Θ to M via $\sigma(\theta) = (\sigma_i(\theta_i))_{i=1}^n$. A strategy profile σ is a *Bayes–Nash equilibrium (BNE)* if for each player i , type θ_i , and strategy σ'_i , we have $g \circ \sigma \succeq_{\theta_i} g \circ (\sigma'_i, \sigma_{-i})$. Following [Jackson \(1991\)](#), we define a BNE to be a *pure* strategy profile.

Definition 1 (BNE implementation). A mechanism (M, g) *BNE implements* an SCC F if the following hold:

1. For each BNE σ of (M, g) , the map $g \circ \sigma$ is a selection from F ;
2. For each selection f from F , there exists a BNE σ of (M, g) such that $g \circ \sigma = f$.

An SCC F is *BNE-implementable* if there exists a mechanism (M, g) that BNE implements F .

[Jackson \(1991\)](#) introduces a necessary condition for BNE implementation, termed Bayesian monotonicity. To state the definition, we need a few preliminaries. A *deception* for player i is a function $\alpha_i: \Theta_i \rightarrow \Theta_i$. We view a deception profile $\alpha = (\alpha_i)_{i=1}^n$ as a map from Θ to Θ via $\alpha(\theta) = (\alpha_i(\theta_i))_{i=1}^n$. Given an SCF f , define for any player i and type $\bar{\theta}_i$ the SCF $f_{i, \bar{\theta}_i}$ by $f_{i, \bar{\theta}_i}(\theta_i, \theta_{-i}) = f(\bar{\theta}_i, \theta_{-i})$.

Definition 2 (Bayesian monotonicity). A social choice correspondence F is *Bayesian monotone* if for each selection f from F , and each deception profile α such that $f \circ \alpha$ is not a selection from F , there exists some player i , payoff type $\bar{\theta}_i \in \Theta_i$, and social choice function f' such that $f' \circ \alpha \succ_{\bar{\theta}_i} f \circ \alpha$ and, for each $\theta_i \in \Theta_i$, we have $f \succeq_{\theta_i} f'_{i, \alpha_i(\bar{\theta}_i)}$.

From the necessity proof of Theorem 1 in [Jackson \(1991\)](#), Bayesian monotonicity is necessary for BNE implementability. The interpretation is that, in order to ensure that it is not an equilibrium for the players to follow the deception profile α , there must be some SCF $f'_{i, \alpha_i(\bar{\theta}_i)}$ that can be used in place of f to reward type $\bar{\theta}_i$ for “blowing the whistle” on their opponents for following α . The conditions require that type $\bar{\theta}_i$ strictly prefers to blow the whistle when the players follow α , but that no type of player i strictly prefers to blow the whistle if there is no deception.

Sequential implementation Following [Aghion et al. \(2012\)](#), we consider sequential implementation using finite multi-stage mechanisms with observed actions. Formally, such a mechanism is a triple $(\mathcal{H}, \mathcal{H}', g)$, where $\mathcal{H}$ is the finite set of nonterminal histories, $\mathcal{H}'$ is the finite set of terminal histories, and $g: \mathcal{H}' \rightarrow A$ is the outcome function. The set $\mathcal{H} \cup \mathcal{H}'$ of all histories contains the initial history $\emptyset$ and is closed under the predecessor-of relation. For each history $h \in \mathcal{H} \cup \mathcal{H}'$, let

$$M(h) = \{m : (h, m) \in \mathcal{H} \cup \mathcal{H}' \text{ and } h \text{ is the direct predecessor of } (h, m)\}.$$

For each $h \in \mathcal{H}'$, we have $M(h) = \emptyset$. For each $h \in \mathcal{H}$, we have $M(h) = \prod_i M_i(h)$, where each set $M_i(h)$ is nonempty. At each history $h \in \mathcal{H}$, the players who are active at h move simultaneously, where player i is active at h if and only if $|M_i(h)| > 1$. All players observe these moves before choosing how to move in the next stage. Write $M_i = \prod_{h \in \mathcal{H}} M_i(h)$ and $M = \prod_i M_i$. Given a profile m in M , denote by $g(m)$ the outcome at the terminal history induced by m . Similarly, for each history $h \in \mathcal{H}$, write $g(m; h)$ for the outcome induced by the profile m after starting at the history h . Note that $g(m; h)$ depends only on the moves prescribed by m at history h and its successors.

Given a multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$, a strategy for player i is a map $\sigma_i: \Theta_i \rightarrow M_i$. View a strategy profile $\sigma = (\sigma_i)_{i=1}^n$ as a map from Θ to M via $\sigma(\theta) = (\sigma_i(\theta_i))_{i=1}^n$. A strategy σ_i can be equivalently viewed as a function on $\Theta_i \times \mathcal{H}$ satisfying $\sigma_i(\theta_i, h) \in M_i(h)$ for all $\theta_i \in \Theta_i$ and $h \in \mathcal{H}$. A belief system for player i is a map $\mu_i: \Theta_i \times \mathcal{H} \rightarrow \Delta(\Theta_{-i})$, which specifies player i 's belief about others' payoff types as a function of their own payoff type and the history. An assessment is a pair (σ, μ) consisting of a strategy profile σ and a belief-system profile μ .

Following [Maskin \(1999\)](#), [Jackson \(1991\)](#), and [Aghion et al. \(2012\)](#), we restrict to pure equilibrium strategy profiles. In order to define belief consistency, however, we must consider behavioral strategies. A behavioral strategy $\tilde{\sigma}_i$ for player i assigns to each pair $(\theta_i, h) \in \Theta_i \times \mathcal{H}$ a move distribution $\tilde{\sigma}_i(\theta_i, h) \in \Delta(M_i(h))$. A behavioral strategy is *totally mixed* if $\text{supp } \tilde{\sigma}_i(\theta_i, h) = M_i(h)$ for all $\theta_i \in \Theta_i$ and $h \in \mathcal{H}$. Given any pure strategy σ_i , we denote by δ_{σ_i} the equivalent behavioral strategy satisfying $\delta_{\sigma_i}(\theta_i, h) = \delta_{\sigma_i(\theta_i, h)}$.

Definition 3 (Sequential equilibrium). An assessment (σ, μ) is a *sequential equilibrium* if the following hold:

1. *Sequential optimality*: For each player i , payoff type $\theta_i \in \Theta_i$, and history $h \in \mathcal{H}$,

$$\sigma_i(\theta_i) \in \operatorname{argmax}_{m_i \in M_i} \sum_{\theta_{-i} \in \Theta_{-i}} u_i(g(m_i, \sigma_{-i}(\theta_{-i}); h), \theta) \mu_i(\theta_i, h)[\theta_{-i}].$$

2. *Consistency*: There exists a sequence, (σ^k, μ^k) , of totally mixed behavioral strategy profiles and belief-system profiles such that for each player i :
 - (a) for each k , the belief system μ_i^k is derived from σ_{-i}^k via Bayes' rule;
 - (b) the sequence (σ_i^k, μ_i^k) converges to $(\delta_{\sigma_i}, \mu_i)$ pointwise on $\Theta_i \times \mathcal{H}$.

Definition 4 (Sequential implementation). A multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$ *sequentially implements* a social choice correspondence F if the following hold.

1. For each sequential equilibrium (σ, μ) of $(\mathcal{H}, \mathcal{H}', g)$, the map $g \circ \sigma$ is a selection from F .
2. For each selection f from F , there exists a sequential equilibrium (σ, μ) of $(\mathcal{H}, \mathcal{H}', g)$ such that $g \circ \sigma = f$.

An SCC F is *sequentially implementable* if there exists a multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$ that sequentially implements F .

4 Information perturbations

We now define information perturbations of the social choice environment $(\Theta, \pi, A, (u_i)_{i=1}^n)$. Rather than perfectly observing their true payoff type, each player observes a private signal about it, which determines their type. Following [Fudenberg et al. \(1988\)](#), we associate each payoff type of each player with a distinguished type that will play the role of the “default” signal in the example from Section 2. Formally, an *information structure* is a triple $(T, t^*, \hat{\pi})$ specifying a finite set $T = \prod_i T_i$ of type profiles, an injective map $t_i^*: \Theta_i \rightarrow T_i$ for each player i , and a prior $\hat{\pi} \in \Delta(\Theta \times T)$ such that for each player i , the marginal $\operatorname{marg}_{T_i} \hat{\pi}$ has full support on T_i . This definition implies that $|T_i| \geq |\Theta_i|$ for each player i .

A player's type determines their beliefs about the profile of payoff types, but the types do not directly affect payoffs. We write $t^*(\theta) = (t_i^*(\theta_i))_{i=1}^n$ for the distinguished type profile associated to the payoff-type profile θ . Define the distribution $\pi \otimes t^* \in$

$\Delta(\Theta \times T)$ by $(\pi \otimes t^*)[\theta, t^*(\theta)] = \pi(\theta)$. Thus, the original setting can be represented by any information structure $(T, t^*, \hat{\pi})$ satisfying $\hat{\pi} = \pi \otimes t^*$.⁶

Given an information structure $(T, t^*, \hat{\pi})$ and a multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$, a strategy for player i is a map $\sigma_i: T_i \rightarrow M_i$. A strategy σ_i can equivalently be viewed as a function on $T_i \times \mathcal{H}$ satisfying $\sigma_i(t_i, h) \in M_i(h)$ for all $t_i \in T_i$ and $h \in \mathcal{H}$. A belief system for player i is a map $\mu_i: T_i \times \mathcal{H} \rightarrow \Delta(\Theta \times T_{-i})$, which specifies player i 's belief about the payoff-type profile and others' types as a function of their own type and the history. An assessment (σ, μ) is a sequential equilibrium if it is sequentially optimal and consistent, where these properties are defined as in Definition 3, mutatis mutandis.

We will consider parametric families of information structures, denoted by $(T, t^*, \pi^\varepsilon)_{\varepsilon \in (0,1)}$. Note that T and t^* do not vary with ε . Let $\|\cdot\|_{\text{TV}}$ denote the total variation norm. That is, for any $\pi_1, \pi_2 \in \Delta(\Theta \times T)$, let $\|\pi_1 - \pi_2\|_{\text{TV}} = \sup_E |\pi_1(E) - \pi_2(E)|$, where the supremum is over all subsets E of $\Theta \times T$.

Definition 5 (Convergence of information structures). A family of information structures, $(T, t^*, \pi^\varepsilon)_{\varepsilon \in (0,1)}$, converges to π if

- (i) $\text{marg}_\Theta \pi^\varepsilon = \pi$ for all ε ;
- (ii) $\lim_{\varepsilon \rightarrow 0} \|\pi^\varepsilon - \pi \otimes t^*\|_{\text{TV}} = 0$.

Part (ii) implies that $\lim_{\varepsilon \rightarrow 0} \|\text{marg}_\Theta \pi^\varepsilon - \pi\|_{\text{TV}} = 0$. We impose the stronger requirement (i) because it strengthens our main non-robustness result, Theorem 1.

