Robust equilibrium outcomes in sequential games under almost common certainty of payoffs

Satoru Takahashi, Olivier Tercieux

Robust Equilibrium Outcomes in Sequential Games under Almost Common Certainty of Payoffs

Satoru Takahashi[†] and Olivier Tercieux[‡]

[This version: April 27, 2020]

Abstract

We analyze the robustness of equilibria in sequential games when there is almost common certainty of payoffs. We show that a generic extensive-form game may have no robust equilibrium behavior, but has at least one robust equilibrium outcome, which is induced by a proper equilibrium in its normal-form representation. Therefore, backward induction leads to a unique robust outcome in a generic perfect-information game. We also discuss close relation between robustness to incomplete information and strategic stability. Finally, we present the implications of our results for the robustness of subgame-perfect implementation.
(*Journal of Economic Literature* Classification Number: C72, C73, D82, D83)

Keywords: incomplete information, robustness, higher-order beliefs, refinements, subgame-perfect implementation

† Department of Economics, National University of Singapore, Singapore, Singapore. Email: ecsst@nus.edu.sg

‡ Paris School of Economics and CNRS, Paris, France. Email: tercieux@pse.ens.fr

Acknowledgement

The paper has been circulated under slightly different titles, such as "Robust Equilibria in Sequential Games under Almost Common Certainty of Payoffs." The authors thank the participants of Paris Workshop on Game Theory and seminar participants at Arizona State University, Columbia, Northwestern and Princeton. Gratefully acknowledged are helpful conversations and comments from Sylvain Chassang, Navin Kartik, Stephen Morris, and Klaus Ritzberger. We also thank two anonymous referees and the Editor, Marciano Siniscalchi, for many useful suggestions. All remaining errors are, of course, solely by the authors.

1. Introduction

In a sequential game, how should a player respond if he observes an off-the-equilibrium-path history? One, perhaps conventional, approach is that he attributes such an off-path history to "trembling hands," idiosyncratic errors made by his opponents in preceding decision nodes, and plays a best response to the original equilibrium strategies in the continuation game. Thus, we are naturally led to solution concepts such as (subgame-)perfect or sequential equilibria. Another, perhaps less conventional, approach is to allow for uncertainty in payoffs and explain the off-path history by the possibility of "crazy types" who maximize payoff functions different from ones for "normal types." According to this approach, the player in the off-path history updates his belief about the opponents' types and maximizes the expected continuation payoff against the opponents' continuation strategies. While this paper focuses on the latter approach, we identify useful formal connections between these two approaches.

Following Fudenberg, Kreps, and Levine (1988), we fix an extensive-form game, and consider incomplete-information perturbations where before playing the same extensive game form, nature chooses a profile of types for players and each player is privately informed of his type. The types can be either "normal" where payoffs on terminal histories are the same as in the original game, or "crazy" where payoffs on terminal histories can be different from ones in the original game. The prior probability of the profile of normal types is close to 1. With this setup, one can ask if the equilibrium *behavior* is robust. That is, an equilibrium of the original extensive-form game is robust if for any incomplete-information perturbation, there exists a Bayesian Nash equilibrium under which normal types play strategies that are close to the original equilibrium.

This robustness test is known to have little bite under static environments. Indeed, in any generic normal-form game, all equilibria are regular, and hence robust (Fudenberg and Tirole (1991, Theorems 14.5, 14.6)). In contrast, sequential games generally have no regular equilibrium since players are indifferent among strategies that differ only off the equilibrium path. In fact, we construct a two-stage perfect-information game whose unique subgame-perfect equilibrium is not robust to incomplete information, as best responses in off-path histories are sensitive to introduction of "crazy types" (Example 1 in Section 2).[1] Hence, this robustness test has a non-trivial bite in sequential games.

This motivates us to weaken the robustness test in order to hopefully restore the existence result. More specifically, motivated by the literature on mechanism design, where the social planner is mainly interested in equilibrium *outcomes*, i.e., distributions over terminal histories, and not in equilibrium behavior *per se*, we weaken the robustness test by requiring that in the incomplete-information perturbations, there exists a Bayesian Nash equilibrium under which normal types induce an outcome that is close to that induced by the original equilibrium.[2] This weakening can have far reaching consequences and will be crucial for our main result. For example, the

[1]This is not a knife-edge case in that the same result holds for a nonempty open set of payoffs on terminal histories.

[2]In other words, we apply the robustness test to a set of equilibria that induce the same outcome. The literature on strategic stability has also adopted set-valued solution concepts by showing that no singleton-valued solution concept satisfies (i) admissibility and iterated dominance, (ii) backward induction and invariance, or (iii) backward induction and forward induction (Kohlberg and Mertens (1986, pp. 1015, 1018, 1029)). In contrast, Morris and Ui

equilibrium outcome of the two-stage game mentioned above is robust to incomplete information. But in general, not every subgame-perfect equilibrium outcome is robust, as we show by another game, which is of imperfect information due to simultaneous moves (Example 2 in Section 2). Hence, one may still be worried about non-existence of robust outcomes.

It turns out that we can show the generic existence of robust outcomes (Theorem 1). That is, every generic extensive-form game has at least one robust equilibrium outcome. This is our first main result. The heart of the proof is to show that every hyperstable component (maximal connected and closed set) of equilibria, defined in Kohlberg and Mertens (1986), is robust to incomplete information. Generic existence follows immediately, given the existence of hyperstable components and the generic finiteness of equilibrium outcomes.

While the generic existence result is useful, and its proof reveals that hyperstability is a sufficient condition for robustness, it does not provide much insight into which equilibrium is *not* robust. Our second main result exhibits necessary conditions for robustness. Indeed, we show that every robust equilibrium outcome is stable in the sense of Kohlberg and Mertens (1986), and induced by some proper equilibrium (Propositions 2 and 3). Moreover, we show by means of an example that neither stability nor properness is sufficient for robustness. These results suggest that the class of perturbations used for our robustness notion is significantly larger than those for stability and for proper equilibria. Of course, this does not weaken the main message in AFHKT since full implementation does fail in nearby environments (i.e., undesirable equilibria may still appear in nearby games). In addition, combined with the generic existence result, the necessity of properness generates an interesting link between the conventional trembling-hand approach and our approach. That is, every generic perfect-information game has a unique robust outcome, which coincides with the backward-induction outcome (Corollary 1). More generally, our necessary and sufficient conditions shed light on how the traditional refinement literature connects with the more recent literature on robustness to incomplete information.

Our main motivation in this paper is to study the role of the standard assumption that the environment is of complete information when the game played is sequential. The assumption that the game is of complete information is particularly prominent in mechanism design and, more specifically, in the implementation literature. The growing body of work on robust mechanism design (e.g., Bergemann and Morris (2012)) which intends to assess the role of (strong) informational assumptions in mechanism design partly motivated our exercise. Sequential mechanisms are of special importance in this literature. For instance, it is well known that many social choice functions cannot be implemented using static mechanisms when Nash equilibrium is used as a solution concept (Maskin (1999), Jackson (1989)). However, in a seminal paper, Moore and Repullo (1988) showed that—under complete information—only weak conditions are needed when using sequential mechanisms and subgame perfection as a solution concept. This result had an important impact beyond the implementation literature and yielded the "implementation critique" of the property right theory of the firm (e.g., Maskin and Tirole (1999)). Recently, Aghion, Fudenberg, Holden,

(2005) introduce a set-valued notion of robustness because even a generic normal-form game may not have a robust equilibrium in the sense of Kajii and Morris (1997a).

Kunimoto, and Tercieux (2012) (AFHKT, hereafter) showed that Moore and Repullo's implementation result may not be maintained when the information of agents is slightly perturbed. In particular, AFHKT showed that in many cases where Moore and Repullo's mechanisms are used, in incomplete-information perturbations, undesirable equilibria may appear and no equilibrium behavior is close to the (unique) subgame-perfect equilibrium behavior of the original game induced by the mechanism.[3] Loosely speaking, the latter result uses a notion of robustness close to the standard notion mentioned above (*à la* Fudenberg and Tirole (1991)). However, what eventually matters in mechanism design is the outcome of the game and the result by AFHKT leaves open the possibility that some desirable equilibrium outcome survives in nearby situations. Using the leading example of AFHKT, we show how slight variations on our arguments can complete the picture: we prove that the original subgame-perfect equilibrium outcome is robust (Proposition 5). Thus, at least for this example, our results allow to save part of the properties of the Moore and Repullo's mechanism proving that the (unique) desirable equilibrium outcome survives in nearby incomplete-information games. In addition, we argue that the logic behind the example extends much beyond this specific instance.

1.1. **Related Literature.** Kajii and Morris (1997a) introduce the notion of robust equilibria as in our paper, but they focus on simultaneous-move games and allow for perturbations where there may not be almost common certainty of the original payoff functions (see Section 6). Chassang and Takahashi (2011) apply their robustness requirement to repeated games and prove a folk theorem in (dynamically) robust equilibria. Weinstein and Yildiz (2007) study the impact of perturbations of higher-order beliefs (formalized by the product topology in the universal type space) in simultaneous-move games. The impact of perturbations of higher-order beliefs in dynamic environments is studied in Chen (2012), Penta (2012), and Weinstein and Yildiz (2013). In all these papers, there need not be almost common certainty of the original payoffs in the perturbations. In static environments, robustness to perturbations assuming almost common certainty of the original payoffs has been studied in Monderer and Samet (1989), Fudenberg and Tirole (1991, Chapter 14), Chen, Di Tillio, Faingold, and Xiong (2010), and Morris, Takahashi, and Tercieux (2012). Fudenberg, Kreps, and Levine (1988) introduce the notion of "elaborations" that respect dynamic structures and investigate the condition under which a given equilibrium behavior can be approximated by a sequence of strict equilibria in nearby elaborations. Our paper studies robustness issues in sequential games where we impose almost common certainty of payoffs.

In the Stackelberg oligopoly and two-stage sequential games more generally, subgame-perfect equilibrium behavior may not be robust if the followers observe the leaders' actions with noise, but generically, every such game has at least one subgame-perfect equilibrium outcome that is robust to noisy monitoring (Bagwell (1995), van Damme and Hurkens (1997), and Güth, Kirchsteiger, and Ritzberger (1998)). Our generic existence result and its proof are similar in spirit to those in Güth,

[3]AFHKT show that in incomplete-information perturbations of the complete-information game induced by the Moore and Repullo mechanism, no equilibrium behavior is close to the (unique) subgame-perfect equilibrium behavior of the original game provided that the mechanism originally implements a non-strategy-proof social choice function. In addition, for any (finite) extensive-form mechanism, if, under complete information, a non-Maskin monotonic social function can be implemented with this mechanism, then, in incomplete-information perturbations, an undesirable equilibrium outcome exists.

Kirchsteiger, and Ritzberger (1998). However, we consider another class of perturbations, and our robustness notion is implied by hyperstability while the weaker notion of essential sets is sufficient in their case. Indeed, we show, by means of an example, that our robustness notion is different from theirs (see Remark 3).

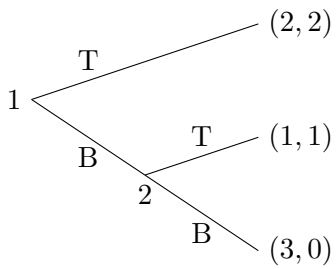
There is a large literature on reputation in repeated games, which studies the impact of introducing "crazy types" on equilibrium behavior and outcomes (see Mailath and Samuleson (2006)). Despite apparent similarity between the reputation literature and our paper, the two differ with respect to the order of limits. Namely, the reputation literature first fixes a small probability of "crazy types" and then takes sufficiently long horizon of the game (or the discount factor sufficiently close to 1) whereas we first fix an extensive-form game, including the horizon, and then take small incomplete-information perturbations.

2. Motivating Examples

In this section, we give a first insight into the robustness test we will study in this paper. Example 1 shows that if we stick to the standard definition of robustness in terms of equilibrium *behavior*, then even a unique subgame-perfect equilibrium (of a perfect-information game) with strict incentives at each node may not be robust. We then explain how the definition of robustness can be modified to restore the robustness of the equilibrium in this specific game. This gives an idea of our main theorem, which will show that any generic extensive-form game has a robust equilibrium *outcome*. In Example 2, we provide an example of a game that has a subgame-perfect equilibrium with strict incentives at each node, which induces an outcome that is not robust in our sense. This shows that the weaker robustness test we propose still has a significant bite.

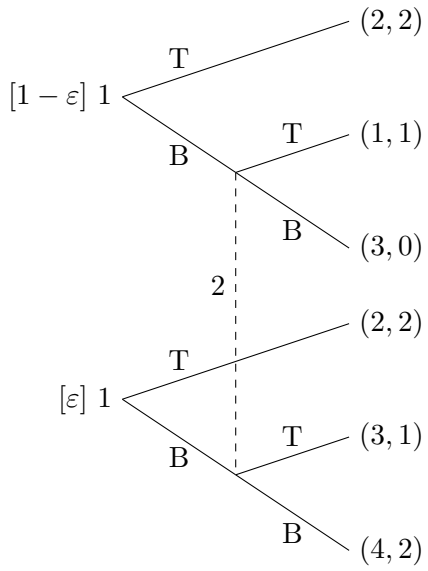
2.1. Example 1.

Let us consider a perfect-information game with two players $i = 1, 2$ and two stages. Player 1 plays first and chooses between Top (T) and Bottom (B). If (and only if) Player 1 plays B, then Player 2 has an opportunity to choose between action Top (T) and Bottom (B). Payoffs are given as follows:



In the above game, the unique subgame-perfect equilibrium is to play action T for each player. We first show that the behavior prescribed by this equilibrium is not robust to incomplete information, i.e., one can perturb the original game by introducing a small amount of uncertainty in such a way that the perturbed game has no equilibrium where Player 1 follows the strategies prescribed by the equilibrium. To see this, let us build an *incomplete-information elaboration* of the above perfect-information game. Player 1 is assumed to have two types: t_1^* and t_1^{crazy} while Player 2 has

a single type t_2^*. The prior probability over the set of possible profiles of types puts a probability $\varepsilon > 0$ on $(t_1^{\text{crazy}}, t_2^*)$ and $1-\varepsilon$ on (t_1^*, t_2^*). Payoffs are given as follows:



When Player 1 is of type t_1^*, the ex-post payoffs for both players are given by the upper tree, and when Player 1 is of type t_1^{crazy}, the ex-post payoffs are given by the lower tree. Hence, with (arbitrarily) small probability, the same game tree is played but with different payoffs on terminal histories. Observe that when Player 1 is of type t_1^{crazy}, he has a dominant strategy to play action B. Now, let us show that for any $\varepsilon > 0$, there cannot be an equilibrium where Player 1 plays T when he is of type t_1^*. Proceed by contradiction and assume there is such an equilibrium. If Player 1 upon receiving t_1^* plays according to the equilibrium, he gets payoffs 2. Now, if Player 1 deviates, by Bayes' rule, as long as ε is strictly positive, Player 2 must believe with probability one that Player 1 is of type t_1^{crazy}. By construction of the incomplete-information game, Player 2 must then be playing B. Coming back to Player 1 of type t_1^*, the deviation must then make him strictly better off. This is a contradiction.

Note that we can extend the above argument and show that there is no equilibrium in mixed strategies where both t_1^* and t_2^* play T with high probabilities. Indeed, it is easy to check that if t_2^* were doing so, then Player 1 would have no incentive to mix and the argument above would apply.[4]

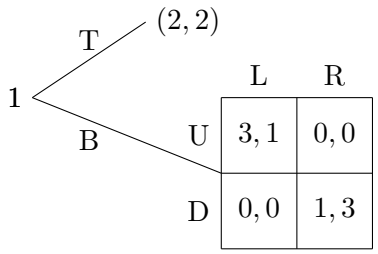
In many contexts, what really matters is not the equilibrium behavior *per se* but the distribution over terminal histories induced by the equilibrium. For instance, this is particularly clear when we consider a mechanism design setting where the designer cares about the allocation selected by the mechanism after a report of agents. Of course, here the selected allocation depends only on the terminal history and not on the specific behavior of agents off the equilibrium path. The designer may want to ensure that a nearby allocation is selected when we slightly perturb the game (e.g.,

[4]In this example, Player 2's payoffs on terminal histories depend on Player 1's type, and hence and Player 2 learns her own payoffs through Player 1's action in the first stage. This feature is not essential to our argument. We can change the example by adding another move by Player 3 after Player 2's choice of B; see Online Appendix C.1 for details.

see AFHKT). In such a case, he would not care if in those perturbed games, agents' equilibrium behavior changes as long as the terminal history induced by equilibrium behavior (and hence, the allocation selected by the mechanism) does not change too much. Given the above argument, it is natural to ask whether in any perturbed game, one can find an equilibrium that induces a distribution over terminal histories—or more simply, an *outcome*—that is close to the outcome induced by a given equilibrium. More specifically, in our original two-player example, can we find an equilibrium in the perturbed game under which t_1^* would play T with a probability that goes to 1 as ε goes to 0? The answer is yes. Indeed, fix $\varepsilon > 0$ and consider a profile of strategies where Player 1 mixes when he is of type t_1^* putting probability $\frac{\varepsilon}{1-\varepsilon}$ on B (and the complement probability on T) and plays B with probability one when he is of type t_1^{crazy}. Player 2 in turn mixes and plays T and B with probability fifty-fifty. It is easily checked that Player 1 is indeed indifferent between his two actions when he is of type t_1^* while, using Bayes' rule, Player 2 believes that Player 1 is of type t_1^* and of type t_1^{crazy} with probability fifty-fifty, and so Player 2 is also indifferent between his two actions. Note that while under this equilibrium, players' behavior is far from the original equilibrium (Player 2 mixes with probability fifty-fifty on each action), this equilibrium induces a distribution over terminal histories that gets closer and closer (as ε vanishes) to the dirac measure on T.

One of our main results will show that any generic sequential game has a sequential equilibrium outcome that is robust to a class of incomplete-information elaborations. Also, it is known that in generic simultaneous-move games, any equilibrium is robust in our sense (Fudenberg and Tirole (1991, Theorems 14.5, 14.6)). Hence, a natural question arises: does the robustness test (informally) formulated here have any bite? I.e., can we find games where some sequential equilibrium outcome is not robust in the sense that for some nearby incomplete-information elaboration, no equilibrium induces a distribution over terminal histories that is close to this outcome? The following example shows that non-robust equilibrium outcomes may exist in generic sequential games.

