ECONOMIC RESEARCH REPORTS

Strategic Delegation By Unobservable Incentive Contracts

By

Levent Kockesen and Efe A. OK

RR# 99-11 May 1999

C.V. STARR CENTER
FOR APPLIED ECONOMICS

NEW YORK UNIVERSITY
FACULTY OF ARTS AND SCIENCE
DEPARTMENT OF ECONOMICS
WASHINGTON SQUARE
NEW YORK, NY 10003-6687

Strategic Delegation By Unobservable Incentive Contracts[*]

Levent Koçkesen[†] Efe A. Ok[‡]

February 1999
(First version: November 1998)

Abstract

Many strategic interactions in the real world take place among delegates empowered to act on behalf of others. Although there may be a multitude of reasons why delegation arises in reality, one intriguing possibility is that it yields a strategic advantage to the delegating party. We analyze the possibility that strategic delegation arises as an equilibrium outcome under completely unobservable incentive contracts within the class of two-person extensive form games with perfect information. We show that delegation *may* arise solely due to strategic reasons in quite general economic environments even under unobservable contracts. Furthermore, under some reasonable restrictions on out-of-equilibrium beliefs and actions of the outside party, strategic delegation is shown to be the *only* equilibrium outcome.

JEL Classification: C72, D80.
Keywords: Strategic Delegation, Unobservable Contracts, Forward Induction.

[*] We thank Kyle Bagwell, JP Benoit, Douglas Gale, Faruk Gul, Boyan Jovanovic, Ehud Kalai, Giuseppe Lopomo, George Mailath, Ariel Rubinstein, Andy Schotter, Rajiv Sethi, and seminar participants in Columbia University, University of Michigan-Ann Arbor, Princeton University, Rutgers University, and University of Texas-Austin for helpful comments and suggestions. The research of the first author is supported by the Alfred P. Sloan Foundation. Both authors also acknowledge the support from the C. V. Starr Center for Applied Economics at New York University.

[†] Department of Economics, New York University, 269 Mercer Street, New York NY 10003. E-mail: kockesen@fasecon.econ.nyu.edu.

[‡] Department of Economics, New York University, 269 Mercer Street, New York NY 10003. E-mail: okefe@fasecon.econ.nyu.edu.

1 Introduction

Many strategic interactions in the real world take place among delegates empowered to act on behalf of others. Managers make strategic decisions that affect profits; sales persons have power over setting prices; and lawyers, sports agents, diplomats and union officials represent their clients and constituents in bargaining processes. An important question is if this could be, at least partly, due to the strategic advantage that delegation may provide to the delegating party. After all, it is well known that signing binding and publicly observable contracts with a third party may serve as a commitment device, and thus yield a strategic advantage. This idea, which goes back at least to Schelling (1960), has been put into use in many areas of economics.[1]

However, the observability of contracts appears to be a precondition for them to play a commitment role, and for this reason, almost all applications of strategic delegation theory are couched in terms of observable contracts. The formalization of this intuition is given by Katz (1991) who showed in his oft-cited paper that if contracts are unobservable, then the Nash equilibrium outcomes of a game with or without delegation are the same. In particular, delegation through unobservable contracts does not change the predicted outcome of games with a unique Nash equilibrium. These observations, in turn, call the empirical relevance of the applied delegation studies into serious question since, in most real-world transactions, the signed contracts are unobservable to the outside parties.

Before deeming the strategic delegation theory useless, however, one should recall that in games where actions are taken in a sequential manner, some type of sequential rationality is naturally imposed on the part of the players and this makes the set of predicted outcomes generically smaller than the set of Nash equilibrium outcomes. Therefore, Katz's result is less informative in such games as it does not let us see whether unobserved delegation could lead to a change in the set of predicted outcomes in games with a sequential structure. A case in point is recently provided by Fershtman and Kalai (1997), who, within the context of an ultimatum bargaining example, demonstrated that delegation could change the equilibrium outcome even under completely unobservable contracts.[2]

It is then important to understand under what conditions, if any, strategic delegation would arise as an equilibrium outcome and change the predicted outcome of a strategic interaction when contracts between the principals and delegates are not observable. In this

[1] A partial list includes applications in industrial organization (Vickers, 1985; Fershtman and Judd, 1987; Sklivas, 1987; Brander and Lewis, 1986; Bolton and Scharfstein, 1990; Hadfield, 1991), in international trade (Brander and Spencer 1983, 1985; Eaton and Grossman, 1986; Gatsios and Karp, 1991; Das, 1997), in bargaining theory (Segendorff, 1998), and in monetary policy (Persson and Tabellini, 1993; Walsh, 1995; Jensen, 1997).

[2] Katz (1991) himself reported a similar result using a simple bargaining example.

paper we study this issue within the context of arbitrary finite two-person extensive form games with a unique sequentially rational equilibrium outcome. The main findings of the present paper will not only strengthen the observation made by Fershtman and Kalai (1997), but will also generalize it to a very large class of games.