To interpret this notion of convergence, consider the event E^* that each player's type is the distinguished type for their true payoff type. Formally, $E^* = \{(\theta, t^*(\theta)) : \theta \in \Theta\}$. Note that $(\pi \otimes t^*)(E^*) = 1$. Suppose that a family $(T, t^*, \pi^\varepsilon)_{\varepsilon \in (0,1)}$ converges to π . Then there exists a function $p: (0, 1) \rightarrow [0, 1]$ satisfying $\lim_{\varepsilon \rightarrow 0} p(\varepsilon) = 1$ such that for each $\varepsilon \in (0, 1)$, the event E^* has ex-ante probability at least $p(\varepsilon)$ and E^* is an evident $p(\varepsilon)$ -belief in the sense of Monderer and Samet (1989, p. 176).⁷

Remark 1 (Fixed payoffs versus payoff perturbations). In the perturbations we consider, the utility functions u_i are fixed, and signals affect behavior only through players' beliefs. A broader class of perturbations would allow the utility functions to depend on players' types and require only that the utility functions associated

⁶In this case, the full marginal support condition implies that $T_i = t_i^*(\Theta_i)$ for all i .

⁷Indeed, let $p(\varepsilon) = \min\{\pi^\varepsilon(E^*), \min_{i, \theta_i} \pi^\varepsilon[E^* | t_i^*(\theta_i)]\}$.

with the distinguished types converge to the baseline utility functions. As in the complete-information case studied by [Fudenberg et al. \(1988\)](#), using this broader class of perturbations can enlarge the set of baseline equilibrium behaviors that can arise as limits of sequential equilibria under perturbations: revealing types may provide direct incentives for an action through their different payoffs, rather than only changing the beliefs induced by their presence. Establishing fragility in the more restrictive fixed-payoff case strengthens our result.

5 Fragility of sequential implementation

We now show that the expanded implementation power from sequential equilibrium relative to Bayes–Nash equilibrium is not robust to small information perturbations.

Theorem 1 (Necessity of Bayesian monotonicity)

Let F be a social choice correspondence that is not Bayesian monotone. If F is sequentially implemented (under π) by some multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$, then there exists a family of information structures, $(T, t^*, \pi^\varepsilon)_{\varepsilon \in (0,1)}$, converging to π and an associated family $(\sigma^\varepsilon, \mu^\varepsilon)_{\varepsilon \in (0,1)}$ of sequential equilibria of $(\mathcal{H}, \mathcal{H}', g)$ such that for some $\theta \in \Theta$ and all $\varepsilon \in (0,1)$, the decision $g(\sigma^\varepsilon(t^*(\theta)))$ is not in $F(\theta)$.

Remark 2 (Integer games). We require that $\mathcal{H} \cup \mathcal{H}'$ is finite, while the mechanisms in [Bergin and Sen \(1998\)](#) and [Baliga \(1999\)](#) ask players to select arbitrary integers at some histories. In the proof, we show that our conclusion still holds for any finite-horizon multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$ with countably many moves, provided that it admits a sequential best response for each player at each history, for all beliefs and opposing pure strategies. This property holds for the mechanisms in [Bergin and Sen \(1998\)](#) and [Baliga \(1999\)](#).

To interpret the conclusion of Theorem 1, recall from the definition of convergence that for each $\theta \in \Theta$, we have $\lim_{\varepsilon \rightarrow 0} \pi^\varepsilon[\theta, t^*(\theta)] = \pi(\theta)$. Thus, the limiting ex-ante probability that the induced decision is outside F is at least $\min_{\theta' \in \Theta} \pi(\theta')$, which is strictly positive.

Here is an outline of the proof. Let F be an SCC that violates Bayesian monotonicity, and suppose that the multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$ sequentially implements F under π . Then there exists a sequential equilibrium $(\hat{\sigma}, \hat{\mu})$ of $(\mathcal{H}, \mathcal{H}', g)$ such that

$g \circ \hat{\sigma}$ is a selection from F , and there exists a deception profile α such that $\hat{\sigma} \circ \alpha$ is a Bayes–Nash equilibrium of $(\mathcal{H}, \mathcal{H}', g)$ but $g \circ \hat{\sigma} \circ \alpha$ is not a selection from F .

For each $\varepsilon \in (0, 1)$, define an information structure and an associated strategy profile such that, for every $\theta \in \Theta$, conditional on the payoff-type profile being θ , the players coordinate on the profile $\hat{\sigma}(\alpha(\theta))$ with probability $1 - \varepsilon$ and on the profile $\hat{\sigma}(\theta)$ with probability ε . Each player’s signal provides them with enough information to coordinate play in this way, but at least one player is sometimes uncertain of their payoff type and hence about which of the two profiles the players are coordinating on. We define the belief system so that if any player observes a deviation by an opponent, they infer that the players are coordinating on $\hat{\sigma}(\theta)$ (though they may remain uncertain of the true payoff-type profile θ). We then derive sequential optimality and consistency from the corresponding properties of $(\hat{\sigma}, \hat{\mu})$ under π .

Theorem 1 is an incomplete-information analogue of the complete-information result of [Aghion et al. \(2012, Theorem 3\)](#). The two constructions have a common structure: The failure of Bayesian/Maskin monotonicity is used to construct a “bad” Bayes–Nash/Nash equilibrium. Then an information structure is defined in order to induce a mixture between the good and bad equilibria. To ensure sequential optimality, each player infers from a detectable opponent deviation that the players are coordinating on the good equilibrium. Despite these similarities, the incomplete-information setting requires a different and more intricate construction.

First, the state realization can no longer serve as a coordination device. In order to coordinate play on the good and bad profiles, we enrich each player’s signal space so that it is strictly larger than their payoff-type space. Each player’s signal indicates how they should play without necessarily revealing their own payoff type. Second, even after a player observes a deviation by an opponent, they may remain uncertain about their opponents’ payoff types. We construct a sequence of trembles that replicates the potentially nondegenerate off-path beliefs about opponents’ payoff types that guarantee sequential optimality under the “good” equilibrium. Third, our construction is global in the sense that it depends on the full deception profile α . By contrast, the failure of Maskin monotonicity yields a single “deception” for all players, which differs from the identity at a single state. Thus, [Aghion et al. \(2012\)](#) reason locally about a fixed pair of states θ' and θ'' ; for every other realization, the state becomes common knowledge.

With additional structure on the social choice environment, Bayesian monotonicity

and incentive compatibility are jointly sufficient for BNE implementation. Given social choice functions f and f' and a subset E of Θ , let $f_E f'$ denote the social choice function that coincides with f on E and with f' on $\Theta \setminus E$.

Definition 6 (Economic environment). The social choice environment $(\Theta, \pi, A, (u_i)_{i=1}^n)$ is *economic* if for each SCF f and each payoff-type profile $\theta \in \Theta$, there exist distinct players i and j and constant SCFs a and b such that for all subsets E of Θ containing θ , we have $a_E f \succ_{\theta_i} f$ and $b_E f \succ_{\theta_j} f$.

Definition 6 is a uniform non-satiation property, which usually holds in exchange economies. A social choice correspondence is incentive compatible if every selection from it is Bayesian incentive compatible in the standard sense. By Jackson (1991, Theorem 1, p. 467), if $n \geq 3$ and the environment is economic, then Bayesian monotonicity and incentive compatibility are jointly sufficient for BNE implementability.⁸ Incentive compatibility is also necessary for sequential implementability (and BNE implementability); otherwise, any selection that is not incentive compatible could never be induced by any BNE since some type would strictly prefer to mimic the strategy of another type. Thus, we obtain the following.

Corollary 1 (Fragility of additional implementation)

Suppose that $n \geq 3$ and the social choice environment is economic. Let F be a social choice correspondence that is not BNE implementable (under π). If F is sequentially implemented (under π) by some multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$, then the conclusion of Theorem 1 holds.

Remark 3 (Bayes–Nash implementation). To highlight the contrast between sequential implementation and Bayes–Nash implementation, note that a standard upper hemicontinuity property of BNE gives the following result. Suppose that a social choice correspondence F is implemented in BNE (under π) by some finite mechanism (M, g) . For any family of information structures, $(T, t^*, \pi^\varepsilon)_{\varepsilon \in (0,1)}$, converging to π and any associated family $(\sigma^\varepsilon)_{\varepsilon \in (0,1)}$ of pure BNE, there exists $\bar{\varepsilon} > 0$ such that $g(\sigma^\varepsilon(t^*(\theta))) \in F(\theta)$ for all $\varepsilon \in (0, \bar{\varepsilon})$ and all $\theta \in \Theta$.⁹

⁸Jackson (1991) requires an additional condition, termed closure, but this condition is already implied by our focus on social choice correspondences, and the associated set of selections, as opposed to general sets of social choice functions.

⁹Otherwise, since T and M are finite, there exists some payoff-type profile $\theta \in \Theta$, strategy profile $\sigma: T \rightarrow M$, and sequence (ε^k) converging to 0 such that for all k , we have $\sigma^{\varepsilon^k} = \sigma$ and $g(\sigma(t^*(\theta))) \notin F(\theta)$. Passing to the limit in the equilibrium conditions, it follows that the profile $\sigma'(\theta) = \sigma(t^*(\theta))$ is a BNE of (M, g) under π , contrary to BNE implementation.

6 More restrictive perturbations

To prove Theorem 1, we constructed information structures with two properties: (a) at least one player is sometimes uncertain of their own payoff type; and (b) conditional on the payoff-type profile, the players' types are not statistically independent. Here, we show that the general result fails if perturbations are required to satisfy “known payoff types” or conditional independence.

6.1 Known payoff types

An information structure $(T, t^*, \hat{\pi})$ has *known payoff types (KPT)* if (a) for each player i , there exists a finite set S_i such that $T_i \subseteq \Theta_i \times S_i$, and (b) $\hat{\pi}(\theta, (\theta'_i, s_i)_{i=1}^n) = 0$ for every $\theta \in \Theta$ and $(\theta'_i, s_i)_{i=1}^n \in T$ such that $\theta \neq (\theta'_i)_{i=1}^n$. Intuitively, each player i 's type $t_i = (\theta_i, s_i)$ reveals their true payoff type θ_i and a supplementary signal $s_i \in S_i$.

Remark 4 (KPT and knowledge of payoffs). The KPT condition should not be interpreted as requiring player i to know their entire realized payoff function. KPT requires that player i observe their own payoff-type component θ_i . Because we allow interdependent values, player i may nevertheless remain uncertain about the utility consequences of a decision: their utility $u_i(a, \theta)$ can depend on θ_{-i} . A stronger known-payoffs property would require that all payoff-type profiles that player i considers possible induce the same function $a \mapsto u_i(a, \theta)$.

Formally, we show that Theorem 1 fails if the perturbed information structures are required to satisfy KPT.

Proposition 1 (KPT counterexample)

There is a social choice environment $(\Theta, \pi, A, (u_i)_{i=1}^n)$ with independent private values, an SCF f that is not Bayesian monotone, and a mechanism $(\mathcal{H}, \mathcal{H}', g)$ that sequentially implements f under π such that the following holds: There exists $\bar{\eta} > 0$ such that for every KPT information structure $(T, t^*, \hat{\pi})$ satisfying $\|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}} < \bar{\eta}$, every sequential equilibrium (σ, μ) of $(\mathcal{H}, \mathcal{H}', g)$ satisfies $g(\sigma(t^*(\theta))) = f(\theta)$ for all $\theta \in \Theta$.

The final conclusion of Proposition 1 implies that

$$\hat{\pi} \{(\theta, t) \in \Theta \times T : g(\sigma(t)) \neq f(\theta)\} \leq \|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}},$$

so in the example, the error probability on the left side converges to 0 for any family of KPT information structures converging to π .