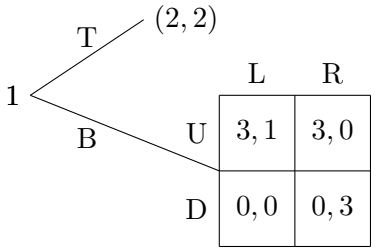
2.2. **Example 2.** Let us consider an outside-option game with two players $i = 1, 2$ and two stages. Player 1 plays first and chooses between Top (T) and Bottom (B). If (and only if) Player 1 plays B, then the players play a battle-of-sexes. Payoffs are given as follows:



In the above game, there are two subgame-perfect equilibrium outcomes. One is that Player 1 plays T, which is sustained by an off-path behavior under which the players coordinate on (D, R) (or on the mixed equilibrium) if Player 1 plays action B. The other is that Player 1 plays B and then in the subgame, the players coordinate on (U, L). In the sequel, we show that the first equilibrium outcome is not robust to incomplete information, i.e., one can build nearby incomplete-information

games under which no (mixed) equilibrium induces a distribution over terminal histories that yields Player 1 to play T with large ex ante probability. As will become clear, this example shows that the robustness test we formulate in this paper has a significant bite. In particular, even a subgame-perfect equilibrium (outcome) with strict incentives at each node may not be robust in our sense.[5]

To see this, let us build and *incomplete-information elaboration* of the above game. Player 1 is assumed to have two types: t_1^* and t_1^{crazy} while Player 2 has a single type t_2^*. The prior probability over the set of possible profile of types puts a probability $\varepsilon > 0$ on $(t_1^{\text{crazy}}, t_2^*)$ and $1-\varepsilon$ on (t_1^*, t_2^*). When Player 1 is the type t_1^*, the ex-post payoffs for both players are given by the above game. However, when Player 1 is type t_1^{crazy}, the ex-post payoffs are given as follows:



Hence, here again, with (arbitrarily) small probability, the same game tree is played but with different payoffs for Player 1. Observe that when Player 1 is of type t_1^{crazy}, he has a strictly dominant strategy to play B and then U.

In the remaining, we show that for any $\varepsilon > 0$, in the above incomplete-information game, there is no equilibrium where Player 1 of type t_1^* plays T with positive probability. First, it is easy to show that t_1^* cannot play T with probability one. To see this, proceed by contradiction and assume that there is an equilibrium under which t_1^* plays T with probability one. Then whenever Player 2 sees that Player 1 has played B, his only belief consistent with Bayes' rule puts probability one on the event that Player 1 is of type t_1^{crazy}. Hence, in case Player 2 sees Player 1 playing action B, he believes that Player 1 will be playing U in the subgame and so he himself plays action L in the subgame. Given that, it is clearly profitable for Player 1 of type t_1^* to deviate from his equilibrium action and play B and then U. This yields a contradiction.

Now, let us show that for any $\varepsilon > 0$, in the above incomplete-information game, there is no equilibrium where Player 1 of type t_1^* mixes over actions T and B. Indeed, for this to be possible, t_1^* has to be indifferent between T and B. Hence, in the subgame, Player 1 must play U. Otherwise, if Player 1 were to put a positive probability on D in the subgame, his expected equilibrium payoff in the subgame would be no more than his expected payoff from playing D, which is at most 1 regardless of Player 2's equilibrium action. Hence, t_1^* would not be indifferent between T and B, which is a contradiction. Now, note that Player 1 plays U in the subgame irrespective of his

[5]Note that the non-robust outcome here does not survive iterative deletion of weakly dominated strategies: for Player 1, playing B and then D is strictly dominated by playing just T in the first stage. Given this, playing L weakly dominates playing R for Player 2. Given this, Player 1 strictly prefers playing B and then U to playing T. Hence, (BU, L) is the only action profile surviving iterative deletion of weakly dominated strategies. More generally, if an outcome is robust, then it must be induced by an equilibrium that survives iterative deletion of weakly dominated strategies, but the converse does not hold in general. See Section 4.1 and Remark 5.

type. Given that, Player 2 has a unique best response to play L, and here again, t_1^* would not be indifferent between T and B. In conclusion, there is no equilibrium that makes Player 1 of type t_1^* play T with positive probability, proving our claim.

3. Framework and Generic Existence

In this section, we formally define our robustness notion as well as present our first main result (generic existence) and its proof.

3.1. **Complete-Information Normal-Form Games.** A (finite) complete-information normal-form game $G = (N, (A_i)_{i\in N}, (v_i)_{i\in N})$ consists of a finite set of players N and for each player $i \in N$, a finite set of player i's actions A_i and his payoff function $v_i \colon A \to \mathbb{R}$ with $A = \prod_{i\in N} A_i$. A mixed action of player i is denoted $\alpha_i \in \Delta(A_i)$, and a profile of mixed actions is denoted $\alpha = (\alpha_i)_{i\in N}$ with no subscript (similar abbreviations will be used throughout the paper).[6] The domain of v_i is extended to mixed action profiles in the standard way. We say that α is a (Nash) equilibrium if $v_i(\alpha) \geq v_i(a_i, \alpha_{-i})$ for any $i \in N$ and $a_i \in A_i$.

3.2. **Complete- and Incomplete-Information Extensive-Form Games.** Following Osborne and Rubinstein (1994), we define a (finite) incomplete-information extensive game form $\Gamma = (N, H, \tau, f_c, (\mathcal{I}_i)_{i\in N})$ as follows. (1) N is a finite set of players. (2) H is a finite set of histories that satisfies (2a) $\emptyset \in H$ and (2b) if $(a^k)_{k=1,\ldots,K} \in H$ and $L < K$ then $(a^k)_{k=1,\ldots,L} \in H$. (3) τ is a function that assigns to each nonterminal history a member of $N \cup \{c\}$ with the interpretation that $\tau(h) \in N$ is the player who takes an action after history h, whereas chance determines the action if $\tau(h) = c$. (4) For each player $i \in N$, $\mathcal{I}_i$ is his information partition on H_i, where H_i is the set of histories after which player i moves. For any history $h \in H_i$, $I_i(h) \in \mathcal{I}_i$ denotes the information set that contains h, and $S_i(h) = \{a_i : (h, a_i) \in H\}$ denotes the set of actions available to player i at h. It is required that $S_i(\cdot)$ be $\mathcal{I}_i$-measurable, i.e., $S_i(\bar{h}) = S_i(h)$ for all $\bar{h} \in I_i(h)$. (5) Finally, f_c is a function that associates with every history h for which $\tau(h) = c$, a probability distribution $f_c(\cdot \mid h)$ on $S_c(h)$. We make the usual restriction to ensure that the extensive game form has perfect recall. Let Z be the set of terminal histories and $g_i \colon Z \to \mathbb{R}$ be the payoff function of player i; the domain of g_i is naturally extended to $\Delta(Z)$. Then $(\Gamma, (g_i)_{i\in N})$ is an incomplete-information extensive-form game. When there is no move by nature, such a game will be called a complete-information extensive-form game. Note that even in a complete-information extensive-form game, there may exist non-trivial information sets for players. This reflects the fact that when a player has an opportunity to choose an action, he may not know what actions his opponents' have chosen in the past. Hence, although formally speaking we do not allow more than one player to move after any history, this allows to capture situations essentially similar to that where players choose actions simultaneously. We say that an extensive-form game is of perfect information if it is a complete-information extensive-form game with $I_i(h) = \{h\}$ for all players i and all histories h.

A pure strategy for player i is a mapping $s_i \colon H_i \to \bigcup_{h\in H_i} S_i(h)$, where for all $h \in H_i$, $s_i(h) \in S_i(h)$ and $s_i(h) = s_i(\bar{h})$ for all $\bar{h} \in I_i(h)$. Let S_i be the set of pure strategies of player i. A

[6] All along the paper, for a given finite set X, we denote $\Delta(X)$ for the set of probability distributions over X.

mixed strategy σ_i is a probability distribution over pure strategies, $\sigma_i \in \Delta(S_i)$. A profile of pure strategies $s = (s_i)_{i \in N}$ induces a distribution $z(s)$ over terminal histories; we extend the domain of $z(\cdot)$ to mixed-strategy profiles. Given an incomplete-information extensive-form game $(\Gamma, (g_i)_{i \in N})$, we will denote $G^{\Gamma,g} = (N, (S_i)_{i \in N}, (g_i^{\mathcal{N}})_{i \in N})$ for its (ex ante) normal-form representation where $g_i^{\mathcal{N}}(s) = g_i(z(s))$. A mixed-strategy profile is a Bayesian Nash equilibrium in $(\Gamma, (g_i)_{i \in N})$ if it is a Nash equilibrium in $G^{\Gamma,g}$.

We will restrict our attention to specific incomplete-information extensive-form games called elaborations. An elaboration of a complete-information extensive-form game $(\Gamma, (g_i)_{i \in N})$ is an incomplete-information game where the first move is done by nature and determines a profile of types. Each player is informed of his own type and then plays the game form Γ. Formally, an elaboration of $(\Gamma, (g_i)_{i \in N})$ with a finite set of types T_i for each player i and a prior $P \in \Delta(T)$ is an incomplete-information extensive-form game $(N, \{\emptyset\} \cup T \times H, \tilde{\tau}, \tilde{f}_c, (\widetilde{\mathcal{I}}_i)_{i \in N}, (u_i)_{i \in N})$ where the set of players is the same as in the complete-information game and before the game is played, nature chooses a profile of types according to the prior probability P, i.e., $\tilde{\tau}(\emptyset) = c$, $S_c(\emptyset) = T = \prod_i T_i$ and $f_c(t \mid \emptyset) = P(t)$. Then the same extensive game form is played, i.e., for all $t \in T$, and $h \in H$, $\tilde{\tau}(t, h) = \tau(h)$ and $\tilde{f}_c(\cdot \mid (t, h)) = f_c(\cdot \mid h)$. In addition, for any history $h' = (t, h)$, player i is privately informed of his own type, i.e., $\widetilde{I}_i(h') = \{t_i\} \times T_{-i} \times I_i(h)$. The set of terminal histories is then $T \times Z$, where Z is the set of terminal histories of the complete-information game, and hence for each player i, payoffs are defined by $u_i \colon Z \times T \to \mathbb{R}$. An elaboration of the complete-information extensive-form game $(\Gamma, (g_i)_{i \in N})$ will be denoted by $U = (\Gamma, P, T, (u_i)_{i \in N})$. We assume that payoffs are uniformly bounded over elaborations, i.e., there is $M \geq 0$ such that for each elaboration $U = (\Gamma, P, T, (u_i)_{i \in N})$, $\max_{(z,t) \in Z \times T} |u_i(z, t)| \leq M$. Note that in an elaboration, a pure strategy for player i is a mapping from T_i to S_i, and a mixed strategy can be identified with two kinds of behavioral strategy: a mapping from T_i to $\Delta(S_i)$ and a mapping from $T_i \times H_i$ to $\bigcup_{h \in H_i} \Delta(S_i(h))$.

3.3. **Robustness to Incomplete Information.** Now, we want to formalize the idea that an elaboration $U = (\Gamma, P, T, (u_i)_{i \in N})$ is close to a complete-information game $(\Gamma, (g_i)_{i \in N})$ if with high probability, the payoffs under U are the same as those under $(\Gamma, (g_i)_{i \in N})$. In order to do so, we follow the approach proposed by Fudenberg, Kreps, and Levine (1988).

Definition 1. *Fix $\varepsilon \geq 0$. $U = (\Gamma, P, T, (u_i)_{i \in N})$ is an ε-elaboration of $(\Gamma, (g_i)_{i \in N})$ if it is an elaboration of $(\Gamma, (g_i)_{i \in N})$ and there is a type profile $t^* \in T$ satisfying $P(t^*) \geq 1 - \varepsilon$ and $u_i(\cdot, t^*) = g_i(\cdot)$ for all $i \in N$.*[7]

Given Example 1, we will not focus on the robustness of equilibrium strategies, as is often done in the literature (see Fudenberg, Kreps, and Levine (1988), Monderer and Samet (1989), and Kajii and Morris (1997a) among others), but on the robustness of equilibrium outcomes—recall our terminology: an outcome is a distribution over terminal histories. In order to formalize our robustness notion, we will measure distance between two distributions over terminal histories, μ and ν, by the max norm: $\|\mu - \nu\| = \max_{z \in Z} |\mu(z) - \nu(z)|$. In the sequel, we abuse notations and use

[7]We could instead require that $\max_{z \in Z} |u_i(z, t^*) - g_i(z)| \leq \varepsilon$; all but one of our results would remain unchanged. See footnote 16.

symbol z as an element of Z as well as the outcome function of the complete-information game, i.e., the function mapping $\prod_i \Delta(S_i)$ to Z.

Definition 2. *Fix a complete-information extensive-form game $(\Gamma, (g_i)_{i\in N})$. An equilibrium outcome $\mu \in \Delta(Z)$ is robust if for any $\delta > 0$, there is $\varepsilon > 0$ such that any ε-elaboration has a Bayesian Nash equilibrium σ satisfying $\|\mu - z(\sigma(t^*))\| < \delta$.*[8]

In the above definition, we use Bayesian Nash equilibrium as a solution concept. In general, nothing precludes that an equilibrium outcome is robust when we use Bayesian Nash equilibria in the ε-elaborations while this outcome would not be robust anymore when we use a stronger solution concept such as perfect Bayesian or sequential equilibrium. As we will discuss later, our generic existence result remains unchanged if we use any solution concept with the existence property coarser than properness (which includes both perfect Bayesian and sequential equilibria), see Remark 1.

Although our robustness notion in Definition 2 shares some similarities with those introduced by Kajii and Morris (1997a,b) as well as Monderer and Samet (1989), there are some essential differences. The most important is that under our definition, when ε vanishes, the prior of an ε-elaboration puts probability 1 on the single profile of types t^* and not on a set of types (all having ex post payoffs identical to those of the complete-information game). This seemingly small difference turns out to have far reaching consequences as discussed in Section 6.

3.4. **The Main Theorem: Existence.** In generic normal-form games, we know that every Nash equilibrium is robust (Fudenberg and Tirole (1991, Theorems 14.5, 14.6)). While we showed in Example 1 that with the usual definition, there may be no robust equilibrium at all, we prove that with the version of robustness given in Definition 2 above based on equilibrium outcomes, we recover a generic existence result.

Given a complete-information game form Γ, let $\mathcal{G}$ be the set of payoff function profiles $(g_i)_{i\in N}$ such that $(\Gamma, (g_i)_{i\in N})$ has finitely many equilibrium outcomes. By Kreps and Wilson (1982, Theorem 2), $\mathcal{G}$ is generic, i.e., $\mathbb{R}^{Z\times N} \setminus \mathcal{G}$ has Lebesgue measure zero in $\mathbb{R}^{Z\times N}$.

Theorem 1. *If $(g_i)_{i\in N} \in \mathcal{G}$, then at least one of the equilibrium outcomes is robust.*

The genericity condition is indispensable. For example, consider the following two-player normal-form game with only one active player:

$$\begin{array}{c|c} & \text{X} \\ \hline \text{T} & 0,0 \\ \hline \text{B} & 0,1 \\ \hline \end{array} \tag{1}$$

It is easy to see that this game has no robust equilibrium outcome. Indeed, for any $\varepsilon \in (0, 1]$, we can construct an elaboration with $T_1 = \{t_1^*\}$, $T_2 = \{t_2^*, t_2^{\mathrm{T}}\}$, and $P(t^*) = 1 - P(t_1^*, t_2^{\mathrm{T}}) = \varepsilon$, $u_1(\cdot, t^*) = u_1(\mathrm{B}, \mathrm{X}, t_1^*, t_2^{\mathrm{T}}) = 0$ and $u_1(\mathrm{T}, \mathrm{X}, t_1^*, t_2^{\mathrm{T}}) = 1$. Then this elaboration has a unique Bayesian Nash

[8]Here, $\sigma(t^*)$ denotes the profile of mixed strategies in the complete-information game form Γ played by t^*.

equilibrium, where Player 1 plays T for sure. Similarly, we can exchange T and B and construct another elaboration where Player 1 plays B for sure in the unique Bayesian Nash equilibrium.[9]

We now comment on a number of implications of the theorem. First, Theorem 1 implies that the unique subgame-perfect equilibrium outcome of Example 1 is robust. Indeed, in this example, there is a unique equilibrium outcome (the dirac measure on T). Similarly, in Example 2, there are two equilibrium outcomes: the dirac measure on T and the dirac measure on $(B,(U,L))$. Since we have shown that the former is not robust, the latter must be by the existence result of Theorem 1.

Before we move to the proof of Theorem 1, let us go back to Example 1 and give a short intuition. Under complete information, the normal-form representation of the game is given by

	T	B
T	$2,2$	$2,2$
B	$1,1$	$3,0$

(Table 1)

Now, if we write the (ex ante) normal-form representation of the ε-elaboration considered in the example, we get

	T	B
(T,T)	$2,2$	$2,2$
(T,B)	$2+\varepsilon, 2-\varepsilon$	$2+2\varepsilon, 2$
(B,T)	$1+\varepsilon, 1+\varepsilon$	$3-\varepsilon, 2\varepsilon$
(B,B)	$1+2\varepsilon, 1$	$3+\varepsilon, 2\varepsilon$

where $(s_1, s_1') \in \{T,B\} \times \{T,B\}$ denotes the strategy of Player 1 playing s_1 when his type is t_1^* and s_1' otherwise. Observe first that for $\varepsilon = 0$, this game is "equivalent" to the normal-form game given in Table 1 above, in the sense that by identifying payoff-equivalent strategies (T,T) and (T,B), and (B,T) and (B,B), we get exactly the game given in Table 1 (up to payoff-irrelevant names of strategies). Hence, for $\varepsilon > 0$, we can see the ε-elaboration as a normal-form game where we first add some redundant pure strategies and then slightly perturb the payoffs. This makes a link between our notion of elaborations that perturbs players' information structure as well as payoffs on terminal histories and a notion of perturbations in the refinement literature that perturb payoffs in normal form. In particular, Kohlberg and Mertens (1986) say that a set E of equilibria of a normal-form game is hyperstable if for any equivalent normal-form game and any small perturbation of payoffs of this equivalent game, there is an equilibrium close to E. They showed that any game has a component (maximal connected and closed set) of equilibria that is hyperstable. For instance, in the game given in Table 1, the set of mixed strategy profiles putting probability one on T for Player 1 and probability p in between $1/2$ and 1 on T for Player 2 is a hyperstable component. As one can see here, all strategy profiles in this component yield the same outcome T (i.e., the dirac measure on the terminal history where Player 1 plays T). Since, as we already explained, an ε-elaboration can be seen as a particular case of a payoff perturbation of an equivalent game, in an ε-elaboration (for ε small), there must exist an equilibrium that is close to this hyperstable component and so which yields an outcome close to T. Since this reasoning holds for any ε-elaboration, the outcome T is actually a robust outcome.