To prove Proposition 1, we return to the example from Section 2. Let $(\mathcal{H}, \mathcal{H}', g)$ denote the mechanism and f the associated SCF from Section 2. In the perturbations from that example, it is critical that each player is sometimes uncertain of their own payoff type because then, after observing an unexpected challenge by their opponent, the player infers that the opponent's signal reveals that the player's payoff type matches the report they submitted. If players know their own payoff types, this kind of inference is impossible.

Since Bergemann and Morris (2005), it has become standard to maintain the known-payoff-type property when varying the information structure. Our results identify one setting in which this restriction changes the conclusion.

6.2 Conditional independence

An information structure $(T, t^*, \hat{\pi})$ is *conditionally independent (CI)* if under $\hat{\pi}$, the types $t_1, \dots, t_n$ are mutually independent, conditional on θ . This independence property implies that $\hat{\pi}(t_{-i}|\theta, t_i) = \hat{\pi}(t_{-i}|\theta)$ whenever $\hat{\pi}(\theta, t_i) > 0$. This is a natural generalization of Fudenberg et al.'s (1988) independent-types condition to our setting. Next, we prove the analogue of Proposition 1 for CI perturbations in place of KPT perturbations.

Proposition 2 (CI counterexample)

There is a social choice environment $(\Theta, \pi, A, (u_i)_{i=1}^n)$ with independent private values, an SCF f that is not Bayesian monotone, and a mechanism $(\mathcal{H}, \mathcal{H}', g)$ that sequentially implements f under π such that the following holds: There exists $\bar{\eta} > 0$ such that for every CI information structure $(T, t^*, \hat{\pi})$ satisfying $\|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}} < \bar{\eta}$, every sequential equilibrium (σ, μ) of $(\mathcal{H}, \mathcal{H}', g)$ satisfies $g(\sigma(t^*(\theta))) = f(\theta)$ for all $\theta \in \Theta$.

To prove Proposition 2, we consider a variant of the example from Section 2. The mechanism still has the report–challenge–defend structure, but now only player 2 can challenge, and only after a particular report by player 1. If there is no challenge, the payoffs now depend on the reports, unlike the example from Section 2. Under CI information structures, players can be uncertain of their own payoff types. Thus, an unexpected challenge can cause a player to update their beliefs about their own payoff type. Conditional independence implies that a player's signal does not change the

conditional distribution of their opponent’s signal given the payoff-type profile. In the example, the “bad” BNE outcome cannot be sustained as a sequential equilibrium because the payoff-type beliefs that can justify defending after a challenge in turn constrain beliefs about the opponent’s signal in a way that makes a stage-1 reporting deviation profitable.

7 Discussion

Theorem 1 shows that the implementation power of sequential equilibrium is not robust to small changes in the players’ information. The construction preserves both the utility functions and the distribution of payoff-type profiles, but it allows players to become uncertain about their own payoff types. If perturbations are required to preserve each player’s knowledge of their own payoff type, then no general fragility result of the form in Theorem 1 is possible. In environments with interdependent values, this restriction should be understood as preserving knowledge of a player’s own payoff-type component, rather than as requiring knowledge of their realized payoff function.

The theorem also permits correlation across players’ signals. In our construction, the common latent variable coordinates the revealing signals and supports the required off-path inferences. The analogous fragility result does not hold if, conditional on the payoff-type profile, players’ types are required to be conditionally independent.

Our analysis of the restrictions provided by sequential equilibrium leaves open whether the restrictions of other equilibrium refinements such as undominated Nash equilibrium, strategic stability (Kohlberg and Mertens, 1986), or strict Nash equilibrium (which implies all these other refinements) are robust to the perturbations we considered. If so, it would be interesting to see if they are also robust to slightly expanded notions of nearby games, such as replacing the condition $\text{marg}_{\Theta} \pi^{\varepsilon} = \pi$ in Definition 5 with its limiting version $\lim_{\varepsilon \rightarrow 0} \text{marg}_{\Theta} \pi^{\varepsilon} = \pi$ as suggested by Frick and Romm (2015).

A Proofs

A.1 Proof of Theorem 1

Fix a social choice correspondence F that is not Bayesian monotone and a multi-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$ that sequentially implements F (under π).

The proof proceeds in three steps. First, we obtain a suitable profile of deceptions. Second, we use this deception profile to define the family of information structures. Third, we construct a sequential equilibrium under each information structure with the desired properties.

Obtaining the deception profile Since F is not Bayesian monotone, there exists a selection f from F and a profile $\alpha = (\alpha_i)_{i=1}^n$ of deceptions such that (i) $f \circ \alpha$ is not a selection from F and (ii) for each player i and type $\bar{\theta}_i$ and each $f': \Theta \rightarrow A$, if $f \succsim_{\theta_i} f'_{i, \alpha_i(\bar{\theta}_i)}$ for all types θ_i in Θ_i , then $f \circ \alpha \succsim_{\bar{\theta}_i} f' \circ \alpha$. Note that α_i is not the identity for at least one player i . Since f is a selection from F and $(\mathcal{H}, \mathcal{H}', g)$ sequentially implements F , there exists a sequential equilibrium $(\hat{\sigma}, \hat{\mu})$ such that $g \circ \hat{\sigma} = f$.

Lemma 1 (Deceptive BNE)

Under π , the profile $\hat{\sigma} \circ \alpha = (\hat{\sigma}_i \circ \alpha_i)_{i=1}^n$ is a Bayes-Nash equilibrium of $(\mathcal{H}, \mathcal{H}', g)$.

Proof. We follow the proof in Jackson (1991). Fix player i , type $\theta_i \in \Theta_i$, and a deviation $m_i \in M_i$. It suffices to show that type θ_i cannot profit by deviating to m_i . Define $\sigma'_i: \Theta_i \rightarrow M_i$ by $\sigma'_i(\theta'_i) = m_i$ for all θ'_i . Define $f': \Theta \rightarrow A$ by $f' = g \circ (\sigma'_i, \hat{\sigma}_{-i})$. By construction, $f = g \circ \hat{\sigma}$. Since $\hat{\sigma}$ is a BNE, we conclude that

$$f = g \circ \hat{\sigma} \succsim_{\theta'_i} g \circ (\sigma'_i, \hat{\sigma}_{-i}) = f' = f'_{i, \alpha_i(\theta_i)},$$

for all types θ'_i . Hence,

$$g \circ (\hat{\sigma} \circ \alpha) = f \circ \alpha \succsim_{\theta_i} f' \circ \alpha = g \circ (\sigma'_i, \hat{\sigma}_{-i} \circ \alpha_{-i}). \quad \square$$

Constructing the information structures For each player i , let

$$T_i = \{(t_i^{\theta_i}, 0) : \theta_i \in \Theta_i\} \cup \{(t_i^{\theta_i}, 1) : \theta_i \in \Theta_i\} \cup \{(t_i^{\theta_i}, 2) : \theta_i \in \Theta_i \setminus \alpha_i(\Theta_i)\}. \quad (2)$$

Let $T = \prod_{i=1}^n T_i$. Define $t_i^*: \Theta_i \rightarrow T_i$ by $t_i^*(\theta_i) = (t_i^{\theta_i}, 0)$.

For each player i , define $\bar{\alpha}_i: T_i \rightarrow \Theta_i$ by

$$\bar{\alpha}_i(t_i^{\theta'_i}, z_i) = \begin{cases} \alpha_i(\theta'_i) & \text{if } z_i = 0, \\ \theta'_i & \text{if } z_i \neq 0. \end{cases} \quad (3)$$

We construct a joint distribution over $\Theta \times T$ that depends on the realization of a latent Bernoulli variable Z with the following properties. If $Z = 0$, then $\bar{\alpha}_i(t_i) = \alpha_i(\theta_i)$. If $Z = 1$, then $\bar{\alpha}_i(t_i) = \theta_i$. And every type profile is consistent with $Z = 1$.

Now we turn to the formal construction. First, we introduce auxiliary kernels. For each $z \in \{0, 1\}$ and each player i , define $d_i^z: \Theta_i \rightarrow \Delta(T_i)$ by letting $d_i^z(\theta_i)$ equal the point mass on $(t_i^{\theta_i}, z)$. Define the product kernel $d^z: \Theta \rightarrow \Delta(T)$ by $d^z(\theta) = \otimes_i d_i^z(\theta_i)$. Next, define $\kappa_i: \Theta_i \rightarrow \Delta(T_i)$ as follows. For $\theta_i \in \alpha_i(\Theta_i)$, let

$$\kappa_i(t_i^{\theta'_i}, z_i | \theta_i) = \begin{cases} 1/2 & \text{if } (\theta'_i, z_i) = (\theta_i, 1), \\ \frac{1}{2|\alpha_i^{-1}(\theta_i)|} & \text{if } \theta'_i \in \alpha_i^{-1}(\theta_i) \text{ and } z_i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For $\theta_i \in \Theta_i \setminus \alpha_i(\Theta_i)$, let

$$\kappa_i(t_i^{\theta'_i}, z_i | \theta_i) = \begin{cases} 1/2 & \text{if } (\theta'_i, z_i) = (\theta_i, 1), \\ 1/2 & \text{if } (\theta'_i, z_i) = (\theta_i, 2), \\ 0 & \text{otherwise.} \end{cases}$$

Define the product kernel $\kappa: \Theta \rightarrow \Delta(T)$ by $\kappa(\theta) = \otimes_i \kappa_i(\theta_i)$.

For each $\varepsilon \in (0, 1)$, define $\pi^\varepsilon \in \Delta(\Theta \times T)$ by

$$\pi^\varepsilon = (1 - \varepsilon)(\pi \otimes d^0) + \varepsilon(\pi \otimes \kappa),$$

where $\otimes$ denotes the product of a measure and a kernel.¹⁰ We include $z_i = 2$ in order to keep the probability of $z_i = 1$ constant across payoff types. This is not essential, but it simplifies the construction of trembles below.

Remark 5 (Interpretation of information structure). Figure 1 shows the stochastic mapping from Z to T_i and then the deterministic mapping from $Z \times T_i$ to Θ_i . Here,

¹⁰That is, $(\pi \otimes \kappa)[\theta, t] = \pi(\theta)\kappa(t|\theta)$ for all $\theta \in \Theta$ and $t \in T$.