[9] The normal-form representation of Example 1 also serves as a non-generic counterexample for Theorem 1.

In the next section, we provide the proof of Theorem 1, which generalizes the above argument to generic extensive-form games.

3.5. **Proof of Theorem 1.** To prove Theorem 1, we will use the notion of hyperstability introduced by Kohlberg and Mertens (1986), or to be more precise, its strengthening proposed by Govindan and Wilson (2005), i.e., the notion of uniform hyperstability.[10] Let us recall its definition.

Following Kohlberg and Mertens (1986), we define an equivalence relation between games. First, say that, in a given normal-form game, two strategies of one player are equivalent if for every profile of other players' strategies they yield the same payoff. A pure strategy of a player is redundant if that player has another pure or mixed strategy that is equivalent. From a normal-form game G, one obtains its reduced form G_* by deleting redundant pure strategies until none remain. The reduced form is unique apart from the payoff-irrelevant names of the remaining pure strategies. Two normal-form games G and G' are equivalent if their reduced forms are the same, i.e., $G_* = G'_*$ (up to payoff-irrelevant names of strategies).

For two normal-form games $G = (N, (A_i)_{i\in N}, (v_i)_{i\in N})$ and $G' = (N, (A_i)_{i\in N}, (v'_i)_{i\in N})$ with the same sets of players and of each player's actions, we define $\|G - G'\| = \max_{i\in N, a\in A} |v_i(a) - v'_i(a)|$.

Now, let us present the notion of uniform hyperstability—which for short we will just call hyperstability.

Definition 3 (Kohlberg and Mertens (1986), Govindan and Wilson (2005))**.** *A subset E of equilibria of a normal-form game G is hyperstable if for every $\delta > 0$, there is $\varepsilon > 0$ such that for every equivalent game G', any G'' satisfying*

$$\left\|G'' - G'\right\| \leq \varepsilon$$

has an equilibrium α'' that is δ-close to some equilibrium $\alpha \in E$.[11]

To contrast the notion of hyperstability with our notion of robustness, note that hyperstability perturbs payoffs over strategy profiles while robustness perturbs payoffs on terminal histories. Hence, hyperstability does not take into account any extensive-form structure: if we see G as the normal-form representation of an extensive-form game $(\Gamma, (g_i)_{i\in N})$, then the notion of hyperstability allows for perturbed games (say with no addition of equivalent strategies) G'' that do not correspond to the normal-form representation of any extensive-form game $(\tilde{\Gamma}, (\tilde{g}_i)_{i\in N})$ satisfying $\tilde{\Gamma} = \Gamma$. Hence, while our notion of "nearby games" respects the extensive-form structure of $(\Gamma, (g_i)_{i\in N})$, i.e., the set of histories of an elaboration is just a duplication of the set of histories in Γ, the notion of perturbation in consideration for hyperstability does not respect any extensive-form structure. In this sense, we impose more discipline by respecting the extensive-form structure of $(\Gamma, (g_i)_{i\in N})$. On the other hand, the notion of hyperstability only allows for perturbations of payoffs and does not perturb the information structure of players. In this sense, our perturbations are wider than

[10] We also follow Govindan and Wilson (2005) and leave aside the minimality condition required in Kohlberg and Mertens (1986).

[11] Strictly speaking, α and α'' lie in different strategy spaces. We define the (pseudo)distance between α and α'' after mapping them to strategy profiles in the reduced form G_*.

just payoff perturbations. So overall, it does not seem obvious a priori how to compare hyperstability and robustness. However, Proposition 1 below gives a precise sense in which a "hyperstable outcome" is robust.

We also recall some basic results that will be used in our proof. An equilibrium component, or simply a component, is a maximal connected and closed set of Nash equilibria. Every finite normal-form game has finitely many components, at least one of which is hyperstable (Kohlberg and Mertens (1986, Proposition 1)). If $(g_i)_{i\in N} \in \mathcal{G}$, then since a strategy profile induces an outcome continuously, and every continuous function from a connected space to a discrete space is constant, the equilibrium outcome is constant over each component.

We are now in a position to prove our theorem. Consider a complete-information extensive-form game $(\Gamma, (g_i)_{i\in N})$ such that $(g_i)_{i\in N} \in \mathcal{G}$. Let $G^{\Gamma,g} = (N, (S_i)_{i\in N}, (g_i^{\mathcal{N}})_{i\in N})$ be its normal-form representation. Let C be a hyperstable component of $G^{\Gamma,g}$. Let $\mu \in \Delta(Z)$ be the unique outcome associated with C, i.e., for any equilibrium $\sigma \in C$, $z(\sigma) = \mu$. We will show that μ is indeed a robust outcome. Hence, Theorem 1 is a corollary of Proposition 1 below.

Proposition 1. *Fix a complete-information extensive-form game $(\Gamma, (g_i)_{i\in N})$. If C is a hyperstable component of $G^{\Gamma,g}$ and $\mu \in \Delta(Z)$ is the unique outcome associated with C, then μ is robust.*[12]

Proof. Fix $\delta > 0$. Let $\varepsilon > 0$ be as in the definition of hyperstability. Fix any $\tilde{\varepsilon}$-elaboration $U = (\Gamma, P, T, (u_i)_{i\in N})$ of $(\Gamma, (g_i)_{i\in N})$ and let $G^{\Gamma,u} = (N, (S_i')_{i\in N}, (u_i^{\mathcal{N}})_{i\in N})$ be its normal-form representation. Denoting S_i for the set of player i's pure strategies in $(\Gamma, (g_i)_{i\in N})$, we have that $S_i' = S_i^{T_i}$ and $u_i^{\mathcal{N}}(s') = \sum_{t\in T} P(t)u_i(z(s'(t)), t)$ for any $s' \in S'$. The positive number $\tilde{\varepsilon}$ is yet to be specified.

We first build a game G' equivalent to $G^{\Gamma,g} = (N, (S_i)_{i\in N}, (g_i^{\mathcal{N}})_{i\in N})$, the normal-form representation of $(\Gamma, (g_i)_{i\in N})$. In order to do so, consider the normal-form game $G' = (N, (S_i')_{i\in N}, (v_i')_{i\in N})$ where for any $s' \in S'$,

$$v_i'(s') = g_i^{\mathcal{N}}(s'(t^*)).$$

Clearly, the game G' is equivalent to $G^{\Gamma,g}$.

Now, we show that $G^{\Gamma,u}$ is in a neighborhood of the game G' (note that they both have the same set of strategies). Indeed,

$$\begin{aligned} u_i^{\mathcal{N}}(s') &= \sum_{t\in T} P(t)u_i(z(s'(t)), t) \\ &= P(t^*)g_i(z(s'(t^*))) + \sum_{t\neq t^*} P(t)u_i(z(s'(t)), t) \\ &= P(t^*)v_i'(s') + \sum_{t\neq t^*} P(t)u_i(z(s'(t)), t). \end{aligned}$$

[12] Govindan and Wilson (2005) show that an equilibrium component is hyperstable if and only if its index is nonzero (Dold (1972), Ritzberger (1994)). Combined with Proposition 1, this provides a sufficient condition for an outcome to be robust in a generic extensive-form game. That is, in a generic extensive-form game $(\Gamma, (g_i)_{i\in N})$, if an equilibrium component C of $G^{\Gamma,g}$ has a nonzero index, then the unique outcome $\mu \in \Delta(Z)$ associated with C is robust.

Since payoffs are uniformly bounded over elaborations, the above term tends to $v_i'(s')$ as $\tilde{\varepsilon}$ tends to 0. Hence, for $\tilde{\varepsilon}$ small enough, $G^{\Gamma,u}$ is indeed in the ε-neighborhood of G'.

Now, by the definition of hyperstability, $G^{\Gamma,u}$ has an equilibrium σ'' that is δ-close to some equilibrium in C. Hence, by the continuity of z, and assuming here again that $\tilde{\varepsilon}$ is small enough, $z(\sigma''(t^*))$ is δ-close to μ. Hence, the $\tilde{\varepsilon}$-elaboration U has a Bayesian Nash equilibrium σ'' satisfying $\|\mu - z(\sigma''(t^*))\| < \delta$. ∎

Remark 1. Hillas (1990, Theorem 4) shows a "continuity property" for various notions of stability. In particular, the continuity property of hyperstability implies that in the proof of Proposition 1, $G^{\Gamma,u}$ has a hyperstable component $C^{\Gamma,u}$ contained in a neighborhood of C. Therefore, in the above proof, we can choose any equilibrium in $C^{\Gamma,u}$ as σ''. Since $C^{\Gamma,u}$ contains a proper equilibrium of $G^{\Gamma,u}$ (Kohlberg and Mertens (1986, Proposition 5)), and hence a sequential equilibrium of U (van Damme (1984), Kohlberg and Mertens (1986, Proposition 0)), this shows that σ'' can be a sequential equilibrium. Hence, our generic existence result would hold if we were to replace the notion of Bayesian Nash equilibrium in Definition 2 by sequential equilibrium or any solution concept with the existence property that is coarser than properness.

Remark 2. Proposition 1 shows that hyperstability is sufficient for robustness. Beyond the class of generic perfect-information games, whether hyperstability is necessary for robustness remains open to us.[13,14]

Remark 3. The proof of our generic existence result, which relies on hyperstability, is similar in spirit to that in Güth, Kirchsteiger, and Ritzberger (1998), which in turn relies on the notion of essentiality. Namely, Güth, Kirchsteiger, and Ritzberger (1998) show that the outcome of an essential set is robust to noisy monitoring, which they call "accessible." We show in Appendix A.3 that there is an essential outcome that is not robust in our sense, implying that our notion of robustness is different from their notion of accessibility.

4. Necessary Conditions

This section presents two necessary conditions for robust equilibrium outcomes: stable sets and proper equilibria.

[13] As shown in the proof of Proposition 1, our notion of elaboration can be seen as a perturbation allowed by hyperstability. However, the latter class of perturbations is larger since it does not require the extensive-form structure to be respected. This may be a sign that hyperstability is not necessary for robustness for some games. Finding such a game is not an easy task though since we lack alternative sufficient conditions (other than hyperstability) to show robustness. Of course, one could weaken the definition of hyperstability by restricting attention to equivalent games G' that duplicate only pure strategies in G. Let us call this weaker notion semi-hyperstability. From the proof of Proposition 1, it is clear that every semi-hyperstable outcome is robust. However, whether semi-hyperstability is strictly weaker than hyperstability remains open to us.

[14] As we will see in the next section, other strategic stability notions that have appeared in the literature, known to be strictly weaker than hyperstability, are not sufficient for robustness. For example, the outcome T in Example 3 in Section 4.1 is fully stable in the sense of Kohlberg and Mertens (1986), but not robust; the outcome Out in the example of Hauk and Hurkens (2002) in Appendix A.3 is essential, but not robust; the outcome T in the example of van Damme (1989) in Online Appendix C.2 contains two stable sets in the sense of Mertens (1989) (as shown by Govindan and Wilson (2001)), but is not robust.

4.1. **Relationship to Stable Sets.** Let us recall the definition of a stable outcome in the sense of Kohlberg and Mertens (1986).[15]

Definition 4 (Kohlberg and Mertens (1986))**.** *Let $(\Gamma, (g_i)_{i\in N})$ be an extensive-form game. An outcome μ is stable if for any $\delta > 0$, there is $\bar{\varepsilon} > 0$ such that for any completely mixed strategy profile $\tilde{\sigma}$ and for any $\boldsymbol{\varepsilon} := (\varepsilon_i)_{i\in N}$ such that $\varepsilon_i < \bar{\varepsilon}$ for any $i \in N$, the normal-form game $G^{\boldsymbol{\varepsilon}} = (N, (S_i)_{i\in N}, (v_i^{\boldsymbol{\varepsilon}})_{i\in N})$ where each player i's payoffs are given by $v_i^{\boldsymbol{\varepsilon}}(s) = g_i^{\mathcal{N}}(((1-\varepsilon_j)s_j + \varepsilon_j\tilde{\sigma}_j)_{j\in N})$ has an equilibrium σ satisfying $\|\mu - z(\sigma)\| < \delta$.*

The following proposition states our first necessary condition for robustness.

Proposition 2. *If μ is robust, then μ is stable.*

We prove Proposition 2 in two steps: robustness implies robustness to canonical elaborations, and robustness to canonical elaborations implies stability. The notion of canonical elaborations is defined as follows, where we say that a type t_i is committed to strategy s_i if type t_i plays for sure action s_i.

Definition 5. *Fix $\varepsilon \geq 0$. $U = (\Gamma, P, T, (u_i)_{i\in N})$ is a canonical ε-elaboration of $(\Gamma, (g_i)_{i\in N})$ if $T_i = \{t_i^*\} \cup \{t_i^{s_i} : s_i \in S_i\}$ for each i, where each type $t_i^{s_i}$ is committed to strategy s_i, $P(t^*) \geq 1-\varepsilon$, and $u_i(\cdot, t^*) = g_i(\cdot)$ for all $i \in N$.*

Definition 6. *A canonical elaboration $U = (\Gamma, P, T, (u_i)_{i\in N})$ of $(\Gamma, (g_i)_{i\in N})$ has independent types and known own payoffs if there exists a profile $(P_i)_{i\in N}$ of distributions such that $P(t) = \prod_{i\in N} P_i(t_i)$ for any $t \in T$, and $u_i(\cdot, t_i^*, t_{-i}) = g_i(\cdot)$ for any $i \in N$ and $t_{-i} \in T_{-i}$.*

We define the notion of robustness to canonical elaborations (canonical elaborations with independent types and known own payoffs, resp.) by replacing ε-elaborations in Definition 2 by canonical ε-elaborations (canonical ε-elaborations with independent types and known own payoffs, resp.).

Proposition 2 can be decomposed into the following two parts.

Lemma 1. (1) *If μ is robust, then μ is robust to canonical elaborations.*
(2) *Outcome μ is stable if and only if μ is robust to canonical elaborations with independent types and known own payoffs.*[16]

The proof of Lemma 1 is given in Appendices A.1 and A.2. Note that Part 1 of Lemma 1 is non-trivial. As we mentioned in the previous section, our definition of elaborations respects the extensive-form structure of the complete-information extensive-form game. Typically, an elaboration "duplicates" the original extensive game form and changes payoffs on terminal histories. Hence, in non-trivial extensive-form games, we do not allow elaborations where players have strictly

[15] Kohlberg and Mertens (1986) define stability on sets of equilibria while we use a definition based on outcomes. An outcome is stable if it is induced by each equilibrium of a stable set in the sense of Kohlberg and Mertens (1986).
[16] For this equivalence result, we use the assumption that $u_i(\cdot, t_i^*, t_{-i}) = g_i(\cdot)$ holds exactly.

dominant strategies. Thus the challenge in the proof of Part 1 of Lemma 1 is to construct an elaboration such that for each strategy s_i, type $t_i^{s_i}$ plays a strategy that is outcome-equivalent to s_i in any equilibrium.

Note that the converse of Proposition 2 fails, i.e., stability is not a sufficient condition for robustness.[17] In fact, in Appendix A.3, we provide a counterexample for the converse of Part 1 of Lemma 1, illustrating the difference between robustness and robustness to canonical elaborations. Moreover, as the next example shows, a stable outcome may not be robust to canonical elaborations with correlated types, thereby demonstrating that the restriction to independent types is indispensable for Part 2 of Lemma 1.[18]

Example 3. Consider a modification of Example 2 in Section 2, where there are now three players: Player 1 plays first, and then Players 2 and 3 play the battle-of-sexes (hence, in the second stage Player 1 is now replaced by new Player 3). As for payoffs, Players 1 and 2 keep the same payoffs as in Example 2 while Player 3 is a "duplication" of Player 1 in the sense for each terminal history, his payoffs are the same as those of Player 1. It is easy to show that the subgame-perfect equilibrium outcome where Player 1 plays T is stable. Now, consider a canonical elaboration where, as in Example 2, Player 1 has two types and 2 has a single type. Further assume that new Player 3 has two types that are perfectly correlated to Player 1's type in the sense that Player 3 is a normal type if and only if Player 1 is a normal type. It is clear that in this context, we can mimic our argument in Section 2 to show that the subgame-perfect equilibrium outcome where Player 1 plays T is not robust to canonical elaborations.[19]

Remark 4. Under Kajii and Morris' (1997a) notion of robustness, Ui (2001) poses a question of whether robustness to canonical elaborations is strictly weaker than robustness to all elaborations. Recently, Pram (2019) shows the equivalence of the two robustness notions if agent-normal-form correlated equilibrium is used as a solution concept for (canonical) elaborations. On the other hand, under the solution concept of Bayesian Nash equilibrium, Takahashi (2019) shows, by means of an example, the non-equivalence of the set-valued versions of the respective robustness notions.

Remark 5. Note that in Example 3, the outcome T is stable, and hence induced by an equilibrium that survives iterative deletion of weakly dominated strategies (Kohlberg and Mertens (1986, Proposition 6)). In this respect, this example is more powerful than Example 2 in which the robustness test eliminates only equilibria that do not survive iterative deletion of weakly dominated strategies.

Remark 6. Generalizing Example 3, we can show that any elaboration of an extensive-form game can be replicated by an elaboration of the "agent-extensive" form, where all information sets are assigned to different players. Therefore, if μ is robust in the agent-extensive form, then it is robust in the original extensive-form game. This result resembles (the converse of) Kohlberg and Mertens

[17] This is consistent with Proposition 1 because stability is a strictly weaker set-valued refinement than hyperstability.
[18] Also, a stable outcome may not be robust to canonical elaborations if players do not know their own payoffs, i.e., $u_i(\cdot, t_i^*, t_{-i}) \neq g_i(\cdot)$ for some $t_{-i} \neq t_{-i}^*$. A concrete example is provided in Online Appendix C.2.
[19] Thus, by Proposition 1, this also shows that this outcome is not hyperstable.

(1986, Proposition 4), which shows that if a set is fully stable in the normal form of an extensive-form game, then it is fully stable in the agent-normal form.

4.2. **Relationship to Proper Equilibria.** Another necessary condition for robustness is properness. Thus, even if we have only assumed Nash equilibrium as a solution concept, our robustness requirement implies that a robust outcome must be consistent with subgame perfection.

Definition 7 (Myerson (1978))**.** *An ε-proper equilibrium of a normal-form game $G = (N, (A_i)_{i\in N}, (v_i)_{i\in N})$ is a totally mixed strategy profile α^ε such that if $v_i(a_i, \alpha^\varepsilon_{-i}) < v_i(a'_i, \alpha^\varepsilon_{-i})$, then $\alpha^\varepsilon_i(a_i) \leq \varepsilon\alpha^\varepsilon_i(a'_i)$. A proper equilibrium is any limit of ε-proper equilibria as ε goes to 0. In an extensive-form game $(\Gamma, (g_i)_{i\in N})$, if σ is a proper equilibrium of the normal-form representation of $(\Gamma, (g_i)_{i\in N})$, then we say that $\mu := z(\sigma)$ is induced by the proper equilibrium.*

The following proposition states our second necessary condition for robustness. The proof is given in Appendix A.4.

Proposition 3. *If μ is robust, then μ is induced by some proper equilibrium.*

Similarly to Proposition 2, the converse of Proposition 3 does not hold. That is, a proper equilibrium may induce a non-robust outcome. For example, in Example 2 in Section 2, outcome T is induced by a proper equilibrium (any equilibrium where Player 1 plays T and Player 2 plays R; regardless of Player 1's off-path behavior), but it is not robust.

Combining Theorem 1 and Proposition 3, we can characterize robustness in perfect-information extensive-form games with distinct payoffs on all terminal histories.

Corollary 1. *If $(\Gamma, (g_i)_{i\in N})$ is a perfect-information extensive-form game such that $g_i(z) \neq g_i(z')$ whenever $z \neq z'$, then the backward-induction outcome is a unique robust equilibrium outcome.*

Proof. A simple backward-induction argument shows that $(\Gamma, (g_i)_{i\in N})$ has finitely many Nash equilibrium outcomes.[20] Thus, by Theorem 1, at least one of the equilibrium outcomes is robust. By Proposition 3, any robust equilibrium outcome is induced by some proper equilibrium, and hence by a sequential equilibrium of $(\Gamma, (g_i)_{i\in N})$ (van Damme (1984), Kohlberg and Mertens (1986, Proposition 0)). Therefore, a robust equilibrium outcome is unique and given by backward induction. ∎

5. Robustness of Subgame-Perfect Implementation

So far we have considered incomplete-information perturbations of a benchmark game which is of complete information. In many applications, complete-information is a common assumption. This is true, in particular, in the implementation literature (e.g., Maskin (1999), Moore and Repullo (1988), Palfrey and Srivastva (1991)...). Recently, several results (Chung and Ely (2003) and AFHKT) casted doubts on the robustness of implementation results obtained in complete-information environments. In this section, we consider the implication of our results for the literature on the robustness of subgame-perfect implementation (AFHKT).

[20] For a proof, see Govindan and McLennan (2001, Section 1.1).

5.1. **The Hart-Moore Example of the Moore-Repullo Mechanism.** Let us consider the leading example in AFHKT initially due to Hart and Moore (2003). This example captures the logic behind Moore and Repullo's (1988) mechanism yielding permissive results for subgame-perfect implementation. A B(uyer) and a S(eller) are willing to trade an indivisible good. When trade occurs, B's utility is $\theta - p$, where p is the price and θ is the item's quality. In turn, the utility of the seller S is simply p. Hence, we implicitly normalize the cost of producing the item to 0. The quality level of the item can take two levels. When it is of high quality, the buyer B values it at $\theta_H = 14$. If it is of low quality, then B values it at $\theta_L = 10$.

Ideally, we would like to have a (sequential) mechanism that ensures full surplus extraction at the unique subgame-perfect equilibrium, i.e., under which the item is always traded, and the buyer B pays the true θ to the seller S. This can be done assuming that the state is commonly known. Indeed, the following mechanism ensures that, whenever a state $\theta \in \{\theta_L, \theta_H\}$ is commonly known, the induced game has a unique subgame-perfect equilibrium under which agents report truthfully their preferences and all B's surplus is extracted.

The mechanism is given as follows:

(1) B announces either a "high"or "low"quality. If B announces "high,"then B pays S a price equal to 14, and the game stops.
(2) If B announces "low" and S does not "challenge" B's announcement, then B pays a price equal to 10, and the game stops.
(3) If S challenges B's announcement, then:
   (a) B pays a fine F to a third party;
   (b) B is offered the good at price 6;
   (c) if B accepts the good, then S receives F from the third party (and also a payment of 6 from B), and the game stops;
   (d) if B rejects at 3b, then S pays F to the third party;
   (e) B and S each get the item with probability 1/2.

When $F > 9$, it is easy to check that, irrespective of the state θ, the game induced by this mechanism has a unique subgame-perfect equilibrium where each player reports truthfully his preferences. Indeed, let us first consider the case where $\theta = \theta_H$. We go through a simple backward induction argument: at stage 3, B has an incentive to accept the offer at price 6 (since, by rejecting, he will end up at stage 3e and get $14/2 - F = 7 - F$, but since the good is worth 14 he gets $14-6-F = 8-F$ by accepting). Now, at stage 2, S will not be fined if she challenges so given that $F > 9$, she will challenge (she receives $F+6 > 15$ in case she challenges and 10 otherwise). Finally, at Stage 1, if B tells the truth and claims the state is "high," he gets the item at price 14, which brings him a utility of 0, while if he lies, Stage 3 is reached and B is fined F. With $F > 9$, B is not willing to lie (he gets $14 - 6 - F < 0$). Similarly, if the state $\theta = \theta_L$. B is not willing to accept the offer at stage 3 (accepting yields $10 - 6 - F$ while by rejecting the offer he gets $10/2 - F = 5 - F$). Given this, S will be fined if she challenges at Stage 2. She will get $5 - F$ if she challenges and 10, otherwise. So S will not challenge at Stage 2. In turn, B is willing to claim that the state is "low" in which case he gets the item at price 10 (while he gets it at price 14 if he lies).

Note that the above game can be seen as a "degenerate" incomplete-information game where the set of types is $\{t_i^L, t_i^H\}$ for each player $i \in \{B, S\}$, the utility function of an agent i at each state θ are as described above, and the prior is given by:

$P:$		t_B^H, t_S^H	t_B^H, t_S^L	t_B^L, t_S^H	t_B^L, t_S^L
	θ_H	$1-\alpha$	0	0	0
	θ_L	0	0	0	α

where α is arbitrary in $(0, 1)$. Here, for the profile of types having strictly positive probability (t_B^H, t_S^H) $(t_B^L, t_S^L$, resp.), whenever it realizes, the state θ_H (θ_L, resp.) is commonly known among the agents. In the sequel, we will let U be this degenerate incomplete information extensive-form games which can be thought of as representing a family of complete-information extensive-form games (one for each $\theta \in \{\theta_L, \theta_H\}$).

There are two nice properties of this mechanism. First, it yields unique implementation in subgame-perfect equilibrium, i.e., for any state of nature, there is a unique subgame-perfect equilibrium, which yields the right outcome (i.e., full surplus extraction). Second, in each state, the unique subgame-perfect equilibrium reports the state truthfully. AFHKT show that one can perturb the above prior in such a way that (1) an undesirable sequential equilibrium arises, and (2) no sequential equilibrium has both players telling the truth with high probability. While these results show a lack of robustness of the implementation result obtained by mechanisms *à la* Moore and Repullo, our main result in this paper suggests that the outcome of the unique subgame-perfect equilibrium should remain in the perturbation. Put in another way, a "good" sequential equilibrium should still exist in the nearby situation. We first recall the argument in AFHKT and then state and discuss the implication of our result in this context.

Consider prior P^ε that satisfies $\|P^\varepsilon - P\| \leq \varepsilon$ as follows:

$P^\varepsilon:$		t_B^H, t_S^H	t_B^H, t_S^L	t_B^L, t_S^H	t_B^L, t_S^L
	θ_H	$(1-\alpha)\left(1-\varepsilon-\varepsilon^2\right)$	$(1-\alpha)\varepsilon$	$(1-\alpha)\varepsilon^2/2$	$(1-\alpha)\varepsilon^2/2$
	θ_L	$\alpha\varepsilon^2/2$	$\alpha\varepsilon^2/2$	$\alpha\varepsilon$	$\alpha\left(1-\varepsilon-\varepsilon^2\right)$

Note that B's type becomes infinitely more accurate than S's type as $\varepsilon \to 0$ in the sense that conditional on observing the two types (t_B^H, t_S^L) $((t_B^L, t_S^H)$, resp.) of players, the probability over the set of states converges to the dirac measure on θ_H (θ_L, resp.). This special feature implies that if S and B were informed of both signals, and the signals disagree, they will conclude that with high probability the state corresponds to B's signal.

We have an incomplete-information extensive-form games where the set of types is $\{t_i^L, t_i^H\}$ for each player $i \in \{B, S\}$, the utility function of an agent i at each state θ are as described above, and the prior is given by P^ε. We denote this game by U^ε. In a natural sense, U^ε perturbs U and the perturbation becomes small as ε vanishes. We recall AFHKT's result as follows.

Proposition 4 (AFHKT, Proposition 1)**.** *Consider the collection of incomplete-information extensive-form games $\{U^\varepsilon\}$. For any collection of strategy profiles $\{\sigma^\varepsilon\}$ where σ^ε is an equilibrium of U^ε, the probability that both players report their preferences sincerely does not go to 1 as ε goes to 0.*[21]

[21] We say that i reports his preferences sincerely if he reports θ_L (θ_H) when his type is t_i^L (t_i^H).

The argument for the above proposition is simple. Indeed, let us proceed by contradiction, and consider a sequence $\{\sigma^\varepsilon\}$ of equilibria of $\{U^\varepsilon\}$ under which the probability that both players report their preferences sincerely goes to 1 as ε vanishes. First, note that for $\varepsilon > 0$ small enough, B must be playing in pure strategies under σ^ε. This implies that B's observed action is fully revealing. Hence, when S has an opportunity to make a move, S must believe with probability 1 that B's type is t_B^L. Given the prior P^ε and that $\varepsilon > 0$ is small, S must also assign a large probability that $\theta = \theta_L$ when getting an opportunity to move (B's type is much more informative than S's type). Now, assume that $t_B = t_B^H$. If B sticks to the equilibrium strategy and announces "high", he gets the item at price 14. If, on the contrary, B deviates, because S believes that $\theta = \theta_L$ is very likely and therefore, S does not challenge and B gets the item at price 10. Thus, B is better-off by misreporting his preferences, a contradiction with our assumption that $\{\sigma^\varepsilon\}$ is a sequence of equilibria.

While the equilibrium behavior of agents cannot be close to truthtelling, our results (e.g., Corollary 1) suggest that the equilibrium outcome should be preserved as ε vanishes. While, strictly speaking, the complete-information benchmark here refers to a family of complete-information extensive-form games (one for θ_L and one for θ_H), as shown in Appendix B, one can easily get similar results in this context. In particular, we can apply a logic similar to that yielding Corollary 1 in order to get the following:

Proposition 5. *For any $\delta > 0$, there is $\varepsilon > 0$ such that the incomplete-information extensive-form games U^ε has an equilibrium such that all B's surplus is extracted with probability at least $1 - \delta$.*[22] *Further, the result remains true for any arbitrary prior P^ε as long as $\|P^\varepsilon - P\| \leq \varepsilon$.*[23]

Proof. This is implied by Corollary 2 below. ∎

In the above, for each state $\theta \in \{\theta_L, \theta_H\}$, under complete information, each player has distinct payoffs on all terminal histories. As will become clear (see, for instance, Corollary 4), this will be enough to apply an analogue of Corollary 1 to the current context.[24]

To recap, even though, as shown in AFHKT, the equilibrium behavior (which reports the state truthfully) is not robust to incomplete information, we show that the equilibrium outcome (full surplus extraction) is robust. Of course, as shown in AFHKT, we do have nearby priors under which we get undesirable equilibrium outcomes (i.e., where a significant fraction of the surplus cannot be extracted). In that respect, our result completes the picture and shows that, at least in this

[22] More precisely, under (t_B^H, t_S^H), B pays the price 14 to S with probability at least $1 - \delta$. Similarly, under (t_B^L, t_S^L), B pays the price 10 to S with probability at least $1 - \delta$.

[23] Chen, Holden, Kunimoto, Sun, and Wilkening (2017) provide a two-stage mechanism which achieves full surplus extraction under complete information. They further prove that under their mechanism, any sequential equilibrium outcome in perturbed environments is close to full surplus extraction, provided that players know their own payoffs. Further, in such private value perturbations, all sequential equilibrium strategy profiles are close to the unique subgame perfect equilibrium strategy profile of the original complete information game. This result depends on the fact that the mechanism has only two stages and on the private value assumption. AFHKT and our results show that under more general perturbations, one can only have some equilibrium outcome close to full surplus extraction.

[24] We further note that AFHKT consider the case of $F = 9$. This case turns out to be non-generic in the sense that two different histories can yield the same terminal payoffs to a player. However, it is easy to check that the induced game for $F = 9$ has finitely many Nash equilibrium outcomes which is what is needed in the proof of Corollary 1.

example, full surplus extraction may still be achieved if agents coordinate on the right equilibrium. In the next section, we explain how our argument can be applied beyond this example.

5.2. **A General Result.** In this section, we show that the above result holds more generally in abstract implementation environments. In order to prove a general statement, we first need to adapt our notions to the implementation context—which we only did implicitly in the previous section.

We will restrict our attention to the following incomplete-information extensive-form game $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$ consisting of a game form Γ, a prior $P \in \Delta(\Theta \times T)$ over payoff-relevant states and types, where T_i is given by a copy of Θ, $T_i = \{t_i^\theta : \theta \in \Theta\}$, and terminal payoffs $g_i \colon Z \times \Theta \to \mathbf{R}$. Nature moves first and determines $(\theta, t) \in \Theta \times T$ according to P. Then each player i is informed of his own type t_i and plays a game form Γ with terminal payoffs $g_i(\cdot, \theta)$. We assume that Θ is finite.

We write $t^\theta = (t_i^\theta)_{i\in N}$. We say that a prior P is a complete-information prior if $P(\theta, t^\theta) > 0$ for any θ, and $P(\theta, t) = 0$ whenever $t \neq t^\theta$. This implies that whenever a profile of types t^θ realizes, θ is commonly known among players. In this case, we say that U is a family of complete-information extensive-form games. Given $\varepsilon \geq 0$, we say that $U' = (\Gamma, P', T, \Theta, (g_i)_{i\in N})$ is an ε-elaboration of the family of complete-information extensive-form games $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$ if $\|P' - P\| \leq \varepsilon$.[25]

Now, moving toward an implementation environment, we let $\mathcal{X}$ be a finite set of allocations and $f \colon \Theta \to \mathcal{X}$ be a social choice function.[26] Agents have utilities over allocations $\mathcal{X}$ given by $g_i^{\mathcal{X}} \colon \mathcal{X} \times \Theta \to \mathbb{R}$ for each player i. A mechanism is a tuple $\mathcal{M} = [\Gamma, \gamma]$ composed of a game form Γ and an allocation function $\gamma \colon Z \to \mathcal{X}$ describing which allocation in $\mathcal{X}$ is achieved after each terminal history in Z.[27] Given a complete-information prior P over $\Theta \times T$, a mechanism $\mathcal{M} = [\Gamma, \gamma]$ induces a family of complete-information extensive-form games $U = (\Gamma, P, T, \Theta, (g_i^{\mathcal{X}}(\gamma(\cdot), \cdot))_{i\in N})$.

Moore and Repullo (1988) and AFHKT considered the notion of full implementation in complete-information environments where the solution concept is subgame-perfect equilibrium.

Definition 8. *Fix a complete-information prior P. Social choice function f is fully implementable if there is a mechanism $\mathcal{M} = [\Gamma, \gamma]$ under which for any θ, any subgame-perfect equilibrium σ^θ of the complete-information extensive-form game $(\Gamma, (g_i^{\mathcal{X}}(\gamma(\cdot), \theta))_{i\in N})$ satisfies $\gamma(z(\sigma^\theta)) = f(\theta)$.*

An important part of the implementation literature focuses on pure equilibria (see for instance, the survey by Jackson (2001)).[28] We impose here the requirement that all subgame-perfect equilibria, including non-pure ones, yield a desirable outcome. As shown in the Online Appendix C.3, this stronger requirement is needed for the result below to hold.

[25] We note that an ε-elaboration perturbs the prior probability P but keeps Θ constant. For instance, in the case of a single complete information game such as the case studied in Section 3, i.e., where Θ is a singleton, any ε-elaboration trivially coincides with the complete-information game. Hence, the class of elaborations we look at does not generalize that introduced in Section 3.

[26] The set of allocations in the example of the previous section was infinite. However, the mechanism used there is finite meaning that the mechanism is only using a finite subset of the set of allocations. Hence, we could have very well started with this finite set of allocations.

[27] By definition, our mechanisms are finite, i.e., the set of terminal histories is finite.

[28] Exceptions are Mezzetti and Renou (2012), Kartik and Tercieux (2012), Serrano and Vohra (2009), and Maskin (1999).

As shown in Appendix B, the logic of our arguments yielding Theorem 1 can be easily applied to an environment like here where the benchmark is a family of complete-information extensive-form games.