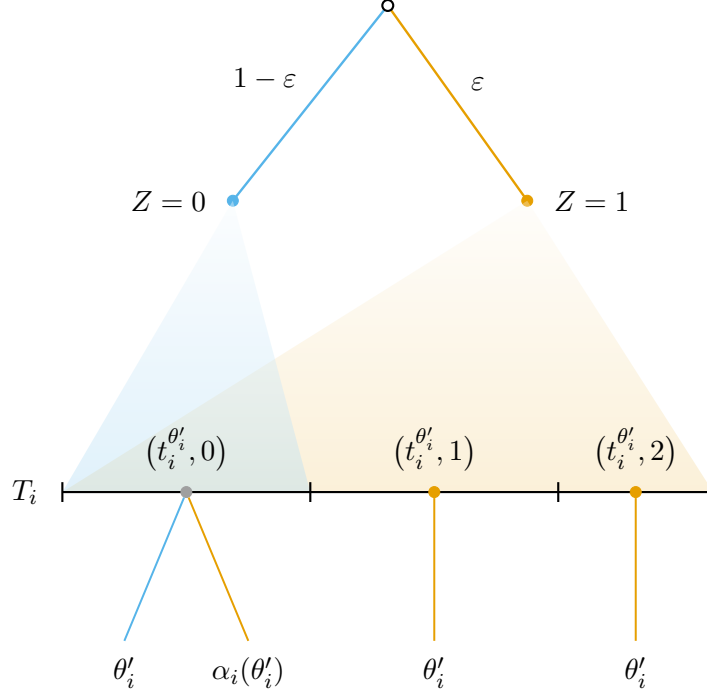


Figure 1. Visualization of the distribution of (Z, t_i, θ_i) under π^ε

the partition of T_i in (2) is indicated by the three line segments.¹¹ Call a type $(t_i^{\theta'_i}, z_i)$ *ambiguous* if $z_i = 0$ (shown as the first segment) and *revealing* if $z_i \neq 0$ (shown as the last two segments). Draw a latent Bernoulli random variable Z . With probability $1 - \varepsilon$, we have $Z = 0$. In this case, (θ, t) is drawn from $\pi \otimes d^0$, so each player's type is ambiguous. With probability ε , we have $Z = 1$. In this case, (θ, t) is drawn from $\pi \otimes \kappa$, so each player's type could be ambiguous or revealing. In fact, the marginal $\text{marg}_T(\pi \otimes \kappa)$ has full support on T , so every type profile is consistent with $Z = 1$. Thus, for small ε , the ambiguous types make $Z = 0$ likely but not certain. The revealing types reveal that $Z = 1$.

The family $(T, t^*, \pi^\varepsilon)_{\varepsilon \in (0,1)}$ converges to the original setting π . Clearly, $\text{marg}_\Theta \pi^\varepsilon = \pi$ for each ε . And by the definition of t^* , we have $\pi \otimes d^0 = \pi \otimes t^*$, so $\|\pi^\varepsilon - \pi \otimes t^*\|_{\text{TV}} \leq \varepsilon$.

Constructing the sequential equilibrium Recall the definition of $\bar{\alpha}$ from (3). Conditional on $Z = 0$, we have $\hat{\sigma}(\bar{\alpha}(t)) = \hat{\sigma}(\alpha(\theta))$. Conditional on $Z = 1$, we have $\hat{\sigma}(\bar{\alpha}(t)) = \hat{\sigma}(\theta)$.

¹¹The third set is nonempty if and only if $\alpha_i(\Theta_i) \neq \Theta_i$.

Lemma 2 (BNE under perturbation)

Under π^ε , the profile $\hat{\sigma} \circ \bar{\alpha} = (\hat{\sigma}_i \circ \bar{\alpha}_i)_{i=1}^n$ is a Bayes–Nash equilibrium of $(\mathcal{H}, \mathcal{H}', g)$.

Proof. Fix player i and a type $t_i = (t_i^{\theta'}, z_i) \in T_i$. To show that player i does not have a profitable deviation, it suffices to show that player i does not have a profitable deviation conditional on each realization of Z .¹² We use the notation $[\cdot]$ to denote the indicator function for the predicate it encloses.

- We have $\pi^\varepsilon[Z = 1|t_i] > 0$ and

$$\pi^\varepsilon(\theta|t_i, Z = 1) = \pi(\theta_{-i}|\bar{\alpha}_i(t_i))[\theta_i = \bar{\alpha}_i(t_i)].$$

In this case, player i is certain that the players are following $\hat{\sigma}(\theta)$, so they have no profitable deviation since $\hat{\sigma}$ is a BNE (under π).

- We have $\pi^\varepsilon[Z = 0|t_i] > 0$ if and only if $z_i = 0$. If $z_i = 0$, then

$$\pi^\varepsilon(\theta|t_i, Z = 0) = \pi(\theta_{-i}|\theta'_i)[\theta_i = \theta'_i].$$

In this case, player i is certain that the players are following $\hat{\sigma}(\alpha(\theta))$, so they have no profitable deviation since $\hat{\sigma} \circ \alpha$ is a BNE (under π) by Lemma 1. $\square$

Next, we modify the profile $\hat{\sigma} \circ \bar{\alpha}$ off-path and construct associated beliefs to obtain a sequential equilibrium $(\sigma, \mu) = (\sigma^\varepsilon, \mu^\varepsilon)$ of $(\mathcal{H}, \mathcal{H}', g)$ such that $g(\sigma(t)) = g(\hat{\sigma}(\bar{\alpha}(t)))$ for all $t \in T$. The proof is then completed upon observing that for each $\theta \in \Theta$, we have

$$g(\sigma^\varepsilon(t^*(\theta))) = g(\hat{\sigma}(\bar{\alpha}(t^*(\theta)))) = g(\hat{\sigma}(\alpha(\theta))) = f(\alpha(\theta)),$$

where the second equality follows from the definition of $\bar{\alpha}$ and the last equality holds because $g \circ \hat{\sigma} = f$. Since $f \circ \alpha$ is not a selection from F , we conclude that $g(\sigma^\varepsilon(t^*(\theta))) = f(\alpha(\theta)) \notin F(\theta)$ for some $\theta \in \Theta$, as desired.

We construct $(\sigma, \mu) = (\sigma^\varepsilon, \mu^\varepsilon)$ as follows. First, we define the belief system μ . Then we use the belief system to modify each strategy $\hat{\sigma}_i \circ \bar{\alpha}_i$ at certain off-path histories to recover sequential optimality.

¹²Here and below, when computing probabilities involving Z , we think of π^ε as being extended to $\Delta(\Theta \times T \times \{0, 1\})$ as $(1 - \varepsilon)(\pi \otimes d^0) \otimes \delta_0 + \varepsilon(\pi \otimes \kappa) \otimes \delta_1$. The symbol $\otimes$ is overloaded: it can denote the product between two measures or between a measure and a kernel.

To begin, we introduce notation to classify histories. For each player i and history $h \in \mathcal{H}$, let $C_i(h)$ be the set of payoff types θ_i in Θ_i such that the sequence of moves taken by player i to reach h is consistent with $\hat{\sigma}_i(\theta_i)$. Let $\text{dev}(h) = \{j : C_j(h) = \emptyset\}$. In words, $\text{dev}(h)$ is the set of players who, at history h , are revealed to have deviated. For each i , let $\text{dev}_i(h) = \text{dev}(h) \setminus \{i\}$.

Fix player i . Define $\mu_i : T_i \times \mathcal{H} \rightarrow \Delta(\Theta \times T_{-i})$ as follows.

- If $\text{dev}_i(h) = \emptyset$, let

$$\mu_i(t_i, h)[\theta, t_{-i}] = \frac{\pi^\varepsilon(\theta, t_{-i}|t_i) \prod_{j \neq i} [\bar{\alpha}_j(t_j) \in C_j(h)]}{\sum_{(\theta', t'_{-i}) \in \Theta \times T_{-i}} \pi^\varepsilon(\theta', t'_{-i}|t_i) \prod_{j \neq i} [\bar{\alpha}_j(t'_j) \in C_j(h)]}, \quad (4)$$

where the denominator is positive because $\text{dev}_i(h) = \emptyset$ and under $\pi^\varepsilon(\cdot|t_i)$, the random variable $\bar{\alpha}_{-i}(t_{-i})$ has full support on Θ_{-i} .

- If $\text{dev}_i(h) \neq \emptyset$, let

$$\mu_i(t_i, h) = \delta_{\bar{\alpha}_i(t_i)} \otimes [\hat{\mu}_i(\bar{\alpha}_i(t_i), h) \otimes \kappa_i^h], \quad (5)$$

where $\kappa_i^h : \Theta_{-i} \rightarrow \Delta(T_{-i})$ is the product kernel $\otimes_{j \neq i} \kappa_{i,j}^h$, where

$$\kappa_{i,j}^h = \begin{cases} d_j^1 & \text{if } j \in \text{dev}_i(h), \\ \kappa_j & \text{if } j \notin \text{dev}_i(h). \end{cases}$$

Intuitively, player i infers that each player j in $\text{dev}_i(h)$ has a type t_j with $z_j(t_j) = 1$ and hence that $Z = 1$. Thus, $\bar{\alpha}(t) = \theta$.

Next, we define the strategy profile σ . Fix player i . Let z_i denote the second-component projection from T_i to $\{0, 1, 2\}$. We define $\sigma_i : T_i \rightarrow M_i$ by modifying $\hat{\sigma}_i \circ \bar{\alpha}_i$ after certain deviations by player i . Let

$$D_i = \{(t_i, h) \in T_i \times \mathcal{H} : \text{dev}_i(h) = \emptyset \text{ and } z_i(t_i) = 0 \text{ and } \bar{\alpha}_i(t_i) \notin C_i(h)\}.$$

For each $(t_i, h) \in (T_i \times \mathcal{H}) \setminus D_i$, let

$$\sigma_i(t_i, h) = \hat{\sigma}_i(\bar{\alpha}_i(t_i), h).$$

We extend σ_i to D_i as follows. For each $t_i \in T_i$, let $D_i(t_i) = \{h \in \mathcal{H} : (t_i, h) \in D_i\}$.

Fixing the opposing strategy $\hat{\sigma}_{-i} \circ \bar{\alpha}_{-i}$, player i faces a single-agent decision problem. In this problem, for each history h in $D_i(t_i)$, observe that every reachable nonterminal successor of h is also in $D_i(t_i)$. Therefore, given the beliefs μ_i , we can define $\sigma_i(t_i)$ on $D_i(t_i)$ to be sequentially optimal by working backwards, starting at the histories that directly precede terminal histories, and breaking ties arbitrarily. By (4), the beliefs μ_i satisfy Bayes' rule with respect to $\hat{\sigma}_{-i} \circ \bar{\alpha}_{-i}$ on $D_i(t_i)$. Thus, the one-shot deviation principle applies; see [Hendon, Jacobsen, and Sloth \(1996\)](#) and [Perea \(2002\)](#).

Our construction ensures the following *path-equivalence property* between σ and $\hat{\sigma} \circ \bar{\alpha}$. For each $t_i \in T_i$ and $m_{-i} \in M_{-i}$, the profiles $(\sigma_i(t_i), m_{-i})$ and $(\hat{\sigma}_i(\bar{\alpha}_i(t_i)), m_{-i})$ induce the same path of play.

Sequential optimality We prove that the strategy σ is sequentially optimal given the belief system μ . The next result allows us to reason, in certain cases, about the strategies $\hat{\sigma}_i \circ \bar{\alpha}_i$ rather than σ_i .

Lemma 3 (Sequential outcome-equivalence)

Fix a history $h \in \mathcal{H}$ and a type $t_i \in T_i$.

(i) For each $m_i \in M_i$, we have

$$\mu_i(t_i, h)[g(m_i, \sigma_{-i}(t_{-i}); h) = g(m_i, \hat{\sigma}_{-i}(\bar{\alpha}_{-i}(t_{-i})); h)] = 1.$$

(ii) If $(t_i, h) \notin D_i$, then for each $m_{-i} \in M_{-i}$,

$$g(\sigma_i(t_i), m_{-i}; h) = g(\hat{\sigma}_i(\bar{\alpha}_i(t_i)), m_{-i}; h).$$

Fix player i . By Lemma 3.i, it suffices to prove that σ_i is sequentially optimal given μ_i and $\hat{\sigma}_{-i} \circ \bar{\alpha}_{-i}$. Fix $(t_i, h) \in T_i \times \mathcal{H}$. If $(t_i, h) \in D_i$, then sequential optimality is immediate, so we may assume $(t_i, h) \notin D_i$. In this case, by Lemma 3.ii, it suffices to show that $\hat{\sigma}_i(\bar{\alpha}_i(t_i))$ is sequentially optimal at (t_i, h) given μ_i and $\hat{\sigma}_{-i} \circ \bar{\alpha}_{-i}$. To prove this, we separate into two cases.