Recall that $\mathcal{G}$ is the set of payoffs $(g_i)_{i\in N}$ over terminal histories such that $(\Gamma,(g_i)_{i\in N})$ has finitely many Nash equilibrium outcomes.

Theorem 2. *Fix a complete-information prior P. Assume that f is fully implementable by a mechanism $\mathcal{M}=[\Gamma,\gamma]$. If $(g_i^{\mathcal{X}}(\gamma(\cdot),\theta))_{i\in N}\in\mathcal{G}$ for any θ, then the following property holds: for any $\delta>0$, there is $\varepsilon>0$ such that any ε-elaboration of $U=(\Gamma,P,T,\Theta,(g_i^{\mathcal{X}}(\gamma(\cdot),\cdot))_{i\in N})$ has a Bayesian Nash equilibrium σ satisfying $\max_{\theta\in\Theta}\left\|f(\theta)-\gamma(z(\sigma(t^\theta)))\right\|<\delta$.*

Note that since $\mathcal{G}$ is generic in $\mathbb{R}^{Z\times N}$, if γ is injective, then the set of utilities $(g_i^{\mathcal{X}}(\cdot,\theta))_{i\in N}$ over alternatives such that $(g_i^{\mathcal{X}}(\gamma(\cdot),\theta))_{i\in N}\in\mathcal{G}$ is also generic in $\mathbb{R}^{\mathcal{X}\times N}$.

As in the proof of Corollary 1, if Γ is a perfect-information extensive game form, a simple backward-induction argument shows that if each player has distinct payoffs on all terminal histories, then U has finitely many Nash equilibrium outcomes. Thus, we get the following corollary that generalizes Proposition 5.

Corollary 2. *Assume that f is fully implementable by a mechanism $\mathcal{M}=[\Gamma,\gamma]$, where Γ is a perfect-information extensive form. If γ is injective, and $g_i^{\mathcal{X}}(x,\theta)\neq g_i^{\mathcal{X}}(x',\theta)$ whenever $x\neq x'$, then the following property holds: for any $\delta>0$, there is $\varepsilon>0$ such that any ε-elaboration of $U=(\Gamma,P,T,\Theta,(g_i^{\mathcal{X}}(\cdot,\gamma(\cdot)))_{i\in N})$ has a Bayesian Nash equilibrium σ satisfying $\max_{\theta\in\Theta}\left\|f(\theta)-\gamma(z(\sigma(t^\theta)))\right\|<\delta$.*

To conclude, fix a social choice function f which is not Maskin monotonic. Given a mechanism which achieves full implementation of f, AFHKT have shown that one can always construct an ε-elaboration (for arbitrary small $\varepsilon>0$) under which we get undesirable equilibrium outcomes (i.e., far from f). Again, our results, in particular, Theorem 2 only completes the picture in showing that in this ε-elaboration, there will also be a desirable equilibrium outcome. We are not claiming that this weakens the conclusion in AFHKT since full implementation does fail in nearby environments.

6. Discussion

6.1. **Multiple Normal Types.** Up to now, we have considered elaborations as defined in Fudenberg, Kreps, and Levine (1988). There are of course alternative definitions of elaborations and hence of robustness. For instance, the literature sometimes considers larger classes of elaborations under which there may exist multiple profiles of types with ex post payoffs that are the same as those of the complete-information game and that do not have vanishing probability in the limit (see for instance, Monderer and Samet (1989) or Kajii and Morris (1997b)).

To discuss this, we use a notion similar to the "robustness to limit-common-knowledge elaborations" used in Kajii and Morris (1997b). In the sequel, we fix a finite set T^*, which will correspond to the profiles of types $t^*\in T^*$ having the same ex post payoffs as under complete information. We also fix a prior distribution P^∞ in $\Delta(T^*)$ which represents the "limit distribution".

Definition 9. *Fix $\varepsilon \geq 0$. $U = (\Gamma, P, T, (u_i)_{i\in N})$ is an ε-elaboration of $(\Gamma, (g_i)_{i\in N})$ with multiple normal types T^* and limit distribution P^∞ if it is an elaboration of $(\Gamma, (g_i)_{i\in N})$, $T \supseteq T^*$, $\|P - P^\infty\| < \varepsilon$ and $u_i(\cdot, t^*) = g_i(\cdot)$ for all $i \in N$ and all $t^* \in T^*$.*[29]

While we are allowing for multiple normal types, the above notion still implies important restrictions on higher-order beliefs. For instance, with T^* fixed, in an ε-elaboration with multiple normal types, players have common p-belief about normal types with p close to 1 when ε is close to 0 (see Monderer and Samet (1989)).

Definition 10. *Fix a complete-information extensive-form game $(\Gamma, (g_i)_{i\in N})$. An equilibrium outcome $\mu \in \Delta(Z)$ is robust to limit-common-knowledge elaborations if for any $\delta > 0$ and any finite T^* and any $P^\infty \in \Delta(T^*)$, there is $\varepsilon > 0$ such that any ε-elaboration $U = (\Gamma, P, T, (u_i)_{i\in N})$ of $(\Gamma, (g_i)_{i\in N})$ with multiple normal types T^* and limit distribution P^∞ has a Bayesian Nash equilibrium σ satisfying $\left\|\mu - \sum_{t\in T} P(t) z(\sigma(t))\right\| < \delta$.*[30]

Recall that a probability distribution P is independent if $P(t) = \prod_{i\in N} P_i(t_i)$ for some collection of probability distributions $(P_i)_{i\in N}$ each over T_i. If in the above definition, we were to restrict our attention to independent limit distributions P^∞, then our main result (Theorem 1) would extend.

However, in general, our result does not extend. In order to show this, let us discuss the following example borrowed from Kajii and Morris (1997b).

Example 4. Consider the three-player two-action normal-form game where each player chooses L or R. Each player gets -1 if he matches the choice of the player preceding him in the cycle $1 \to 2 \to 3 \to 1$ and 1 otherwise:

	L	R
L	$-1,-1,-1$	$-1,1,1$
R	$1,1,-1$	$1,-1,1$

L

	L	R
L	$1,-1,1$	$1,1,-1$
R	$-1,1,1$	$-1,-1,-1$

R

In this game, there is a unique equilibrium, where each player mixes and puts probability a half on each action. Also, this equilibrium constitutes a hyperstable component as a singleton. However, as shown by Kajii and Morris (1997b), this equilibrium is not robust to limit-common-knowledge elaborations. Hence, our core intermediary step—Proposition 1—does not hold with this more demanding notion of robustness. For completeness, let us report briefly their argument. Consider the following sequence of elaborations where Player $i \in \{1, 2\}$ has two types t_i^* and t_i^{**} under which i's ex post payoffs are given by the above complete-information game. Player 3 has two types t_3^* and t_3^{crazy}. Under t_3^*, his ex-post payoffs are given by the above complete-information game while under t_3^{crazy}, we assume that Player 3 has a strictly dominant strategy to play action L. The prior probability assigns weight $2/3$ to (t_1^*, t_2^*, t_3^*), $1/3 - \varepsilon$ to $(t_1^{**}, t_2^{**}, t_3^*)$, and ε to $(t_1^*, t_2^*, t_3^{\text{crazy}})$. First, it should be clear that no type t_1^*, t_2^*, or t_3^* can play a pure strategy. Now,

[29] We abuse notations and write $\|P - P^\infty\|$ for $\max_{t\in T^*} |P(t) - P^\infty(t)|$.

[30] Proposition 6 below would remain true if we were to impose the stronger condition that $P(\{t \in T : \|\mu - z(\sigma(t))\| < \delta\}) > 1 - \delta$.

since t_2^* plays in mixed strategies, knows that his payoffs are given by the complete-information game, and assigns probability one to t_1^*, we must have $\sigma_1(t_1^*)(\mathrm{L}) = 1/2$. Hence, since t_1^* plays in mixed strategies and knows that his payoffs are given by the complete-information game, he must assign probability $1/2$ to Player 3 choosing action L. An easy computation shows that this yields $\sigma_3(t_3^*)(\mathrm{L}) = 1/2 - 3/4\varepsilon$. Now, given that $\sigma_3(t_3^*)(\mathrm{L}) < 1/2$ and that t_1^{**} assigns probability one to t_3^*, we must have $\sigma_1(t_1^{**})(\mathrm{L}) = 1$ and so a similar argument yields $\sigma_2(t_2^{**})(\mathrm{L}) = 0$. Finally, since t_3^* mixes and knows that his payoffs are given by the complete-information game, this type must assign probability $1/2$ to Player 2 playing L. Here again, an easy computation shows that this yields $\sigma_2(t_2^*)(\mathrm{L}) = 3/4 - 3/4\varepsilon$. This provides a full description of the unique equilibrium of this elaboration. As ε tends to 0, the ex ante distribution over outcomes induced by this equilibrium, i.e., $\sum_{t\in T} P(t)z(\sigma(t))$, tends to the outcome μ defined by $\mu(\mathrm{L},\mathrm{L},\mathrm{L}) = \mu(\mathrm{L},\mathrm{L},\mathrm{R}) = 1/8$, $\mu(\mathrm{L},\mathrm{R},\mathrm{L}) = \mu(\mathrm{L},\mathrm{R},\mathrm{R}) = 5/24$, $\mu(\mathrm{R},\mathrm{L},\mathrm{L}) = \mu(\mathrm{R},\mathrm{L},\mathrm{R}) = 1/8$, and $\mu(\mathrm{R},\mathrm{R},\mathrm{L}) = \mu(\mathrm{R},\mathrm{R},\mathrm{R}) = 1/24$. This is not the unique Nash equilibrium of the complete-information game, but a correlated equilibrium.

While our result does not hold in general, if we restrict our attention to perfect-information extensive-form games, Theorem 1 (in fact, Corollary 1) can be extended.

Proposition 6. *If $(\Gamma, (g_i)_{i\in N})$ is a perfect-information extensive-form game such that $g_i(z) \neq g_i(z')$ whenever $z \neq z'$, then the backward-induction outcome is a unique equilibrium robust to limit-common-knowledge elaborations.*

Proof. Denote by s the unique subgame-perfect equilibrium of $(\Gamma, (g_i)_{i\in N})$. Consider the limit 0-elaboration $U = (\Gamma, P^\infty, T^*, (u_i)_{i\in N})$ with multiple normal types T^* and prior distribution P^∞, and let $G^{\Gamma,u}$ be its normal-form representation. Without loss of generality, we assume that $\sum_{t^*_{-i}} P^\infty(t_i^*, t^*_{-i}) > 0$ for all $i \in N$ and $t_i^* \in T_i^*$. Since every player has distinct payoffs on all terminal histories in $(\Gamma, (g_i)_{i\in N})$, a simple backward-induction argument shows that the strategy profile σ^∞ with $\sigma_i^\infty(t_i^*) = s_i^*$ for all $i \in N$ and $t_i^* \in T_i^*$ is a unique sequential equilibrium of U (irrespective of the belief system of players). By Kohlberg and Mertens (1986, Proposition 1), $G^{\Gamma,u}$ has a hyperstable component. Any hyperstable component contains a proper equilibrium (Kohlberg and Mertens (1986, Proposition 5)), and hence σ^∞ (van Damme (1984), Kohlberg and Mertens (1986, Proposition 0)). Therefore, the equilibrium component of $G^{\Gamma,u}$ containing σ^∞ is hyperstable. Also, one can show that $G^{\Gamma,u}$ has finitely many Nash equilibrium outcomes. Indeed, given our assumption that there is no indifference over terminal payoffs, an inductive argument shows that on the equilibrium path players must play pure actions and the pure action chosen is type independent. Hence, there must be finitely many Nash outcomes. Therefore, the backward-induction outcome is hyperstable, and hence, by an argument similar to the proof of Proposition 1, robust to limit-common-knowledge elaborations.

The uniqueness in the statement of Proposition 6 simply comes from Corollary 1 and the observation that robustness to limit-common-knowledge elaborations is stronger than the robustness requirement used in Corollary 1. ∎

One source of difficulty when we move to the more stringent notion of robustness to limit-common-knowledge elaborations is that the limit of equilibria in elaborations with multiple normal

types may not be a Nash equilibrium of the complete-information game. In general, it is a correlated equilibrium. In this sense, the right concept in this context seems to be correlated equilibrium for the complete-information game and some Bayesian version of it for elaborations, such as agent-normal-form correlated equilibrium in Pram (2019). How to extend Theorem 1 to the more demanding robustness test defined in the present section while using correlated equilibrium as a solution concept is an interesting question left for further research.

6.2. **Varying Normal Types.** One can consider an even more stringent notion of robustness as follows:

Definition 11. *Fix a complete-information extensive-form game* $(\Gamma, (g_i)_{i\in N})$. *An equilibrium outcome* $\mu \in \Delta(Z)$ *is robust to all elaborations if for any* $\delta > 0$, *there is* $\varepsilon > 0$ *such that any* ε*-elaboration* $U = (\Gamma, P, T, (u_i)_{i\in N})$ *of* $(\Gamma, (g_i)_{i\in N})$ *with multiple normal types* T^* *and prior distribution* P^∞ *has a Bayesian Nash equilibrium* σ *satisfying* $\left\|\mu - \sum_{t\in T} P(t)z(\sigma(t))\right\| < \delta$.

The difference from Definition 10 is that ε must be chosen uniformly irrespective of (the cardinality of) T^*. This seemingly small difference has a large implication on robustness due to the fact that the players no longer have common p-belief about normal types with p close to 1 (Rubinstein (1989), Kajii and Morris (1997a, b)).

Strictly speaking, the robustness notion in Definition 11 is slightly stronger than that in Kajii and Morris (1997a), as they require each normal type *knows* his own payoffs, i.e., $u_i(\cdot, t_i^*, t_{-i}) = g_i(\cdot)$ for all $i \in I$, $t_i^* \in T_i^*$, and $t_{-i} \in T_{-i}$. Nonetheless, one of their sufficient conditions for robustness, namely their Proposition 3.2, extends to the robustness notion in Definition 11. That is, if the normal-form representation of an extensive-form game has a unique correlated equilibrium outcome, then it is robust to all elaborations. For example, it is easy to see in Example 1 in Section 2 that the set of correlated equilibria coincides with the set of Nash equilibria, each of which assigns probability 1 on Player 1 playing T. Therefore, the outcome associated with Player 1 playing T is robust to all elaborations. We do not know whether this result extends beyond Example 1, say, whether the backward-induction outcome of a generic perfect-information game is robust to all elaborations.

Appendix A. Proofs of Lemma 1 and Proposition 3 and an Example

A.1. **Proof of Part 1 of Lemma 1.** The proof of this part is close in spirit to the proof of Lemma 6 in Chen (2012). Let $(\Gamma, (g_i)_{i\in N})$ be a complete-information extensive-form game. In order to prove Part 1 of Part 1 of Lemma 1, we will need the following lemma.

Lemma 2. *Given any extensive-form game, for any player i and any pure behavioral strategy s_i, there exists $g_i^{s_i}\colon Z \to \mathbb{R}$ under which for any strategy s_i' and any s_{-i}, if $s_i'(h) \neq s_i(h)$ for some subhistory h of $z(s_i, s_{-i})$, then $g_i^{s_i}(z(s_i, s_{-i})) > g_i^{s_i}(z(s_i', s_{-i}))$.*

Proof. Fix i and strategy s_i. Define $g_i^{s_i}\colon Z \to \mathbb{R}$ as follows:

$$g_i^{s_i}(z) = \begin{cases} g_i(z) & \text{if } z = z(s_i, s_{-i}) \text{ for some } s_{-i}, \\ \min_z g_i(z) - 1 & \text{otherwise.} \end{cases}$$

Fix any s_i' and s_{-i}. Assume that $s_i'(h) \neq s_i(h)$ for some subhistory h of $z(s_i, s_{-i})$. This implies that $z(s_i', s_{-i}) \neq z(s_i, s_{-i}')$ for all s_{-i}'; indeed, if there were s_{-i}' satisfying $z(s_i', s_{-i}) = z(s_i, s_{-i}')$, then for any subhistory $h' \in H_i$ of $z(s_i', s_{-i})$, we would have $s_i'(h') = s_i(h')$, and thus $z(s_i', s_{-i}) = z(s_i, s_{-i})$, a contradiction with the assumption that $s_i'(h) \neq s_i(h)$ for some subhistory h of $z(s_i, s_{-i})$. Finally, because $z(s_i', s_{-i}) \neq z(s_i, s_{-i}')$ for all s_{-i}', by construction, we get $g_i^{s_i}(z(s_i, s_{-i})) = g_i(z(s_i, s_{-i})) > \min_z g_i(z) - 1 = g_i^{s_i}(z(s_i', s_{-i}))$ as claimed. ∎

Recall that μ is robust to incomplete information. Fix $\delta > 0$. Fix any canonical ε-elaboration $U^c = (\Gamma, P^c, T^c, (u_i^c)_{i\in N})$ with $T_i^c = \{t_i^*\} \cup \{t_i^{s_i} : s_i \in S_i\}$. We show that if ε is small enough, then there exists an equilibrium that induces an outcome δ-close to μ.

Construct the following elaboration $U = (\Gamma, P, T, (u_i)_{i\in N})$, where $T_i = T_i^c$ and

$$u_i(\cdot, t) = \begin{cases} u_i^c(\cdot, t_i^*, t_{-i}) & \text{if } t_i = t_i^*, \\ g_i^{s_i}(\cdot) & \text{if } t_i = t_i^{s_i}, \end{cases}$$

$$P(t) = \begin{cases} (1-\varepsilon)P^c(t) & \text{if } t_i = t_i^* \text{ for some } i, \\ (1-\varepsilon)P^c(t) + \varepsilon/|S| & \text{otherwise.} \end{cases}$$

Given that U is a 2ε-elaboration, we know that if ε is small enough, then there exists an equilibrium σ with $z(\sigma)$ being δ-close to μ. In the sequel, recall that a mixed strategy σ_i in the elaboration is identified with a behavioral strategy, and $\sigma_i(t_i)$ with a probability distribution over pure strategies. In particular, a pure strategy s_i has positive probability under $\sigma_i(t_i)$ if and only if type t_i plays $s_i(h)$ with positive probability for any history $h \in H_i$.

$H(s_i)$ denotes the set of all histories induced by s_i and some s_{-i}.

Lemma 3. *For any player i and any strategy s_i, if s_i' has positive probability under $\sigma_i(t_i^{s_i})$, then for any history $h \in H_i \cap H(s_i)$, we have $h \in H(s_i')$ and $s_i'(h) = s_i(h)$.*

Proof. We prove this by induction on the size of histories. First, consider player i who moves at the initial history $\emptyset$. Then, since the initial history is a subhistory of any history, by Lemma 2, if pure

strategy s_i' satisfies $s_i'(\emptyset) \neq s_i(\emptyset)$, then we have $g_i^{s_i}(z(s_i, s_{-i})) > g_i^{s_i}(z(s_i', s_{-i}))$ for any s_{-i}. Hence, since $t_i^{s_i}$ knows that his payoffs are given by $g_i^{s_i}(\cdot)$, if $s_i'(\emptyset)$ receives strictly positive probability under $\sigma_i(t_i^{s_i})$, we must have $s_i(\emptyset) = s_i'(\emptyset)$.

Now, suppose that our claim holds for every player j and every history with length less than or equal to m where $m \geq 0$, i.e., (IH) for any j and s_j, if s_j' has positive probability under $\sigma_j(t_j^{s_j})$, then for any $h \in H_j \cap H(s_j)$ with length $\leq m$, we have $h \in H(s_j')$ and $s_j'(h) = s_j(h)$. Fix any history $h \in H_i \cap H(s_i)$ with length $m+1$. We show that s_i' cannot have positive probability under $\sigma_i(t_i^{s_i})$, either in the case where $h \notin H(s_i')$ or in the case where $s_i'(h) \neq s_i(h)$. First, if $h \notin H(s_i')$, then $s_i'(h) \neq s_i(h)$ for a proper subhistory $h' \in H_i \cap H(s_i)$ of h. Since the length of h' is smaller than m, it follows from (IH) that s_i' does not have positive probability under $\sigma_i(t_i^{s_i})$. Second, let us consider the case where $s_i'(h) \neq s_i(h)$.

Step 1. Under (IH), for any s_i, there exist s_{-i} and s'_{-i} such that $\sigma_{-i}(t_{-i}^{s'_{-i}})(s_{-i}) > 0$ and h is a subhistory of $z(s_i, s_{-i})$.

To show this, observe that since $h \in H_i \cap H(s_i)$, h is a subhistory of $z(s_i, s'_{-i})$ for some s'_{-i}. For each $j \neq i$, take any s_j that has positive probability under $\sigma_j(t_j^{s'_j})$. By (IH), $s_j(h') = s_j'(h')$ for any $h' \in H_j \cap H(s_j)$ with length $\leq m$. Hence, h is a subhistory of $z(s_i, s_{-i})$.

Step 2.

$$\sum_{t_{-i}} P(t_{-i} \mid t_i^{s_i}) \sum_{s_{-i}} \sigma_{-i}(t_{-i})(s_{-i}) \left[g_i^{s_i}(z(s_i, s_{-i})) - g_i^{s_i}(z(s_i', s_{-i})) \right] > 0. \tag{2}$$

Let us fix s_{-i} and consider two cases (i) if $s_i'(h') \neq s_i(h')$ for some subhistory h' of $z(s_i, s_{-i})$, then, by construction, $g_i^{s_i}(z(s_i, s_{-i})) > g_i^{s_i}(z(s_i', s_{-i}))$; (ii) if $s_i'(h') = s_i(h')$ for any subhistory h' of $z(s_i, s_{-i})$, then $z(s_i, s_{-i}) = z(s_i', s_{-i})$ and hence $g_i^{s_i}(z(s_i, s_{-i})) = g_i^{s_i}(z(s_i', s_{-i}))$. By Step 1 and $P(t_{-i}^{s'_{-i}} \mid t_i^{s_i}) > 0$ for any s'_{-i}, case (i) happens with positive probability. Therefore (2) holds. ∎

We get the following corollary.

Corollary 3. *For any player i and any strategy s_i, if s_i' has positive probability under $\sigma_i(t_i^{s_i})$, then s_i and s_i' are outcome-equivalent, i.e., $z(s_i, s_{-i}) = z(s_i', s_{-i})$ for all s_{-i}.*

Proof. Theorem 1 in Kuhn (1953) yields that for any s_i, s_i', if we have $h \in H(s_i')$ and $s_i'(h) = s_i(h)$ for any history $h \in H(s_i) \cap H_i$, then s_i and s_i' are outcome-equivalent. ∎

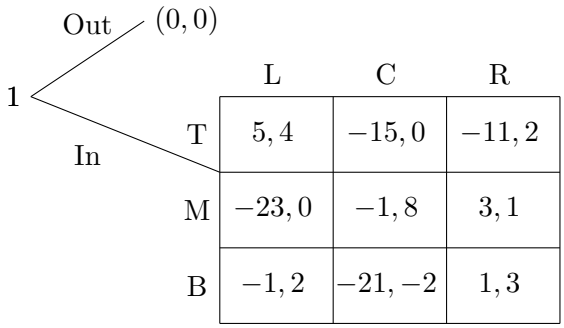
Now, build the modified strategy profile $\hat{\sigma}$ under which $\hat{\sigma}_i(t_i^*) = \sigma_i(t_i^*)$ while for any s_i, $\hat{\sigma}_i(t_i^{s_i}) = s_i$. By the above corollary, this is an equilibrium of the 2ε-elaboration U. Finally, observe that this is also an equilibrium of the canonical ε-elaboration U^c, where each $t_i^{s_i}$ is forced to play s_i, so the proof of Part 1 of Lemma 1 is completed.

A.2. **Proof of Part 2 of Lemma 1.** Assume that μ is robust to canonical elaborations with independent types and known own payoffs. Fix any $\delta > 0$. Fix any collection of $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i \in N}$ and any profile $(\tilde{\sigma}_i)_{i \in N}$ of completely mixed strategies. Build the following canonical elaboration.

Consider player i's type space $T_i = \{t_i^*\} \cup \{t_i^{s_i} : s_i \in S_i\}$, where type t_i^* knows that his own payoff is g_i, and type $t_i^{s_i}$ is the type of player i who is committed to strategy s_i. The prior probability P over T comes from the product of priors P_i over each T_i where $P_i(t_i^{s_i}) = \varepsilon_i \tilde{\sigma}_i(s_i)$ for all s_i. Given that this game is a canonical ε-elaboration with independent types and known own payoffs with $\varepsilon = 1 - \prod_{i \in N}(1 - \varepsilon_i)$, we know that when all ε_i are small enough, there must exist an equilibrium σ in U satisfying $\|\mu - z(\sigma(t^*))\| < \delta$. Also, $\sigma(t^*)$ is an equilibrium of the game where each player i's payoffs are given by $v_i^{\boldsymbol{\varepsilon}}(s) = g_i^{\mathcal{N}}(((1 - \varepsilon_j)s_j + \varepsilon_j \tilde{\sigma}_j)_{j \in N})$.

Conversely, assume that μ is stable. Fix any $\delta > 0$. Fix any canonical ε-elaboration collection $U = (\Gamma, P, T, (u_i)_{i \in N})$ with independent types and known own payoffs. Let $\varepsilon_i = 1 - P_i(t_i^*)$ for each $i \in N$, and if $\varepsilon_i > 0$, then let $\tilde{\sigma}_i(s_i) = P_i(t_i^{s_i})/\varepsilon_i$ for each $s_i \in S_i$; otherwise, let $\tilde{\sigma}_i$ be arbitrary. Given that $\varepsilon_i \leq \varepsilon$, when ε is small enough, there must exist an equilibrium σ in the game where each player i's payoffs are given by $v_i^{\boldsymbol{\varepsilon}}(s) = g_i^{\mathcal{N}}(((1-\varepsilon_j)s_j + \varepsilon_j \tilde{\sigma}_j)_{j \in N})$ satisfying $\|\mu - z(\sigma(t^*))\| \leq \delta$. (Use the upper hemicontinuity of equilibria if $\tilde{\sigma}$ is not completely mixed.) Also, σ is an equilibrium of U for normal types.

A.3. **A Counterexample for the Converse of Part 1 of Lemma 1.** Let us consider the following outside-option game proposed by Hauk and Hurkens (2002). There are two players $i = 1, 2$. At stage 1, Player 1 chooses between Out or In. If he chooses Out, the game ends. If he chooses In, we reach a second stage where players play a 3×3 normal-form game. Payoffs are given as follows:



	L	C	R
T	$5, 4$	$-15, 0$	$-11, 2$
M	$-23, 0$	$-1, 8$	$3, 1$
B	$-1, 2$	$-21, -2$	$1, 3$

Claim 1. *The outcome* Out *is robust to canonical elaborations.*

In order to prove the above result, we will use the concept of essential set.

Definition 12. *A subset E of equilibria of a normal-form game G is essential if for every $\delta > 0$, there is $\varepsilon > 0$ such that every game G' satisfying*

$$\|G' - G\| \leq \varepsilon$$

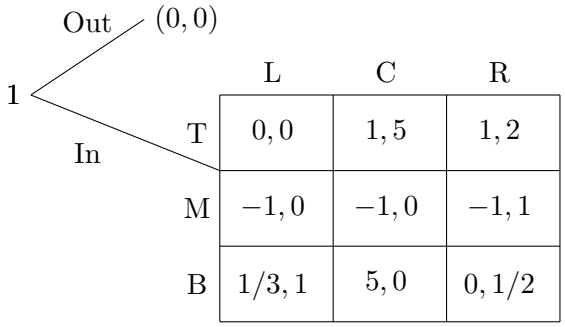
has an equilibrium α' that is δ-close to some equilibrium $\alpha \in E$.

For a given generic extensive-form game $(\Gamma, (g_i)_{i \in N})$, by mimicking the proof of Proposition 1, we get the following result.

Lemma 4. *Fix a complete-information extensive-form game $(\Gamma, (g_i)_{i\in N})$. If C is an essential component of $G^{\Gamma,g}$ and $\mu \in \Delta(Z)$ is the unique outcome associated with C, then μ is robust to canonical elaborations.*

Claim 1 follows from Lemma 4 since the equilibrium component with which the outcome Out is associated is essential (Hauk and Hurkens (2002, Proposition 1)).

We now show that Out is not a robust outcome. In order to do so, let us fix $\varepsilon \geq 0$ and build the following ε-elaboration $U^\varepsilon = (\Gamma, P, T, (u_i)_{i=1,2})$ of the above complete-information game. Player 1's set of types is $\{t_1^*, t_1^{\text{crazy}}\}$ while Player 2's set of types is a singleton $\{t_2^*\}$. The prior probability P over the set of possible profiles of types is as follows: $P(t_1^{\text{crazy}}, t_2^*) = \varepsilon$ and $P(t_1^*, t_2^*) = 1 - \varepsilon$. When Player 1 is the type t_1^*, the ex-post payoffs for both players are given by the above game i.e., $u_i(\cdot, (t_1^*, t_2^*)) = g_i(\cdot)$ for each $i = 1, 2$. However, when Player 1 is type t_1^{crazy}, the ex-post payoffs are given as follows:



Claim 2. *For each $\varepsilon > 0$, in the ε-elaboration U^ε, there is no equilibrium σ where $\sigma_1(t_1^*, \emptyset) = \text{Out}$.*[31]

Proof. Proceed by contradiction and assume that $\sigma_1(t_1^*, \emptyset) = \text{Out}$. Note that for type t_1^{crazy}, action In is a strictly dominant strategy for Player 1 (mixing in the stage game with full support on T and B) so that at any equilibrium $\sigma_1(t_1^{\text{crazy}}, \emptyset) = \text{In}$. Hence, if Player 2 sees Player 1 playing In, his only belief consistent with Bayes' rule puts probability one on t_1^{crazy}. Thus $(\sigma_1(t_1^{\text{crazy}}, \text{In}), \sigma_2(t_2^*, \text{In}))$ must be a Nash equilibrium of the 3×3 game given t_1^{crazy}. It is easily checked that this game has a unique equilibrium: (B, L). Hence, $\sigma_2(t_2^*, \text{In}) = \text{L}$. Thus, for type t_1^*, Player 1 by deviating from σ_1 and playing In and T guarantees a payoff of 5, which is greater than 0, i.e., what he would get if he were to follow the equilibrium strategy. This contradicts the assumption that σ is an equilibrium. ∎

Given the above claim, if Out is a robust outcome, then there must exist $\varepsilon > 0$ small enough such that in the ε-elaboration U^ε, $\sigma_1(t_1^*, \emptyset)$ puts strictly positive probability on Out and strictly positive probability on In as well. Hence, Player 1 of type t_1^* must be indifferent between Out and In. In order for this to be possible, Player 2 must be playing a (mixed) strategy in the set $[P_1, P_2] \cup [P_3, P_4] \cup (P_4, P_5]$ where each P_i are points in the simplex $\Delta(\{\text{L}, \text{C}, \text{R}\})$ defined as follows: $P_1 = (3/4, 1/4, 0)$, $P_2 = (11/16, 0, 5/16)$, $P_3 = (1/2, 0, 1/2)$, $P_4 = (31/282, 10/282, 241/282)$ and

[31] Here, $\sigma_i(t_i, h)$ denotes the mixed action played by type t_i at history h.

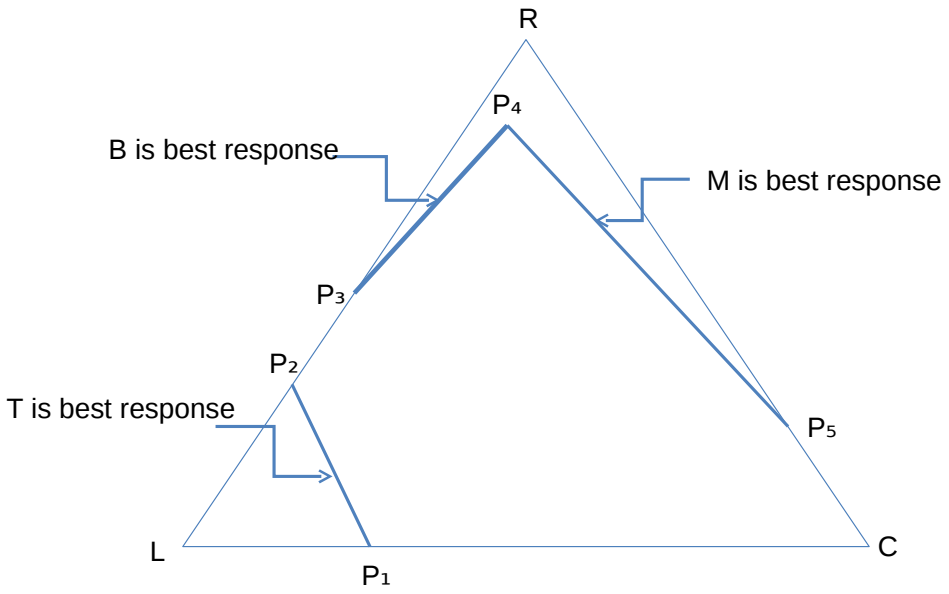


FIGURE 1. Best responses of t_1^* to Player 2's mixed strategies in the subgame

$P_5 = (0, 3/4, 1/4)$. For each of Player 2's strategy in $[P_1, P_2] \cup [P_3, P_4] \cup (P_4, P_5]$, Player 1 of type t_1^*'s best response(s) in the subgame G yields him a payoff of 0 and hence, he is indifferent between In and Out at the first stage. These best responses are summarized in Figure 1 below.

Now, the following three claims complete the proof. Below, β denotes the probability assigned to t_1^* by Player 2's beliefs after Player 1 has played In.

Claim 3. *For each $\varepsilon > 0$, in the ε-elaboration U^ε, there is no equilibrium σ where $\sigma_2(t_2^*, \text{In}) \in [P_1, P_2]$.*

Proof. Let us assume that for some $\varepsilon > 0$, in the ε-elaboration U^ε, there is an equilibrium σ where $\sigma_2(t_2^*, \text{In}) \in [P_1, P_2]$. The proof is splitted into two cases. First, consider the case where $\sigma_2(t_2^*, \text{In}) \in (P_1, P_2]$. In such a case, t_1^*'s best response is T while t_1^{crazy}'s best response is either T or B or both. Note that since $\sigma_2(t_2^*, \text{In}) \in (P_1, P_2]$, Player 2 is playing action R with strictly positive probability. We claim that whatever is the mixed action $(\alpha, 0, 1-\alpha)$ chosen by t_1^{crazy}, and whatever is β (the probability assigned to t_1^* by Player 2's beliefs after Player 1 has played In), R cannot be a best response for Player 2. Indeed, Player 2's payoffs can be described as follows:

- action L yields $4\beta + [0\alpha + 1(1-\alpha)](1-\beta)$;
- action C yields $0\beta + [5\alpha + 0(1-\alpha)](1-\beta)$;
- action R yields $2\beta + [2\alpha + 1/2(1-\alpha)](1-\beta)$.

Now, it is easily checked that (irrespective of α and β) for Player 2, the mixed strategy putting probability 4/7 on L and 3/7 on C yields strictly higher payoffs than action R, which is a contradiction.