1. Suppose $\text{dev}_i(h) = \emptyset$ and $\bar{\alpha}_i(t_i) \in C_i(h)$. Since $\hat{\sigma} \circ \bar{\alpha}$ is a BNE (Lemma 2) and, conditional on t_i , the history h is reached with positive probability under $\hat{\sigma} \circ \bar{\alpha}$, it follows that $\hat{\sigma}_i(\bar{\alpha}_i(t_i))$ is sequentially optimal at (t_i, h) .

2. Otherwise, by the definition of D_i , we must have $\text{dev}_i(h) \neq \emptyset$ or $z_i(t_i) \neq 0$. Since $\hat{\sigma}_i(\bar{\alpha}_i(t_i))$ is sequentially optimal at $(\bar{\alpha}_i(t_i), h)$ given $\hat{\mu}_i$ and $\hat{\sigma}_{-i}$, it suffices to prove that $\mu_i(t_i, h)$ assigns probability 1 to the event that $\bar{\alpha}(t) = \theta$ and $\text{marg}_{\Theta_{-i}} \mu_i(t_i, h) = \hat{\mu}_i(\bar{\alpha}_i(t_i), h)$. If $\text{dev}_i(h) \neq \emptyset$, apply (5). If $\text{dev}_i(h) = \emptyset$ and $z_i(t_i) \neq 0$, then $\pi^\varepsilon[Z = 1|t_i] = 1$, so $\pi^\varepsilon(\cdot|t_i)$ assigns probability 1 to the event that $\bar{\alpha}(t) = \theta$. Fix $\theta_{-i} \in \Theta_{-i}$. Let $\theta = (\bar{\alpha}_i(t_i), \theta_{-i})$. By (4), we have

$$\begin{aligned}
\mu_i(t_i, h)[\theta] &= \frac{\sum_{t_{-i} \in T_{-i}} \pi^\varepsilon(\theta, t_{-i}|t_i) \prod_{j \neq i} [\bar{\alpha}_j(t_j) \in C_j(h)]}{\sum_{(\theta', t'_{-i}) \in \Theta \times T_{-i}} \pi^\varepsilon(\theta', t'_{-i}|t_i) \prod_{j \neq i} [\bar{\alpha}_j(t'_j) \in C_j(h)]} \\
&= \frac{\sum_{t_{-i} \in T_{-i}} \pi^\varepsilon(\theta, t_{-i}|t_i) \prod_{j \neq i} [\theta_j \in C_j(h)]}{\sum_{(\theta', t'_{-i}) \in \Theta \times T_{-i}} \pi^\varepsilon(\theta', t'_{-i}|t_i) \prod_{j \neq i} [\theta'_j \in C_j(h)]} \\
&= \frac{\pi^\varepsilon(\theta|t_i) \prod_{j \neq i} [\theta_j \in C_j(h)]}{\sum_{\theta' \in \Theta} \pi^\varepsilon(\theta'|t_i) \prod_{j \neq i} [\theta'_j \in C_j(h)]} \\
&= \frac{\pi(\theta_{-i}|\bar{\alpha}_i(t_i)) \prod_{j \neq i} [\theta_j \in C_j(h)]}{\sum_{\theta'_{-i} \in \Theta_{-i}} \pi(\theta'_{-i}|\bar{\alpha}_i(t_i)) \prod_{j \neq i} [\theta'_j \in C_j(h)]} \\
&= \hat{\mu}_i(\bar{\alpha}_i(t_i), h)[\theta_{-i}],
\end{aligned}$$

where the last equality follows from the consistency of $\hat{\mu}_i$.

Consistency To prove consistency, we construct a sequence of totally mixed strategies such that for each player i , the probability of trembling after observing $z_i \neq 1$ becomes arbitrarily small relative to the probability of any sequence of trembles after observing $z_i = 1$.

Since $(\hat{\sigma}, \hat{\mu})$ is a sequential equilibrium, there exists a sequence, $(\hat{\sigma}^k, \hat{\mu}^k)$, of totally mixed behavioral strategy profiles and belief-system profiles such that for each player i , the following hold: (a) for each k , the belief system $\hat{\mu}_i^k$ is derived from $\hat{\sigma}_{-i}^k$ via Bayes' rule; and (b) the sequence $(\hat{\sigma}_i^k, \hat{\mu}_i^k)$ converges to $(\delta_{\hat{\sigma}_i}, \hat{\mu}_i)$ pointwise on $\Theta_i \times \mathcal{H}$. For each $\theta_i \in \Theta_i$ and $h \in \mathcal{H}$, let $\hat{L}_i^k(\theta_i, h)$ denote the probability that type θ_i makes the sequence of moves to reach h under strategy $\hat{\sigma}_i^k$. Thus,

$$\hat{\mu}_i^k(\theta_i, h)[\theta_{-i}] = \frac{\pi(\theta_{-i}|\theta_i) \prod_{j \neq i} \hat{L}_j^k(\theta_j, h)}{\sum_{\theta'_{-i} \in \Theta_{-i}} \pi(\theta'_{-i}|\theta_i) \prod_{j \neq i} \hat{L}_j^k(\theta'_j, h)}.$$

Since $\hat{\sigma}_i^k \rightarrow \delta_{\hat{\sigma}_i}$, we have

$$\lim_k \hat{L}_i^k(\theta_i, h) = [\theta_i \in C_i(h)]. \tag{6}$$

Next, we introduce a parameter sequence that will uniformly bound the tremble probabilities of any player j with $z_j(t_j) \neq 1$. For each k , let

$$\lambda^k = \frac{1}{k} \cdot \min_{i, \theta_i, h} \hat{L}_i^k(\theta_i, h), \quad (7)$$

where the minimum is over all players i , types $\theta_i \in \Theta_i$ and histories $h \in \mathcal{H}$. Observe that $\lambda^k > 0$ because each strategy $\hat{\sigma}_i^k$ is totally mixed, and the minimum is over a finite set.

Define σ_i^k as follows. For each $t_i \in T_i$ and $h \in \mathcal{H}$, let

$$\sigma_i^k(t_i, h) = \begin{cases} \hat{\sigma}_i^k(\bar{\alpha}_i(t_i), h) & \text{if } z_i(t_i) = 1, \\ (1 - \lambda^k)\delta_{\sigma_i(t_i, h)} + \lambda^k \hat{\sigma}_i^k(\bar{\alpha}_i(t_i), h) & \text{if } z_i(t_i) \neq 1. \end{cases}$$

These mixtures are computed in $\Delta(M_i(h))$. By construction, σ_i^k is totally mixed. Moreover, $\sigma_i^k \rightarrow \delta_{\sigma_i}$. This is easily verified by cases. If $z_i(t_i) = 1$, then $D_i(t_i) = \emptyset$ so this holds because $\hat{\sigma}_i^k \rightarrow \delta_{\hat{\sigma}_i}$. If $z_i(t_i) \neq 1$, this holds because $\lambda_k \rightarrow 0$.

Since σ_{-i}^k is totally mixed, we can define $\mu_i^k: T_i \times \mathcal{H} \rightarrow \Delta(\Theta \times T_{-i})$ using Bayes' rule. Let $L_i^k(t_i, h)$ denote the probability that player i with type t_i makes the sequence of moves to reach h under strategy σ_i^k . Let

$$\mu_i^k(t_i, h)[\theta, t_{-i}] = \frac{\pi^\varepsilon(\theta, t_{-i}|t_i) \prod_{j \neq i} L_j^k(t_j, h)}{\sum_{(\theta', t'_{-i}) \in \Theta \times T_{-i}} \pi^\varepsilon(\theta', t'_{-i}|t_i) \prod_{j \neq i} L_j^k(t'_j, h)}. \quad (8)$$

To prove that $\mu_i^k \rightarrow \mu_i$, we record two properties of the likelihoods that follow from the definition of σ_i^k . For each $t_i \in T_i$ and $h \in \mathcal{H}$, if $z_i(t_i) = 1$, then

$$L_i^k(t_i, h) = \hat{L}_i^k(\bar{\alpha}_i(t_i), h). \quad (9)$$

For each $t_i \in T_i$ and $h \in \mathcal{H}$, since $\sigma_i^k \rightarrow \delta_{\sigma_i}$, the path-equivalence property gives

$$\lim_k L_i^k(t_i, h) = [\bar{\alpha}_i(t_i) \in C_i(h)]. \quad (10)$$

We need one additional result. Let $[n] = \{1, \dots, n\}$. For each $h \in \mathcal{H}$, let $\text{ob}_i(h) =$

$([n] \setminus \{i\}) \setminus \text{dev}_i(h)$. For any $t_i \in T_i$ and $h \in \mathcal{H}$, let

$$P_i(t_i) = \{(\theta, t_{-i}) \in \Theta \times T_{-i} : (\pi \otimes \kappa)[\theta, t] > 0\},$$

$$G_i(h) = \{(\theta, t_{-i}) \in \Theta \times T_{-i} : z_j(t_j) = 1 \ \forall j \in \text{dev}_i(h) \text{ and } \bar{\alpha}_\ell(t_\ell) \in C_\ell(h) \ \forall \ell \in \text{ob}_i(h)\}.$$

Let $E_i(t_i, h) = P_i(t_i) \cap G_i(h)$. Finally, let

$$\tilde{\mu}_i^k(t_i, h) = \delta_{\bar{\alpha}_i(t_i)} \otimes [\hat{\mu}_i^k(\bar{\alpha}_i(t_i), h) \otimes \kappa_i^h].$$

Lemma 4 (Inference after deviations)

For each player i , type $t_i \in T_i$, and history $h \in \mathcal{H}$ satisfying $\text{dev}_i(h) \neq \emptyset$, we have

$$\lim_k \mu_i^k(t_i, h)(E_i(t_i, h)) = \lim_k \tilde{\mu}_i^k(t_i, h)(E_i(t_i, h)) = 1.$$

Fix $(t_i, h) \in T_i \times \mathcal{H}$. We prove that $\mu_i^k(t_i, h) \rightarrow \mu_i(t_i, h)$. If $\text{dev}_i(h) = \emptyset$, simply apply (10) to (8) and compare the result with (4). Therefore, we may assume $\text{dev}_i(h) \neq \emptyset$. The consistency of $\hat{\mu}$ implies that $\tilde{\mu}_i^k(t_i, h) \rightarrow \mu_i(t_i, h)$. Therefore, it suffices to prove that $\|\mu_i^k(t_i, h) - \tilde{\mu}_i^k(t_i, h)\|_{\text{TV}} \rightarrow 0$. We introduce notation for the numerators in the definitions of $\mu_i^k(t_i, h)$ and $\tilde{\mu}_i^k(t_i, h)$. For each $(\theta, t_{-i}) \in \Theta \times T_{-i}$, let

$$\begin{aligned} w_i^k(t_i, h)[\theta, t_{-i}] &= \pi^\varepsilon(\theta, t_{-i}|t_i) \prod_{j \neq i} L_j^k(t_j, h), \\ \tilde{w}_i^k(t_i, h)[\theta, t_{-i}] &= [\theta_i = \bar{\alpha}_i(t_i)] \pi(\theta_{-i}|\theta_i) \kappa_i^h[t_{-i}|\theta_{-i}] \prod_{j \neq i} \hat{L}_j^k(\theta_j, h). \end{aligned} \tag{11}$$