Second, consider the case where $\sigma_2(t_2^*, \text{In}) = P_1$. In such a case, t_1^*'s best response is T while t_1^{crazy}'s best response is B. Player 2's unique best response is L regardless of Player 1's type, which is also a contradiction. ∎

Claim 4. *For each $\varepsilon > 0$, in the ε-elaboration U^ε, there is no equilibrium σ where $\sigma_2(t_2^*, \text{In}) \in [P_3, P_4]$.*

Proof. Let us assume that for some $\varepsilon > 0$, in the ε-elaboration U^ε, there is an equilibrium σ where $\sigma_2(t_2^*, \text{In}) \in [P_3, P_4]$. In such a case, t_1^*'s best response is either B or both M and B while t_1^{crazy}'s best response is T. Note that since $\sigma_2(t_2^*, \text{In}) \in [P_3, P_4]$, Player 2 is playing action L with strictly positive probability. Clearly, R is strictly better than L regardless of Player 1's type or t_1^*'s choice between M and B, which a contradiction. ∎

Claim 5. *For each $\varepsilon > 0$, in the ε-elaboration U^ε, there is no equilibrium σ where $\sigma_2(t_2^*, \text{In}) \in (P_4, P_5]$.*

Proof. Let us assume that for some $\varepsilon > 0$, in the ε-elaboration U^ε, there is an equilibrium σ where $\sigma_2(t_2^*, \text{In}) \in (P_4, P_5]$. In such a case, t_1^*'s best response is M while the best response of t_1^{crazy} is either T or B or both. Note that since $\sigma_2(t_2^*, \text{In}) \in (P_4, P_5]$, Player 2 is playing action R with strictly positive probability. It is easily checked, as in the first case in the proof of Claim 3, that for Player 2, the mixed strategy putting probability 4/7 on L and 3/7 on C is strictly better than action R regardless of Player 1's type or t_1^{crazy}'s choice between T and B, which is a contradiction. ∎

A.4. **Proof of Proposition 3.** Let $(\Gamma, (g_i)_{i \in N})$ be a complete-information extensive-form game. Define the quotient set of S_i, $\tilde{S}_i = \{[s_i] : s_i \in S_i\}$, where $[s_i]$ is the set of strategies that are outcome-equivalent to s_i, i.e., $[s_i] = \{s_i' : z(s_i, s_{-i}) = z(s_i', s_{-i}) \text{ for all } s_{-i}\}$. For each i and each nonempty subset B_i of $\tilde{S}_i$, define $g_i^{B_i} \colon Z \to \mathbb{R}$ by

$$g_i^{B_i}(z) = \begin{cases} g_i(z) & \text{if } z = z(s_i, s_{-i}) \text{ for some } s_i \text{ with } [s_i] \in B_i \text{ and } s_{-i}, \\ \min_z g_i(z) - 1 & \text{otherwise.} \end{cases}$$

Let type $t_i^{B_i}$ be the type of player i who knows that his payoffs are given by $g_i^{B_i}(\cdot)$. In particular, $t_i^* := t_i^{\tilde{S}_i}$ knows that his payoffs are given by $g_i(\cdot)$.

For any sufficiently small $\varepsilon > 0$, we consider the following elaboration $U = (\Gamma, P, T, (u_i)_{i \in N})$ of $(\Gamma, (g_i)_{i \in N})$, where player i's type space is $T_i = \{t_i^{B_i} : \emptyset \neq B_i \subseteq \tilde{S}_i\}$ and $P(t) = \prod_i P_i(t_i)$ with

$$P_i(t_i^{B_i}) = \frac{\varepsilon^{-|B_i|}}{(1+\varepsilon^{-1})^{|\tilde{S}_i|} - 1}.$$

Since

$$\begin{aligned} P(t^*) &= \prod_i \frac{\varepsilon^{-|\tilde{S}_i|}}{(1+\varepsilon^{-1})^{|\tilde{S}_i|}-1} \\ &> \prod_i \left(1-(2^{|\tilde{S}_i|}-2)\varepsilon\right) \\ &> 1-\sum_i (2^{|\tilde{S}_i|}-2)\varepsilon, \end{aligned}$$

U is an ε'-elaboration with $\varepsilon' \to 0$ as $\varepsilon \to 0$.

Let μ be a robust outcome. Then the elaboration U has an equilibrium σ with $z(\sigma(t^*))$ that is close to μ. We argue below that the induced strategy profile

$$\bar{\sigma}_i(s_i) = \frac{1}{|[s_i]|} \sum_{B_i} P_i(t_i^{B_i}) \sum_{s_i' \in [s_i]} \sigma_i(t_i^{B_i})(s_i')$$

is an ε''-proper equilibrium of the normal-form representation $G^{\Gamma,g} = (N, (S_i)_{i\in N}, (g_i^{\mathcal{N}})_{i\in N})$ of $(\Gamma, (g_i)_{i\in N})$, where $\varepsilon'' \to 0$ as $\varepsilon \to 0$. In addition, the outcome $z(\bar{\sigma})$ is close to μ (and converges to μ as ε vanishes), hence μ is induced by a proper equilibrium.

Note that one can extend Corollary 3 in the proof of Part 1 of Lemma 1 and show that if s_i has positive probability under $\sigma_i(t_i^{B_i})$, then $[s_i] \in B_i$.

Suppose that $g_i^{\mathcal{N}}(s_i, \bar{\sigma}_{-i}) > g_i^{\mathcal{N}}(s_i', \bar{\sigma}_{-i})$. By the above extension of Corollary 3, for any B_i' such that $\sigma_i(t_i^{B_i'})(s_i') > 0$, we have that $[s_i] \notin B_i'$ and $[s_i'] \in B_i'$. Let $B_i = B_i' \cup \{[s_i]\}$. Then, since $\varepsilon P_i(t_i^{B_i}) = P_i(t_i^{B_i'})$, and $\sigma_i(t_i^{B_i})$ assigns positive probabilities only on $[s_i]$, we have $|[s_i]|\varepsilon\bar{\sigma}_i(s_i) \geq \bar{\sigma}_i(s_i')$. Therefore, $\bar{\sigma}$ is a $\max_i \max_{s_i} |[s_i]|\varepsilon$-proper equilibrium.

Appendix B. Proof of Theorem 2

B.1. **Adaptation to Families of Complete-Information Extensive-Form Games.** In this section, we explain how we can adapt our arguments yielding Theorem 1 and Corollary 1 to the environment where the benchmark is a family of complete-information extensive-form games. First, we adapt our robustness notion to this environment.

Definition 13. *Fix a family of complete-information extensive-form games $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$. An equilibrium outcome $\mu\colon \Theta \to \Delta(Z)$ is robust if for any $\delta > 0$, there is $\varepsilon > 0$ such that any ε-elaboration of U has a Bayesian Nash equilibrium σ satisfying $\left\|\mu(\theta) - z(\sigma(t^\theta))\right\| < \delta$ for any $\theta \in \Theta$.*

We note that Proposition 1 holds with virtually the exact same proof. Hence, we obtain

Proposition 7. *Fix a family of complete-information extensive-form game U. If C is a hyper-stable component of the normal-form representation of U and $\mu\colon \Theta \to \Delta(Z)$ is the unique outcome associated with C, then μ is robust.*[32]

[32]We note that this proposition would hold when C is only required to be an essential set (see Definition 12) since, as mentioned in footnote 25, elaborations defined in Section 5 keep the set Θ constant and only perturb the prior probability P. Indeed, the normal-form representation of an ε-elaboration has the same set of strategies as the family

Proof. Fix $\delta > 0$. Let $\varepsilon > 0$ be as in the definition of hyperstability. Fix any $\tilde{\varepsilon}$-elaboration $\tilde{U} = (\Gamma, \tilde{P}, T, \Theta, (g_i)_{i\in N})$ of $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$ and let $(N, (S_i)_{i\in N}, (\tilde{g}_i^{\mathcal{N}})_{i\in N})$ be its normal-form representation. Similarly, let $(N, (S_i)_{i\in N}, (g_i^{\mathcal{N}})_{i\in N})$ be the normal-form representation of U. Note that the set of strategies in both U and $\tilde{U}$ is the same and for any $s \in \{S_i\}_{i\in N}$, $g_i^{\mathcal{N}}(s) = \sum_{(\theta,t)\in\Theta\times T} P(\theta,t) g_i(z(s(t)),\theta)$ while $\tilde{g}_i^{\mathcal{N}}(s) = \sum_{(\theta,t)\in\Theta\times T} \tilde{P}(\theta,t) g_i(z(s(t)),\theta)$. The positive number $\tilde{\varepsilon}$ is yet to be specified.

It is clear that game $(N, (S_i)_{i\in N}, (\tilde{g}_i^{\mathcal{N}})_{i\in N})$ is in a neighborhood of the game $(N, (S_i)_{i\in N}, (g_i^{\mathcal{N}})_{i\in N})$. Indeed,

$$\begin{aligned}\left|\tilde{g}_i^{\mathcal{N}}(s) - g_i^{\mathcal{N}}(s)\right| &\leq \sum_{(\theta,t)\in\Theta\times T} \left|\tilde{P}(\theta,t) - P(\theta,t)\right| \left|g_i(z(s(t)),\theta)\right| \\ &\leq \tilde{\varepsilon} \sum_{(\theta,t)\in\Theta\times T} \left|g_i(z(s(t)),\theta)\right|.\end{aligned}$$

Since Z and Θ are finite, the above term tends to 0 as $\tilde{\varepsilon}$ tends to 0. Hence, for $\tilde{\varepsilon}$ small enough, $(N, (S_i)_{i\in N}, (\tilde{g}_i^{\mathcal{N}})_{i\in N})$ is indeed in the ε-neighborhood of $(N, (S_i)_{i\in N}, (g_i^{\mathcal{N}})_{i\in N})$.

Now, by the definition of hyperstability, $(N, (S_i)_{i\in N}, (\tilde{g}_i^{\mathcal{N}})_{i\in N})$ has an equilibrium $\tilde{\sigma}$ that is δ-close to some equilibrium in C. Hence, by the continuity of z, and assuming here again that $\tilde{\varepsilon}$ is small enough, $z(\tilde{\sigma}(t^\theta))$ is δ-close to $\mu(\theta)$ for each θ. Hence, the $\tilde{\varepsilon}$-elaboration $\tilde{U}$ has a Bayesian Nash equilibrium $\tilde{\sigma}$ satisfying $\left\|\mu(\theta) - z(\tilde{\sigma}(t^\theta))\right\| < \delta$ for any $\theta \in \Theta$. ∎

The following result is a straightforward implication of Proposition 7 together with the fact that there exists a connected hyperstable component in any normal-form game (Kohlberg and Mertens (1986, Proposition 1)).

Theorem 3. *In a family of complete-information extensive-form games $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$, if $(g_i(\cdot,\theta))_{i\in N} \in \mathcal{G}$ for any θ, then at least one of the equilibrium outcomes is robust.*

Now, assume that Γ is a perfect-information extensive game form. If, for each $\theta \in \Theta$, $(\Gamma, (g_i(\cdot,\theta))_{i\in N})$ has a unique subgame-perfect equilibrium (as is the case when each player has distinct payoffs on all terminal histories), it is clear that $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$ must have a unique subgame-perfect equilibrium as well, which plays the unique subgame-perfect equilibrium of $(\Gamma, (g_i(\cdot,\theta))_{i\in N})$ at each type profile t^θ. But, then, given that a hyperstable component must contain a subgame-perfect equilibrium by Kohlberg and Mertens (1986, Proposition 5), this unique subgame-perfect equilibrium must be contained in a hyperstable component of the normal-form representation of U. If U has finitely many Nash equilibrium outcomes (which, as we already noted, is ensured when each player has distinct payoffs on all terminal histories), again Proposition 7 and the fact that there exists a hyperstable component (and that all components are connected) yield the following analogue of Corollary 1 in the environment where the benchmark is a family of complete-information extensive-form games.

---

of complete-information extensive-form games $U = (\Gamma, P, T, \Theta, (g_i)_{i\in N})$. Put in another way, contrary to the proof of Proposition 1, we do not have to use the notion of equivalent games for the proof of Proposition 7.

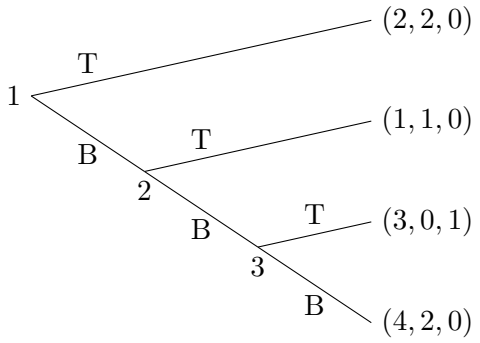
Corollary 4. *Fix a family of complete-information extensive-form games* $U = (\Gamma, P, T, \Theta, (g_i)_{i \in N})$. *If* Γ *is a perfect-information extensive game form, and* $g_i(z, \theta) \neq g_i(z', \theta)$ *whenever* $z \neq z'$, *then the backward-induction outcome for each* θ *is a robust equilibrium outcome.*[33]

B.2. **Proof of Theorem 2.** Let C be any hyperstable component of $U = (\Gamma, P, T, \Theta, (g_i^{\mathcal{X}}(\gamma(\cdot), \cdot))_{i \in N})$. By assumption, there are finitely Nash equilibrium outcomes and so (by connectedness of C), the equilibrium outcome is constant over C. Since, as we already mentioned, C contains a subgame-perfect equilibrium and under full implementation in subgame-perfect equilibrium, all subgame-perfect equilibria must yield the right outcome f, we know that the equilibrium outcome associated with C is the right outcome f. By Proposition 7, f is robust. Thus, we obtain Theorem 2.

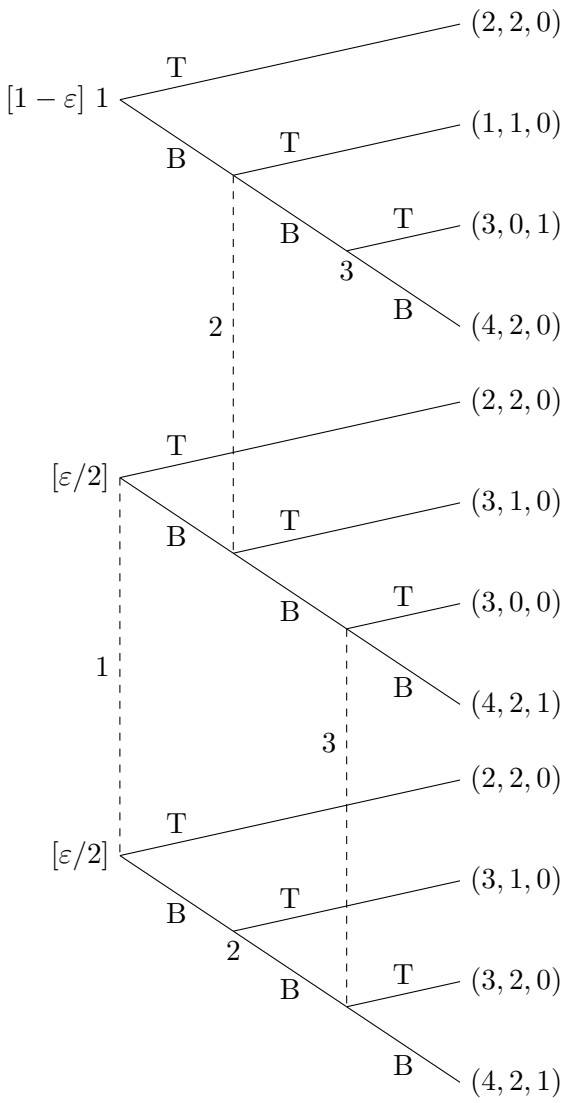
[33] As we mentioned in footnote 25, elaborations defined in Section 5 keep the set Θ constant and only perturb the prior probability P. Hence, we cannot replicate the proof of Proposition 3 here, and this is why we are not claiming uniqueness in the statement of Corollary 4.

Appendix C. Further Examples (not for publication)

C.1. **A variation on Example 1 where players know their own payoffs.** In Example 1 in Section 2, Player 2's payoffs on terminal histories depend on Player 1's type, and hence and Player 2 learns her own payoffs through Player 1's action in the first stage. In the sequel, we show that we can modify this example and get the same result while ensuring that each player knows his own payoffs. Consider the following perfect-information extensive-form game:



The unique subgame-perfect equilibrium is for each player to play T. Now consider an incomplete-information elaboration such that each player has two types, t_i^* and t_i^{crazy}, (t_1^*, t_2^*, t_3^*) occurs with probability $1-\varepsilon$, and $(t_1^{\text{crazy}}, t_2^*, t_3^{\text{crazy}})$ and $(t_1^{\text{crazy}}, t_2^{\text{crazy}}, t_3^{\text{crazy}})$ occur with probability $\varepsilon/2$ each. (Other combinations never occur.) Payoffs are given as follows:

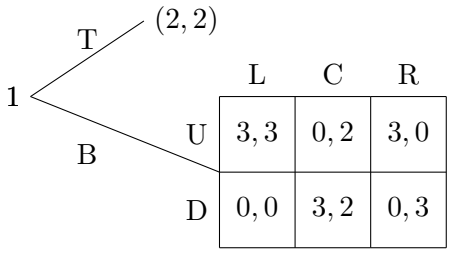


Here, the top tree is for (t_1^*, t_2^*, t_3^*), the middle for $(t_1^{\text{crazy}}, t_2^*, t_3^{\text{crazy}})$, and the bottom for $(t_1^{\text{crazy}}, t_2^{\text{crazy}}, t_3^{\text{crazy}})$. Note that in this example, all players know their own payoffs. By an argument similar to Example 1, we can show that there is no equilibrium where Player 1 of type t_1^* plays T for sure. Indeed, assume there is such an equilibrium, Player 1's crazy type plays B. Given this, Player 2's crazy type plays B as well and so finally, all crazy types play B. Then observing Player 1's action B at the beginning of the second stage, Player 2 of type t_2^* believes with probability one that Player 1 is of type t_1^{crazy}, and hence, given the prior probability, that Player 3 is of type t_3^{crazy}, who will play B in the third stage. Hence Player 2 of type t_2^* plays B, which makes Player 1 of type t_1^* strictly prefer B to T, a contradiction.

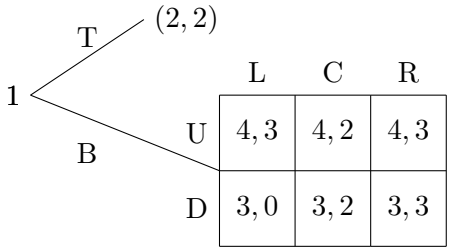
Note that we can extend the above argument and show that there is no equilibrium in mixed strategies where both t_1^* and t_2^* play T with high probabilities.

C.2. **Stable outcomes may not be robust to canonical elaborations even with independent types.** To prove that a stable outcome may not be robust to canonical elaborations when

types are drawn independently, we study the following outside-option game borrowed from van Damme (1989). There are two players $i = 1, 2$ and two stages. Player 1 plays first and chooses between Top (T) and Bottom (B). If (and only if) Player 1 plays B, then the players play a 2×3 normal-form game. Payoffs are given as follows:



By van Damme (1989), the outcome corresponding to action T for Player 1 is stable. We show that this outcome is not robust to canonical elaborations. In order to do so, let us build a canonical elaboration of the above game. Player 1 is assumed to have two types: t_1^* and t_1^{crazy} while Player 2 has a single type t_2^*. The prior probability over the set of possible profiles of types puts a probability $\varepsilon > 0$ on $(t_1^{\text{crazy}}, t_2^*)$ and $(1 - \varepsilon)$ on (t_1^*, t_2^*). When Player 1's type is t_1^*, the ex-post payoffs for both players are given by the above game. Player 1 with type t_1^{crazy} is committed to play B and then U. However, when Player 1 is type t_1^{crazy}, the ex-post payoffs are given as follows:



Using arguments similar to those developed in Section 2, let us first show that for any $\varepsilon > 0$, in the above incomplete information elaboration, there is no equilibrium where Player 1 plays T with probability 1. To see this, proceed by contradiction and assume that we do have such an equilibrium. Recall that t_1^{crazy} is committed to play B and then U. Hence, if Player 2 sees Player 1 playing B, then under the equilibrium, by Bayes' rule, he assigns probability one to Player 1 being of type t_1^{crazy} and playing U in the subgame. Hence, Player 2 at equilibrium does not play C (it is dominated at $(t_1^{\text{crazy}}, t_2^*)$). Therefore, if Player 1 is of type t_1^*, by deviating from his equilibrium strategy playing B and U guarantees a payoff of 3 which is greater than 2, i.e., what he would get if he were to follow his equilibrium strategy, which shows a contradiction.

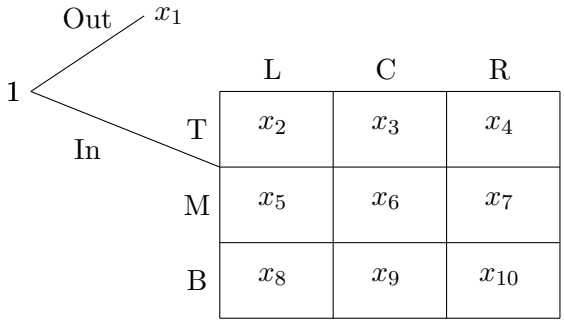
Given the above argument, if there is an equilibrium that yields an outcome close to T (i.e., which converges to T as ε vanishes), Player 1 has to be mixing when his type is t_1^*. So, at equilibrium, he must be indifferent between T and B. One can check that, for this indifference to be possible, Player 2 must be playing a mixed action that is either a convex combination of $P_1 = (2/3, 1/3, 0)$ and $P_2 = (0, 1/3, 2/3)$—in which case, type t_1^* plays U on the second stage and is indifferent between T and B—or a convex combination of $P_3 = (1/3, 2/3, 0)$ and $P_4 = (0, 2/3, 1/3)$—in which case, type

t_1^* plays D on the second stage and is indifferent between T and B. Note that in any case, Player 2 puts a positive weight on C.

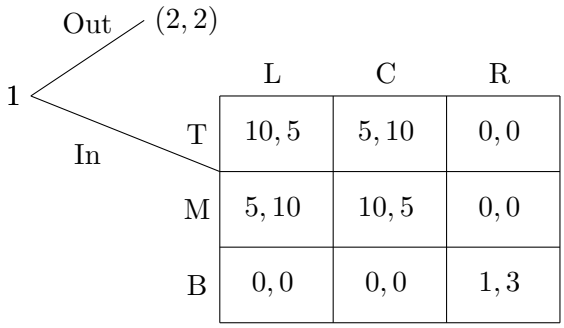
So if Player 2 plays on the segment $[P_1, P_2]$, then in the second stage, he believes that either "Player 1 is of type t_1^* and plays U" or "Player 1 is of type t_1^{crazy} and plays U." It is easily checked that L is strictly better than C regardless of Player 1's type, which is a contradiction with the fact that C is played with positive weight by Player 2. Similarly, if Player 2 plays on the segment $[P_3, P_4]$, then in the second stage, he believes that either "Player 1 is of type t_1^* and plays D" or "Player 1 is of type t_1^{crazy} and plays U." It is easily checked that R is strictly better than C regardless of Player 1's type, which is also a contradiction with the fact that C is played with positive weight by P2.

C.3. **Lack of robustness of full implementation in pure subgame-perfect equilibrium.** In the sequel, we consider the notion of (full) implementation of a social choice function when the solution concept is given by *pure* subgame perfect equilibrium. We show that under this weaker requirement (compared to that in Definition 8), Theorem 2 does not hold.

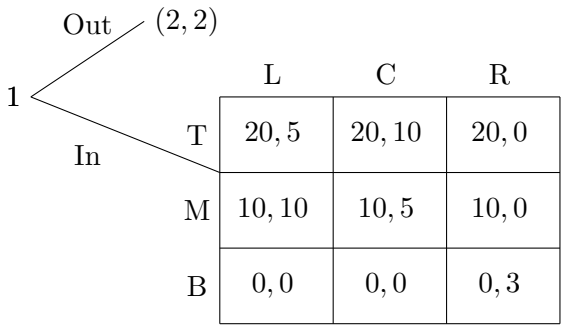
Consider the following example. There are two states of nature θ' and θ''. There are two players 1 and 2. There are ten alternatives $\mathcal{X} = \{x_1, \ldots, x_{10}\}$. We consider a social choice function given by $f(\theta') = x_1$ and $f(\theta'') = x_3$. Ex-post payoffs are yet to be specified. We consider a two-stage mechanism: Player 1 plays first and chooses between Out and In. If (and only if) Player 1 plays In, then the players play a 3-by-3 static game. The description of the mechanism is completed through the following extensive game form



Now, we specify ex-post payoffs over alternatives at each state by a description of the extensive-form game at each state. Ex-post payoffs at state θ' are given by



while ex-post payoffs at state θ'' are given by



It is easy to check that at each state $\theta \in \{\theta', \theta''\}$, when the state θ is commonly known, there is a unique *pure* subgame-perfect equilibrium in the induced extensive-form game yielding alternative $f(\theta)$. So the above mechanism fully implements the social choice function f when the solution concept is pure subgame-perfect equilibrium. However, one can mimic the argument in Example 2 to show that for any $\varepsilon > 0$ there is an ε-elaboration of the above family of complete-information extensive-form games with no equilibrium yielding an outcome close to x_1 at the profile of type $t^{\theta'}$. Hence, Theorem 2 does not hold when the solution concept for implementation is pure subgame-perfect equilibrium.[34]

In order to see this, let us first specify the complete information prior P as

$P:$		$t_1^{\theta'}, t_2^{\theta'}$	$t_1^{\theta'}, t_2^{\theta''}$	$t_1^{\theta''}, t_2^{\theta'}$	$t_1^{\theta''}, t_2^{\theta''}$
	θ'	$1-\alpha$	0	0	0
	θ''	0	0	0	α

where α is arbitrary in $(0, 1)$. We denote by U the family of complete-information extensive-form games defined here. Now, for any $\varepsilon > 0$, let us build the following prior P^ε satisfying $\|P^\varepsilon - P\| \leq \varepsilon$

$P^\varepsilon:$		$t_1^{\theta'}, t_2^{\theta'}$	$t_1^{\theta'}, t_2^{\theta''}$	$t_1^{\theta''}, t_2^{\theta'}$	$t_1^{\theta''}, t_2^{\theta''}$
	θ'	$1-\alpha$	0	0	0
	θ''	0	0	ε	$\alpha - \varepsilon$

.

In the sequel, we denote U^ε for the ε-elaboration of U where the prior is given by P^ε. Note that, conditional on observing the two types $t_1^{\theta''}, t_2^{\theta'}$ of players, the probability distribution over Θ equals the dirac measure on θ''. This special feature implies that if players 1 and 2 knew that they were respectively of types $t_1^{\theta''}$ and $t_2^{\theta'}$, they will conclude that with probability 1 the state is θ''.[35]

In the remaining, we show that for any $\varepsilon > 0$, U^ε has no equilibrium where Player 1 of type $t_1^{\theta'}$ plays Out with positive probability. First, it is easy to show that $t_1^{\theta'}$ cannot play Out with probability one. To see this, proceed by contradiction, and assume that there is an equilibrium under which $t_1^{\theta'}$ plays Out with probability one. Since Player 1 has a dominant strategy to play In when his type is $t_1^{\theta''}$, whenever Player 2 (irrespective of his type) sees that Player 1 has played In, his only belief consistent with Bayes' rule puts probability one on the event that Player 1 is of

[34] We note that at each state of nature of the above extensive-form games, there are finitely many Nash equilibrium outcomes.

[35] The above prior can be made full support replacing 0 by ε^2. The essence of our argument remains the same.

type $t_1^{\theta''}$. Hence, in case Player 2 sees Player 1 playing action In, he believes that Player 1 will be playing T in the subgame (since the only state consistent with $t_1^{\theta''}$ is θ'') and so he himself plays action C in the subgame. Given that, it is clearly profitable for Player 1 of type $t_1^{\theta'}$ to deviate from his equilibrium action and play In and then M in the subgame. This yields a contradiction.

Now, let us show that for any $\varepsilon > 0$, U^ε has no equilibrium where Player 1 of type $t_1^{\theta'}$ mixes over actions Out and In. Indeed, for this to be possible, $t_1^{\theta'}$ has to be indifferent between Out and In. Hence, in the subgame, Player 1 of type $t_1^{\theta'}$ cannot be playing B. Otherwise, if Player 1 of type $t_1^{\theta'}$ were to put a positive probability on B in the subgame, his expected equilibrium payoff in the subgame would be no more than his expected payoff from playing B, which is at most 1 regardless of Player 2's equilibrium strategy. Hence, $t_1^{\theta'}$ would not be indifferent between Out and In, which is a contradiction. Hence, Player 1 does not play B in the subgame irrespective of his type. Given this, irrespective of his type, Player 2 does not play R in the subgame, and here again, $t_1^{\theta'}$ would not be indifferent between Out and In. In conclusion, there is no equilibrium that makes Player 1 of type $t_1^{\theta'}$ playing Out with positive probability which proves the claim.

Remark 7. Note that the argument remains true for a nonempty open set of payoffs over terminal nodes.

Remark 8. The reason why Theorem 2 fails when using pure subgame-perfect equilibrium as a solution concept is because hyperstable components may contain only non-pure subgame-perfect equilibria. Hence, even though all pure subgame-perfect equilibria yield the desired outcome given by the social choice function, it may be that none of them is contained in a hyperstable component. However, with full implementation in subgame-perfect equilibrium, we are sure that all subgame-perfect equilibria—both pure and non-pure—contained in a hyperstable set yield the desired outcome.

References

[1] Aghion, P., D. Fudenberg, R. Holden, T. Kunimoto, and O. Tercieux (2012) "Subgame-Perfect Implementation under Information Perturbations," *Quarterly Journal of Economics*, 127, 1843–1881

[2] Bagwell, K. (1995) "Commitment and Observability in Games," *Games and Economic Behavior*, 8, 271–280.

[3] Bergemann, D. and S. Morris (2012) *Robust Mechanism Design: The Role of Private Information and Higher Order Beliefs*, World Scientific.

[4] Chassang, S. and S. Takahashi (2011) "Robustness to Incomplete Information in Repeated Games," *Theoretical Economics*, 6, 49–93.

[5] Chen, Y.-C. (2012) "A Structure Theorem for Rationalizability in the Normal Form of Dynamic Games," *Games and Economic Behavior*, 75, 587–597.

[6] Chen, Y.-C., Di Tillio A., Faingold, E., and S. Xiong (2010) "Uniform topologies on types," *Theoretical Economics*, 5, 445–478.

[7] Chen, Y.-C., Holden, R., Kunimoto, T., Sun, Y., and T. Wilkening (2017) "Getting Dynamic Implementation to Work," mimeo.

[8] Chung, K., and J. Ely (2013) "Implementation with Near-Complete Information," *Econometrica*, 71 857–871.

[9] van Damme, E. (1984) "A Relation between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games," *International Journal of Game Theory*, 13, 1–13.

[10] van Damme, E. (1989) "Stable equilibria and forward induction," *Journal of Economic Theory*, 48, 476–496.

[11] van Damme, E. and S. Hurkens (1997) "Games with Imperfectly Observable Commitment," *Games and Economic Behavior*, 21, 282–308.

[12] Dekel, E. and D. Fudenberg (1990) "Rational Play Under Payoff Uncertainty," *Journal of Economic Theory*, 52, 243–267.

[13] DeMichelis, S. and F. Germano (2000) "On the Indices of Zeros of Nash Fields," *Journal of Economic Theory*, 94, 192–217.

[14] Dold, A. (1972) *Lectures on Algebraic Topology*, Springer-Verlag.

[15] Fudenberg, D., D. Kreps, and D. K. Levine (1988) "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44, 354–380.

[16] Fudenberg, D. and D. K. Levine (1989) "Reputation and Equilibrium Selection in Games with a Patient Player," *Econometrica*, 57, 759–778.

[17] Fudenberg D. and J. Tirole (1991) *Game Theory*, MIT Press.

[18] Govindan, S. and A. McLennan (2001) "On the Generic Finiteness of Equilibrium Outcome Distributions in Game Forms," *Econometrica*, 69, 455–471.

[19] Govindan, S. and R. Wilson (2001) "Maximal Stable Sets of Two-Player Games," *International Journal of Game Theory*, 30, 557–566.

[20] Govindan, S. and R. Wilson (2005) "Essential Equilibria," *Proceeding of the National Academy of Science*, 102, 15706–15711.

[21] Güth, W., G. Kirchsteiger, and K. Ritzberger (1998) "Imperfectly Observable Commitments in n-Player Games," *Games and Economic Behavior*, 23, 54–74.

[22] Harsanyi, J. (1967) "Games with Incomplete Information Played by Bayesian Players. Part I: The Basic Model," *Management Science*, 14, 159–182.

[23] Hart, O. and J. Moore (2003) "Some (Crude) Foundations for Incomplete Contracts," mimeo.

[24] Hauk, E. and S. Hurkens (2002) "On Forward Induction and Evolutionary and Strategic Stability," *Journal of Economic Theory*, 106, 66–90.

[25] Hillas, J. (1990) "On the Definition of the Strategic Stability of Equilibria," *Econometrica*, 58, 1365–1390.

[26] Jackson, M. (1989) "Bayesian Implementation," *Econometrica*, 59, 461–477.

[27] M. Jackson (2001) "A Crash Course in Implementation Theory," *Social Choice and Welfare*, 18, 655-708.

[28] Kajii, A. and S. Morris (1997a) "The Robustness of Equilibria to Incomplete Information," *Econometrica*, 65, 1283–1309.

[29] Kajii, A. and S. Morris (1997b) "Refinements and Higher Order Beliefs: A Unified Survey," mimeo.

[30] Kartik, N. and O. Tercieux (2012) "A Note on Mixed-Nash Implementation," mimeo.

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

[32] Kreps, D. M. and R. Wilson (1982) "Sequential Equilibria," *Econometrica*, 50, 863–894.

[33] Kuhn, H. W. (1953) "Extensive Games and the Problem of Information" in Kuhn, H. W., Tucker, A. W. (eds.), *Contributions to the Theory of Games, Volume II*, Princeton University Press, 193–216.

[34] Mailath, G. J., and Samuelson, L. (2006) *Repeated Games and Reputations: Long-Run Relationships*, Oxford University Press.

[35] Maskin E. (1999) "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, 66, 23–38.

[36] Maskin, E. and J. Tirole, "Two Remarks on the Property-Rights Literature," *Review of Economic Studies*, 66, 139–149.

[37] Mertens, J.-F. (1989) "Stable Equilibria|A Reformulation, Part I: Definition and Basic Properties," *Mathematics of Operations Research*, 14, 575–624.

[38] Mezzetti, C. and L. Renou (2012) "Implementation in Mixed Nash Equilibrium," *Journal of Economic Theory*, 147, 2357-2375.

[39] Moore, J. and R. Repullo "Subgame Perfect Implementation," *Econometrica*, 56, 1191–1220.

[40] Monderer, D. and D. Samet (1989) "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior*, 1, 170–190.

[41] Morris, S., Takahashi, S., and O. Tercieux (2012) "Robust Rationalizability under Almost Certainty of Payoffs," *Japanese Economic Review*, 63, 57–67.

[42] Morris, S. and T. Ui (2005) "Generalized Potentials and Robust Sets of Equilibria," *Journal of Economic Theory*, 124, 45–78.

[43] Myerson, R. B. (1978) "Refinements of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7, 73–80.

[44] Osborne, M. and A. Rubinstein (1994) *A Course in Game Theory*, MIT Press.

[45] Palfrey, T.R., and S. Srivastava (1991) "Nash Implementation Using Undominated Strategies," *Econometrica*, 59, 479-501.

[46] Penta, A. (2012) "Higher Order Uncertainty and Information: Static and Dynamic Games," *Econometrica*, 2, 631–660.

[47] Pram, K. (2019) "On the Equivalence of Robustness to Canonical and General Elaborations," *Journal of Economic Theory*, 180, 1–10.

[48] Ritzberger, K. (1994) "The Theory of Normal Form Games from the Differentiable Viewpoint," *International Journal of Game Theory*, 23, 207–236.

[49] Rubinstein, A. (1989) "The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge"," *American Economic Review*, 79, 385–391.

[50] Selten, R. (1975) "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25–55.

[51] Serrano, R. and R. Vohra (2010) "Multiplicity of mixed equilibria in mechanisms: A unified approach to exact and approximate implementation," *Journal of Mathematical Economics*, 46, 775-785.

[52] Shapley, L. (1974) "A Note on the Lemke-Howson Algorithm," *Mathematical Programming Study*, 1, 175–189.

[53] Takahashi, S. (2019) "Non-Equivalence between All and Canonical Elaborations," mimeo.

[54] Ui, T. (2001) "Robust Equilibria of Potential Games," *Econometrica*, 69, 1373–1380.

[55] Weinstein, J. and M. Yildiz (2007) "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements," *Econometrica*, 75, 365–400.

[56] Weinstein, J. and M. Yildiz (2013) "Robust Predictions in Infinite-Horizon Games|An Unrefinable Folk Theorem," *Review of Economic Studies*, 80, 365–394.