We check that $w_i^k(t_i, h)$ and $\tilde{w}_i^k(t_i, h)$ are strictly positive on $E_i(t_i, h)$. The likelihoods appearing in (11) are strictly positive because σ^k and $\hat{\sigma}^k$ are totally mixed. Since $\pi^\varepsilon[Z = 1|t_i] > 0$, we have $\pi^\varepsilon(\theta, t_{-i}|t_i) > 0$ for each $(\theta, t_{-i}) \in P_i(t_i)$. And for each $(\theta, t_{-i}) \in E_i(t_i, h) = P_i(t_i) \cap G_i(h)$, we have $\bar{\alpha}(t) = \theta$ and $\kappa_{i,j}^h[t_j|\theta_j] > 0$ for all $j \neq i$. We show that

$$\max_{(\theta, t_{-i}) \in E_i(t_i, h)} \left| \frac{w_i^k(t_i, h)[\theta, t_{-i}]}{\tilde{w}_i^k(t_i, h)[\theta, t_{-i}]} - \frac{\pi^\varepsilon[Z = 1|t_i]}{2^{|\text{dev}_i(h)|}} \right| \rightarrow 0. \tag{12}$$

This limit $\pi^\varepsilon[Z = 1|t_i]/2^{|\text{dev}_i(h)|}$ is strictly positive and does not depend on (θ, t_{-i}) ,

so we conclude that

$$\|\mu_i^k(t_i, h)[\cdot|E_i(t_i, h)] - \tilde{\mu}_i^k(t_i, h)[\cdot|E_i(t_i, h)]\|_{\text{TV}} \rightarrow 0.$$

By Lemma 4, it follows that $\|\mu_i^k(t_i, h) - \tilde{\mu}_i^k(t_i, h)\|_{\text{TV}} \rightarrow 0$.

Now we prove (12). Since $E_i(t_i, h)$ is finite, it suffices to prove pointwise convergence over $E_i(t_i, h)$. Fix $(\theta, t_{-i}) \in E_i(t_i, h)$. The definition of $G_i(h)$ implies that $z_j(t_j) = 1$ for all $j \in \text{dev}_i(h)$. Since $\text{dev}_i(h) \neq \emptyset$, we have

$$\begin{aligned} \pi^\varepsilon(\theta, t_{-i}|t_i) &= \pi^\varepsilon[Z = 1|t_i](\pi \otimes \kappa)[\theta, t_{-i}|t_i] \\ &= \pi^\varepsilon[Z = 1|t_i]\pi(\theta_{-i}|\theta_i)\kappa_{-i}(t_{-i}|\theta_{-i}). \end{aligned} \tag{13}$$

Substitute in (13) and apply (9) for each player $j \in \text{dev}_i(h)$ to get

$$w_i^k(t_i, h)[\theta, t_{-i}] = \pi^\varepsilon[Z = 1|t_i]\pi(\theta_{-i}|\theta_i)\kappa_{-i}(t_{-i}|\theta_{-i}) \prod_{j \in \text{dev}_i(h)} \hat{L}_j^k(\theta_j, h) \prod_{\ell \in \text{ob}_i(h)} L_\ell^k(t_\ell, h).$$

Dividing by $\tilde{w}_i^k(t_i, h)[\theta, t_{-i}]$, the likelihood factors $\hat{L}_j^k(\theta_j, h)$ cancel for each $j \in \text{dev}_i(h)$. For all $\ell \in \text{ob}_i(h)$, we have $\kappa_{i,\ell}^h = \kappa_\ell$. For all $j \in \text{dev}_i(h)$, we have $\kappa_j(t_j|\theta_j) = 1/2$ and $\kappa_{i,j}^h[t_j|\theta_j] = 1$. Therefore,

$$\frac{w_i^k(t_i, h)[\theta, t_{-i}]}{\tilde{w}_i^k(t_i, h)[\theta, t_{-i}]} = \pi^\varepsilon[Z = 1|t_i]2^{-|\text{dev}_i(h)|} \prod_{\ell \in \text{ob}_i(h)} \frac{L_\ell^k(t_\ell, h)}{\hat{L}_\ell^k(\theta_\ell, h)}.$$

For each $\ell \in \text{ob}_i(h)$, we have $\bar{\alpha}_\ell(t_\ell) = \theta_\ell \in C_\ell(h)$, so $L_\ell^k(t_\ell, h) \rightarrow 1$ by (10) and $\hat{L}_\ell^k(\theta_\ell, h) \rightarrow 1$ by (6). Hence, (12) follows.

A.2 Proof of Lemma 3

Fix i and $t_i \in T_i$. We begin by proving a stronger, sequential version of the path-equivalence property. Let $D_i(t_i) = \{h \in \mathcal{H} : (t_i, h) \in D_i\}$. Fix $h \notin D_i(t_i)$. We claim that starting from h , if player i uses $\hat{\sigma}_i(\bar{\alpha}_i(t_i))$, then play will never reach a history in $D_i(t_i)$, no matter how their opponents play. Indeed, if $\text{dev}_i(h) \neq \emptyset$, then $D_i(t_i)$ does not contain any successor of h . If $z_i(t_i) \neq 0$, then $D_i(t_i) = \emptyset$. If $\text{dev}_i(h) = \emptyset$ and $z_i(t_i) = 0$, then $\bar{\alpha}_i(t_i) \in C_i(h)$ and this property is preserved as long as player i follows $\hat{\sigma}_i(\bar{\alpha}_i(t_i))$.

From sequential path-equivalence, part [ii](#) is immediate. Now we prove part [i](#). Fix $h \in \mathcal{H}$. For each j , let $D_j(h) = \{t_j \in T_j : (t_j, h) \in D_j\}$. By applying sequential path-equivalence to each player $j \neq i$, it suffices to show that $\mu_i(t_i, h)[t_j \in D_j(h)] = 0$ for each $j \neq i$. Fix $j \neq i$. If $\text{dev}_i(h) = \emptyset$, then from [\(4\)](#), we have $\mu_i(t_i, h)[\bar{\alpha}_j(t_j) \in C_j(h)] = 1$, so we are done. Therefore, we may assume $\text{dev}_i(h) \neq \emptyset$. We separate into cases. If $j \notin \text{dev}_i(h)$, then $\text{dev}_j(h) \neq \emptyset$, so $D_j(h) = \emptyset$. If $j \in \text{dev}_i(h)$, then from [\(5\)](#), we have $\mu_i(t_i, h)[z_j(t_j) = 1] = 1$, so we are done.

A.3 Proof of Lemma 4

Fix player i , type $t_i \in T_i$, and history $h \in \mathcal{H}$ with $\text{dev}_i(h) \neq \emptyset$. Fix $(\theta, t_{-i}) \in (\Theta \times T_{-i}) \setminus E_i(t_i, h)$. We show that $\mu_i^k(t_i, h)[\theta, t_{-i}] \rightarrow 0$ and $\tilde{\mu}_i^k(t_i, h)[\theta, t_{-i}] \rightarrow 0$. Since the set $(\Theta \times T_{-i}) \setminus E_i(t_i, h)$ is finite, the result then follows.

First, suppose that there exists $j \in \text{dev}_i(h)$ such that $z_j(t_j) \neq 1$. Since $\kappa_{i,j}^h = d_j^1$, we immediately get $\tilde{\mu}_i^k(t_i, h)[\theta, t_{-i}] = 0$ for each k . To analyze $\mu_i^k(t_i, h)[\theta, t_{-i}]$, consider the type $t'_j = (t_j^{\alpha_j(t_j)}, 1)$. Let $t'_{-i} = (t'_j, t_{-\{j,i\}})$ and $\theta' = \bar{\alpha}(t_i, t'_{-i})$. By [\(9\)](#), we have $L_j^k(t'_j, h) = \hat{L}_j^k(\bar{\alpha}_j(t_j), h)$. Since $j \in \text{dev}_i(h)$, it follows from the path-equivalence property that player j with type t_j must tremble at least once to reach h , so $L_j^k(t_j, h) \leq \lambda^k$. Therefore,

$$\frac{L_j^k(t_j, h)}{L_j^k(t'_j, h)} \leq \frac{\lambda_k}{\hat{L}_j^k(\bar{\alpha}_j(t_j), h)} \leq \frac{1}{k}. \quad (14)$$

Since $\pi^\varepsilon[Z = 1|t_i] > 0$ and $\theta' = \bar{\alpha}(t_i, t'_{-i})$, we have $\pi^\varepsilon(\theta', t'_{-i}|t_i) > 0$. By Bayes' rule,

$$\mu_i^k(t_i, h)[\theta, t_{-i}] \leq \frac{1}{k} \cdot \frac{\pi^\varepsilon(\theta, t_{-i}|t_i)}{\pi^\varepsilon(\theta', t'_{-i}|t_i)} \cdot \mu_i^k(t_i, h)[\theta', t'_{-i}]. \quad (15)$$

Pass to the limit in k to conclude that $\mu_i^k(t_i, h)[\theta, t_{-i}] \rightarrow 0$.

We may therefore suppose that $z_j(t_j) = 1$ for every $j \in \text{dev}_i(h)$. Since $\text{dev}_i(h) \neq \emptyset$, we have $\pi^\varepsilon(\theta, t_{-i}|t_i) = \pi^\varepsilon(Z = 1|t_i)(\pi \otimes \kappa)(\theta, t_{-i}|t_i)$. Thus, if $(\theta, t_{-i}) \notin P_i(t_i)$, we have $\mu_i^k(t_i, h)[\theta, t_{-i}] = 0$ and $\tilde{\mu}_i^k(t_i, h)[\theta, t_{-i}] = 0$ for all k . Therefore, we may assume that $(\theta, t_{-i}) \in P_i(t_i)$. Since $(\theta, t_{-i}) \notin E_i(t_i, h) = P_i(t_i) \cap G_i(h)$ and $z_j(t_j) = 1$ for every $j \in \text{dev}_i(h)$, it follows that $\bar{\alpha}_\ell(t_\ell) \notin C_\ell(h)$ for some $\ell \in \text{ob}_i(h)$. Since $\ell \in \text{ob}_i(h)$, the set $C_\ell(h)$ is nonempty. Choose $\theta'_\ell \in C_\ell(h)$. Let $t'_\ell = (t_\ell^{\theta'_\ell}, 1)$. Let $t'_{-i} = (t'_\ell, t_{-\{\ell,i\}})$ and $\theta' = \bar{\alpha}(t_i, t'_{-i})$. By construction, (θ', t'_{-i}) is in $P_i(t_i)$ and t' agrees with t over $\text{dev}_i(h)$,

so $w_i^k(t_i, h)[\theta', t'_{-i}]$ and $\tilde{w}_i^k(t_i, h)[\theta', t'_{-i}]$ are both strictly positive. We have

$$\mu_i^k(t_i, h)[\theta, t_{-i}] = \frac{\pi^\varepsilon(\theta, t_{-i}|t_i)}{\pi^\varepsilon(\theta', t'_{-i}|t_i)} \frac{L_\ell^k(t_\ell, h)}{L_\ell^k(t'_\ell, h)} \mu_i^k(t_i, h)[\theta', t'_{-i}].$$

Since $\bar{\alpha}_\ell(t_\ell) \notin C_\ell(h)$ and $\bar{\alpha}_\ell(t'_\ell) = \theta'_\ell \in C_\ell(h)$, we use (10) to conclude that $L_\ell^k(t_\ell, h)/L_\ell^k(t'_\ell, h)$ converges to 0. Hence, $\mu_i^k(t_i, h)[\theta, t_{-i}] \rightarrow 0$. Turning to $\tilde{\mu}_i^k$, since t'_{-i} and t_{-i} only differ in position ℓ , we have

$$\tilde{\mu}_i^k(t_i, h)[\theta, t_{-i}] = \frac{\pi(\theta_{-i}|\theta_i)}{\pi(\theta'_{-i}|\theta'_i)} \frac{\kappa_\ell(t_\ell|\theta_\ell)}{\kappa_\ell(t'_\ell|\theta'_\ell)} \frac{\hat{L}_\ell^k(\theta_\ell, h)}{\hat{L}_\ell^k(\theta'_\ell, h)} \tilde{\mu}_i^k(t_i, h)[\theta', t'_{-i}].$$

Since $\theta_\ell = \bar{\alpha}_\ell(t_\ell) \notin C_\ell(h)$ and $\theta'_\ell \in C_\ell(h)$, we use (6) to conclude that $\hat{L}_\ell^k(\theta_\ell, h)/\hat{L}_\ell^k(\theta'_\ell, h)$ converges to 0, so $\tilde{\mu}_i^k(t_i, h)[\theta, t_{-i}] \rightarrow 0$.

A.4 Proof of Remark 2

If $M_i(h)$ is countable for each $h \in \mathcal{H}$ then $\mathcal{H}$ is countable. If in addition $(\mathcal{H}, \mathcal{H}', g)$ admits a sequential best response for each player at each history for all beliefs and opposing pure strategies,¹³ the result goes through: The existence of best responses ensures that we can construct sequentially optimal strategies on D_i , and the only other part of the proof where finiteness is used is in defining λ^k . If $\mathcal{H}$ is countably infinite, modify this step as follows. Choose a bijection $K: \mathcal{H} \rightarrow \mathbb{N}$. For each k , let $\mathcal{H}^k = \{h \in \mathcal{H} : K(h) \leq k\}$. Then let

$$\lambda^k = \frac{1}{k} \cdot \min_{i \in [n], \theta_i \in \Theta_i, h \in \mathcal{H}^k} \hat{L}_i^k(\theta_i, h). \quad (16)$$

Thus, $\lambda^k > 0$. For each fixed history h , we have $h \in \mathcal{H}^k$ for all $k \geq K(h)$, so the argument goes through. Note, however, that the convergence may not be uniform over $T_i \times \mathcal{H}$.

¹³This property is satisfied, for instance, by the integer-game mechanisms in [Bergin and Sen \(1998\)](#) and [Baliga \(1999\)](#).

A.5 Proof of Proposition 1

Consider the social choice environment from Section 2, the SCF f given by $f(\theta) = x^\theta$, and the three-stage mechanism $(\mathcal{H}, \mathcal{H}', g)$ introduced in Section 2. In Section 2.1, we proved that $(\mathcal{H}, \mathcal{H}', g)$ sequentially implements f . Now we check that f is not Bayesian monotone. Consider any deception profile α such that α_1 and α_2 are both bijections, and at least one is different from the identity. Since α is not the identity, we have $f \circ \alpha \neq f$. Suppose that there exists some player i , payoff type $\bar{\theta}_i$, and social choice function f' such that $f' \circ \alpha \succ_{\bar{\theta}_i} f \circ \alpha$ and, for each $\theta_i \in \Theta_i$, we have $f \succsim_{\theta_i} f'_{i, \alpha_i(\bar{\theta}_i)}$. In particular, taking $\theta_i = \bar{\theta}_i$ and using indifference over the image of f , we get

$$f' \circ \alpha \succ_{\bar{\theta}_i} f \circ \alpha \sim_{\bar{\theta}_i} f \succsim_{\bar{\theta}_i} f'_{i, \alpha_i(\bar{\theta}_i)}.$$

Since π is uniform,

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i(f'(\alpha_i(\bar{\theta}_i), \alpha_{-i}(\theta_{-i})), \bar{\theta}_i) > \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f'(\alpha_i(\bar{\theta}_i), \theta_{-i}), \bar{\theta}_i).$$

This is a contradiction because the two summations are the same (since $\alpha_{-i}(\Theta_{-i}) = \Theta_{-i}$).

Let $\bar{\eta} = 1/20$. Let $(T, t^*, \hat{\pi})$ be a KPT information structure satisfying $\|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}} < \bar{\eta}$. Let $\eta = \|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}}$. For each $\theta \in \Theta$, we have $\hat{\pi}[\theta, t^*(\theta)] \geq 1/4 - \eta > 0$, so type $t_i^*(\theta_i)$ knows that their payoff type is θ_i . Let (σ, μ) be a sequential equilibrium. We claim that for each player i and payoff type θ_i , type $t_i^*(\theta_i)$ reports θ_i in the first stage. From the claim, it then follows that for each $\theta'_i, \theta_{-i} \in \{0, 1\}$, if player i reports θ'_i then type $t_{-i}^*(\theta_{-i})$ of player $-i$ assigns to the event $\theta_i = \theta'_i$ probability strictly more than $2/3$ since

$$\frac{\hat{\pi}(t^*(\theta'_i, \theta_{-i}))}{\hat{\pi}(t^*(\theta_{-i})) - \hat{\pi}(t^*(1 - \theta'_i, \theta_{-i}))} \geq \frac{1/4 - \eta}{1/4 + \eta} > 2/3.$$

Hence, player $-i$ does not challenge when they receive the distinguished signal (since a challenged player defends if and only if their report is truthful). We conclude that $g(\sigma(t^*(\theta))) = f(\theta)$ for all $\theta \in \Theta$.

It remains to prove the claim. Suppose that for some player i and payoff type θ_i , type $t_i^*(\theta_i)$ reports $1 - \theta_i$ under σ_i . For each $\theta_{-i} \in \Theta_{-i}$, if player i reports $1 - \theta_i$, then type $t_{-i}^*(\theta_{-i})$ of player $-i$ assigns to the event $\theta_i \neq \theta'_i$ probability strictly more than

1/3 since

$$\frac{\hat{\pi}(t^*(\theta_i, \theta_{-i}))}{\hat{\pi}(t^*(\theta_{-i}))} \geq \frac{1/4 - \eta}{1/2 + \eta} > 1/3.$$

Hence, player $-i$ challenges the report $1 - \theta_i$ whenever they receive a distinguished signal. If challenged, player i 's payoff from reporting $1 - \theta_i$ is at most $-3 + 2 = -1$. Otherwise, their payoff is at most 2. Type $t_i^*(\theta_i)$ believes that their opponent received a distinguished signal with probability strictly more than $2/3$ since $(1/2 - \eta)/(1/2 + \eta) > 2/3$. Therefore, their expected payoff is strictly less than 0. But by reporting truthfully, not challenging, and defending if challenged, player i guarantees a payoff of 0, yielding a contradiction.

A.6 Proof of Proposition 2

We modify the social choice environment from Section 2. As before, there are two players, denoted $i = 1, 2$. Each player i observes their payoff type $\theta_i \in \Theta_i = \{0, 1\}$, which is drawn uniformly and independently. But now, only player 2 can challenge, and only when player 1 reports $\theta'_1 = 1$. Formally, the set of decisions is

$$A = \left\{ x^{\theta', c_2, d_1} : \theta' \in \Theta \text{ and } c_2, d_1 \in \{0, 1\} \right\}.$$

Utility functions will be specified below. Let $(\mathcal{H}, \mathcal{H}', g)$ be the following three-stage mechanism.

1. Each player i reports $\theta'_i \in \Theta_i$. If $\theta'_1 = 0$, the game ends and the decision is $x^{\theta'} := x^{\theta', 0, 0}$. Otherwise, play proceeds to stage 2.
2. Players observe the reports from stage 1. Player 2 chooses whether to challenge ($c_2 = 1$) or not ($c_2 = 0$). If player 2 does not challenge, the game ends and the decision is $x^{\theta'} := x^{\theta', 0, 0}$. Otherwise, play proceeds to stage 3.
3. Player 1 observes the challenge decision from stage 2. Player 1 chooses whether to defend ($d_1 = 1$) or not ($d_1 = 0$). The game ends and the decision is $x^{\theta', 1, d_1}$.

The players have private values given by the following tables: For player 1,

$u_1(x^{\theta'}, 0)$	$\theta'_2 = 0$	$\theta'_2 = 1$	$u_1(x^{\theta'}, 1)$	$\theta'_2 = 0$	$\theta'_2 = 1$	(17)
$\theta'_1 = 0$	0	0	$\theta'_1 = 0$	0	-2	
$\theta'_1 = 1$	0	-1	$\theta'_1 = 1$	-1	0	

For player 2,

$$\begin{array}{c|cc} u_2(x^{\theta'}, 0) & \theta'_2 = 0 & \theta'_2 = 1 \\ \hline \theta'_1 = 0 & 0 & -1 \\ \theta'_1 = 1 & 0 & -1 \end{array} \quad \begin{array}{c|cc} u_2(x^{\theta'}, 1) & \theta'_2 = 0 & \theta'_2 = 1 \\ \hline \theta'_1 = 0 & -1 & 0 \\ \theta'_1 = 1 & 0 & -1 \end{array}. \quad (18)$$

For all $\theta' \in \Theta$ and $c_2, d_1 \in \{0, 1\}$, let

$$\begin{aligned} u_1(x^{\theta', c_2, d_1}, \theta_1) &= u_1(x^{\theta'}, \theta_1) + c_2 v_1(\theta'_1, d_1, \theta_1), \\ u_2(x^{\theta', c_2, d_1}, \theta_2) &= u_2(x^{\theta'}, \theta_2) + c_2 w_2(d_1), \end{aligned} \quad (19)$$

where

$$v_1(\theta'_1, d_1, \theta_1) = \begin{cases} -1 + d_1 & \text{if } \theta_1 = \theta'_1, \\ -3 - d_1 & \text{if } \theta_1 \neq \theta'_1, \end{cases}$$

and

$$w_2(d_1) = 2 - 3d_1.$$

Upon being challenged, it is strictly optimal for player 1 to defend if they reported truthfully and not to defend if they misreported. For player 2, challenging changes their utility by 2 if player 1 does not defend but by -1 if player 1 defends. Define the SCF $f: \Theta \rightarrow A$ by $f(\theta) = x^\theta$. Define the deception profile $\alpha = (\alpha_1, \alpha_2)$ by $\alpha_1(\theta_1) = 1 - \theta_1$ and $\alpha_2(\theta_2) = 0$.

To prove Proposition 2, we begin with a lemma about the set of BNE.

Lemma 5 (BNE outcomes)

Under π , every BNE of $(\mathcal{H}, \mathcal{H}', g)$ induces either the SCF f or $f \circ \alpha$.

Proof of Lemma 5. Consider an arbitrary BNE σ . First, we check that no challenge occurs on path. Recall that player 2 can challenge only if player 1 reports $\theta'_1 = 1$. Suppose that this report is on path. If type $\theta_1 = 0$ reports $\theta'_1 = 0$, then player 2 concludes that the report $\theta'_1 = 1$ is truthful, and hence player 2 cannot challenge this report (since player 1 will defend). If type $\theta_1 = 0$ reports $\theta'_1 = 1$ and player 2 ever challenges this report, then type $\theta_1 = 0$ gets utility strictly below 0, and hence type $\theta_1 = 0$ can profitably deviate to $\theta'_1 = 0$ to secure a payoff of 0.

It remains to show that, under σ , the players either (i) always report truthfully, or (ii) report according to α . Type $\theta_2 = 0$ can guarantee payoff 0 by reporting $\theta'_2 = 0$

(and never challenging). Since there are no challenges on path, type $\theta_2 = 0$ cannot report $\theta'_2 = 1$ since this would yield payoff -1 . We consider two cases according to the report of type $\theta_2 = 1$.

- If type $\theta_2 = 1$ reports $\theta'_2 = 1$, then both types of player 1 uniquely maximize the expectation of u_1 by reporting truthfully, and hence must do so (since defending after reporting truthfully and being challenged yields the same payoff as not being challenged). Thus, (i) holds.
- If type $\theta_2 = 1$ reports $\theta'_2 = 0$, then type $\theta_1 = 1$ strictly prefers to report $\theta'_1 = 0$. It follows that type $\theta_1 = 0$ reports $\theta'_1 = 1$; otherwise, type $\theta_2 = 1$ could profitably deviate to the report $\theta'_2 = 1$. $\square$

Using Lemma 5, we now prove Proposition 2 in three steps.

1. We show that $(\mathcal{H}, \mathcal{H}', g)$ sequentially implements f . By Lemma 5, it suffices to show that f can be induced by a sequential equilibrium of $(\mathcal{H}, \mathcal{H}', g)$, but $f \circ \alpha$ cannot. For f , consider the following strategy profile: both players report truthfully; player 2 never challenges; and, when challenged, player 1 defends if and only if they reported truthfully. Sequential optimality is clear in stage 3, and hence in stage 2. For stage 1, optimality follows from computing expected payoffs, under the uniform prior, in (17) and (18).

For $f \circ \alpha$, suppose for a contradiction that there exists a sequential equilibrium that induces $f \circ \alpha$. Consider the report history $\theta' = (1, 0)$. Player 2 is certain that player 1 misreported. By sequential optimality, player 1 will not defend a challenge, so challenging is strictly optimal for player 2, hence $f \circ \alpha$ is not induced.

2. We show that f is not Bayesian monotone. First note that $f \circ \alpha \neq f$. Player 1 cannot be a whistle-blower because each type of player 1 gets expected utility 0 under $f \circ \alpha$, and no outcome gives any type of player 1 a strictly positive utility. For player 2, suppose for a contradiction that there exists some payoff type $\bar{\theta}_2$ and some social choice function f' such that $f' \circ \alpha \succ_{\bar{\theta}_2} f \circ \alpha$ and, for each $\theta_2 \in \Theta_2$, we have $f \succ_{\theta_2} f'_{2, \alpha_2(\bar{\theta}_2)}$. In particular, we take $\theta_2 = \bar{\theta}_2$. Since both types of player 2 are indifferent between f and $f \circ \alpha$, we have

$$f' \circ \alpha \succ_{\bar{\theta}_2} f \circ \alpha \sim_{\bar{\theta}_2} f \succ_{\bar{\theta}_2} f'_{2, \alpha_2(\bar{\theta}_2)}.$$

Since π is uniform, plugging in the function α gives

$$\sum_{\theta_1 \in \Theta_1} u_2(f'(1 - \theta_1, 0), \bar{\theta}_2) > \sum_{\theta_1 \in \Theta_1} u_2(f'(\theta_1, 0), \bar{\theta}_2).$$

This is a contradiction because the two summations are the same.

3. We complete the proof. Since the game is finite, there exists $\eta_0 > 0$ such that, for any information structure $(T, t^*, \hat{\pi})$ satisfying $\|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}} < \eta_0$, every sequential equilibrium (σ, μ) of $(\mathcal{H}, \mathcal{H}', g)$ is such that the strategy profile $\bar{\sigma}$ defined by $\bar{\sigma}_i(\theta_i) = \sigma_i(t_i^*(\theta_i))$ is a BNE under π . Let $\bar{\eta} = \min\{\eta_0, 1/16\}$.

Let $(T, t^*, \hat{\pi})$ be a CI information structure such that $\|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}} < \bar{\eta}$. Let $\eta = \|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}}$. Let (σ, μ) be a sequential equilibrium. Since $\|\hat{\pi} - \pi \otimes t^*\|_{\text{TV}} < \eta_0$, the strategy profile $\bar{\sigma}$ defined by $\bar{\sigma}_i(\theta_i) = \sigma_i(t_i^*(\theta_i))$ is a BNE under π . By Lemma 5, the SCF induced by $\bar{\sigma}$ is either f or $f \circ \alpha$. Suppose for a contradiction that $\bar{\sigma}$ induces $f \circ \alpha$. Since $\bar{\sigma}$ is a BNE under π , it follows that $\bar{\sigma}_1(0)$ must specify defending, if challenged, after report profile $(1, 0)$. By the sequential optimality of σ , type $t_1^*(0)$ must assign positive probability to the event $\theta_1 = 1$. We check that type $t_1^*(0)$ can profit by deviating to the report $\theta'_1 = 0$, which cannot be challenged.

- If $\theta_1 = 0$, then it can be seen from (17) that this deviation weakly increases player 1's payoff.
- If $\theta_1 = 1$, and player 2's signal is either $t_2^*(0)$ or $t_2^*(1)$, then player 2 reports 0 and does not challenge, so by (17), player 1's gain is 1. Player 1's worst gain from the deviation is -2 . By conditional independence, the probability that player 2 does not have either distinguished signal, conditional on each signal for player 1 and each state, is at most $\eta/(1/4 - \eta) < 1/3$, where we used the fact that $\eta < 1/16$. Thus, player 1's expected gain from deviating is strictly more than $(2/3) - 2(1/3) = 0$.

Since $t_1^*(0)$ assigns positive probability to the event $\theta_1 = 1$, this deviation is strictly profitable, which is a contradiction.

References

AGHION, P., D. FUDENBERG, R. HOLDEN, T. KUNIMOTO, AND O. TERCIEUX (2012): "Subgame-Perfect Implementation under Information Perturbations," *The Quarterly Journal of Economics*, 127, 1843–1881. [2, 4, 12, 16]

- BALIGA, S. (1999): “Implementation in Economic Environments with Incomplete Information: The Use of Multi-Stage Games,” *Games and Economic Behavior*, 27, 173–183. [4, 15, 32]
- BATTIGALLI, P. AND E. CATONINI (2026): “Monotonicity and Robust Implementation Under Forward-Induction Reasoning,” *Econometrica*, 94, 505–536. [6]
- BATTIGALLI, P. AND M. SINISCALCHI (1999): “Interactive Beliefs, Epistemic Independence and Strong Rationalizability,” *Research in Economics*, 53, 247–273. [6]
- (2003): “Rationalization and Incomplete Information,” *Advances in Theoretical Economics*, 3, 1–46. [6]
- (2007): “Interactive Epistemology in Games with Payoff Uncertainty,” *Research in Economics*, 61, 165–184. [6]
- BERGEMANN, D. AND S. MORRIS (2005): “Robust Mechanism Design,” *Econometrica*, 73, 1771–1813. [19]
- BERGIN, J. AND A. SEN (1998): “Extensive Form Implementation in Incomplete Information Environments,” *Journal of Economic Theory*, 80, 222–256. [4, 15, 32]
- BÖRGERS, T. (1994): “Weak Dominance and Approximate Common Knowledge,” *Journal of Economic Theory*, 64, 265–276. [5]
- BRUSCO, S. (1995): “Perfect Bayesian Implementation,” *Economic Theory*, 5, 419–444. [4]
- (2006): “Perfect Bayesian Implementation in Economic Environments,” *Journal of Economic Theory*, 129, 1–30. [4]
- CHEN, Y.-C., R. HOLDEN, T. KUNIMOTO, Y. SUN, AND T. WILKENING (2023): “Getting Dynamic Implementation to Work,” *Journal of Political Economy*, 131, 285–387. [5]
- CHUNG, K.-S. AND J. C. ELY (2003): “Implementation with Near-Complete Information,” *Econometrica*, 71, 857–871. [5]
- DEKEL, E. AND D. FUDENBERG (1990): “Rational Behavior with Payoff Uncertainty,” *Journal of Economic Theory*, 52, 243–267. [5]
- DUGGAN, J. (1998): “An Extensive Form Solution to the Adverse Selection Problem in Principal/Multi-Agent Environments,” *Review of Economic Design*, 3, 167–191. [4]

- ECCLES, P. AND N. WEGNER (2016): “Robustness of Subgame Perfect Implementation,” Staff Working Paper 601, Bank of England. [5]
- FRICK, M. AND A. ROMM (2015): “Rational Behavior under Correlated Uncertainty,” *Journal of Economic Theory*, 160, 56–71. [5, 20]
- FUDENBERG, D., D. M. KREPS, AND D. K. LEVINE (1988): “On the Robustness of Equilibrium Refinements,” *Journal of Economic Theory*, 44, 354–380. [4, 5, 13, 15, 19]
- HENDON, E., H. J. JACOBSEN, AND B. SLOTH (1996): “The One-Shot-Deviation Principle for Sequential Rationality,” *Games and Economic Behavior*, 12, 274–282. [26]
- JACKSON, M. O. (1991): “Bayesian Implementation,” *Econometrica*, 59, 461–477. [2, 4, 8, 11, 12, 17, 21]
- KOHLBERG, E. AND J.-F. MERTENS (1986): “On the strategic stability of equilibria,” *Econometrica*, 54, 1003–1037. [20]
- MASKIN, E. (1999): “Nash Equilibrium and Welfare Optimality,” *The Review of Economic Studies*, 66, 23–38. [4, 12]
- MONDERER, D. AND D. SAMET (1989): “Approximating common knowledge with common beliefs,” *Games and Economic Behavior*, 1, 170–190. [14]
- MOORE, J. AND R. REPULLO (1988): “Subgame Perfect Implementation,” *Econometrica*, 56, 1191–1220. [4, 6]
- MÜLLER, C. (2016): “Robust Virtual Implementation under Common Strong Belief in Rationality,” *Journal of Economic Theory*, 162, 407–450. [5]
- PALFREY, T. R. AND S. SRIVASTAVA (1991): “Nash Implementation Using Undominated Strategies,” *Econometrica*, 59, 479–501. [5]
- PEREA, A. (2002): “A Note on the One-deviation Property in Extensive Form Games,” *Games and Economic Behavior*, 40, 322–338. [26]
- PIERMONT, E. AND P. ZUAZO-GARIN (2026): “Misspecified Information in Dynamic Games,” Working paper. [6]
- WANG, Z. (2026): “Robust Predictions in Dynamic Games with Incomplete Information,” Working paper. [6]
- WEINSTEIN, J. AND M. YILDIZ (2007): “A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements,” *Econometrica*, 75, 365–400. [5]