We call any finite two-person extensive form game a *principals-only game* when each player (principal) plays the game himself. For easy future reference, let us call the sequentially rational equilibrium outcome of any principals-only game the *status quo* outcome. Given any principals-only game, we specify a *delegation game* as follows: in the first stage, one of the principals decides whether to play the game himself or to offer, at a cost, an incentive contract to an agent, which specifies the payoff to the agent as a function of the outcome of the game.[3] The agent, in turn, either accepts or rejects the offer. If the agent rejects the offer, the game is played between the principal and the outside party and the delegate receives her outside option. If she accepts, then the game is played between the delegate and the outside party and the delegate receives the payoff as specified by the contract.[4] The crucial point that distinguishes this scenario from the ones commonly considered in the literature is that in the present model the outside party does *not* observe the contract offered and knows only whether he is facing the principal or the delegate.[5]

Our main objective is to understand the nature of the perfect Bayesian equilibria of the delegation game. The first observation is that, provided that the cost of hiring an agent is relatively low, delegation may obtain in equilibrium. More importantly, the outcome induced in the principals-only game by delegation can be different from the status quo outcome (while it must be a Nash equilibrium outcome of the principals-only game). This observation shows that, even under fully unobservable contracts, the act of delegation may possess commitment powers that would alter the outcome that would have obtained in the absence of delegation. While the power of unobservable delegation is suspect in simultaneous-move games as shown

[3] In this paper, we study the scenarios in which only one party has the option to delegate. Understanding this simpler scenario is a prerequisite for a proper analysis of the more complicated (but more realistic) case of two-sided delegation. While we shall comment briefly on how our results modify in this case, we should refer the reader to Koçkesen (1998) for an extensive analysis of the issue.

[4] To concentrate on the strategic elements of delegation we assume that (1) the agent's sole function is to make decisions which does not require any effort, and (2) the principal and the agent are symmetrically informed throughout the game.

[5] We continue to assume, however, that no renegotiation of the contract between the principal and the delegate takes place after the outside party takes his first action. There is an important literature which analyzes commitment effects of delegation when renegotiations are allowed (see Dewatripont, 1988 and Caillaud et al., 1995 among others). This literature shows that delegation has commitment effects only if there is asymmetric information between the principal and the delegate when the initial contract is signed. As noted before, we posit symmetric information between the principal and the delegate throughout. We will have more to say on the renegotiation issue in Section 6 of the paper.

by Katz (1991), this is not necessarily the case in sequential principals-only games.

However, this finding bears an "it is possible that ..." sort of a statement, and hence provides only limited support for the presence of strategic delegation under unobservability. Yet, if we strengthen our equilibrium concept in a reasonable manner, we can understand the strategic consequences of unobserved delegation substantially better. For instance, it is possible to use a forward induction type argument to refine the equilibrium to show that delegation is essentially inevitable if the status quo payoff of the delegating party is not already the best that he can obtain within a potentially large set of Nash equilibrium payoffs of the principals-only game. The idea is simply that forward induction reinstates the commitment power of delegation since, under the forward induction hypothesis, the outside party interprets a delegation decision also as a signal about the contract that is signed. (Why is the principal paying an agent to play the game in place of him, unless he was not planning to instruct the agent to play in a manner that improves his situation over the status quo even after paying the cost of hiring an agent?)

Unfortunately, forward induction type arguments that yield the above conclusion run into formal difficulties in delegation games, as we shall explain in the sequel. However, it turns out that we can obtain the same result by using a simple and intuitive equilibrium refinement, the *well-supported equilibrium*, which is based on imposing certain reasonable restrictions on the out-of-equilibrium beliefs and behavior of the players. (Similar refinements are proposed by McLennan, 1985, and Hillas, 1994, and discussed extensively in Kreps, 1989). Our main result is that if there exists a Nash equilibrium outcome of the principals-only game in which *(i)* the delegating principal receives a payoff strictly greater than his status quo payoff, *(ii)* the outside party behaves sequentially rationally, then in *any* well-supported equilibrium (which always exists in such games), the principal will certainly choose to delegate rather than playing the game himself, provided that the cost of hiring an agent is not too high. Moreover, this will alter the status-quo outcome in a way that is (strictly) beneficial for the delegating party.

These results demonstrate the empirical relevance of studying strategic delegation. After all, important strategic interactions such as bargaining processes involved in international trade and policy arena, in settlements of civil lawsuits, as well as in union-firm negotiations; situations involving incentives to deter entry; and oligopolistic industries with a leader-follower structure can be reasonably modeled as finite extensive form games. Since our main result admits almost any such game as a principals-only game, it provides formal support to the assertion that, in a very general class of economic settings, strategic aspects of delegation would play an important role in contract design, even if the contracts were completely unobservable.

The paper is organized as follows. Section 2 analyzes a simple delegation game to provide motivation for our inquiry and develop the intuition behind our main result. In Section 3, we introduce the basic nomenclature and formally introduce the equilibrium refinement that we propose here for a particular class of games. We also examine the behavior of our refinement in this section by means of several examples. Section 4 formally describes the economic environment within which we analyze the main question of the paper. In turn, we present our main results in Section 5, and discuss some potential extensions along with some open questions in Section 6. The proofs of the results appear in Section 7. We conclude with an appendix that provides a definition of the notion of well-supported equilibrium as it applies to arbitrary games.

2 Motivation: A Simple Bargaining Example

In order to illustrate the main intuition behind our results we shall first analyze a simple ultimatum bargaining game in which player 1 gives either a low offer to player 2 (denoted l) or a high offer (denoted h), and player 2 either accepts (denoted y) or rejects (denoted n) the offer. If player 1 offers l and player 2 accepts, payoffs are \$2 and \$1 for player 1 and player 2, respectively. If player 1 offers h and player 2 accepts, then player 1 receives \$1 and player 2 receives \$2. If an offer is rejected, both players receive a payoff of zero. We refer to this game as the *principals-only game*, and note that it has two Nash equilibrium outcomes (l, y) and (h, y). However, only one of these outcomes is not based on incredible threats by player 2 (i.e., it is the outcome of a subgame perfect equilibrium): player 1 offers l and player 2 accepts.

Now, assume that one of the players (principals) has the option of hiring a third player (whom we call the *agent* (or the *delegate*) and denote by A) to play the game for him. More precisely, a player can either play the game himself, that is, not hire a delegate (this action is denoted $\neg D$), or he can offer a contract to the agent, at a cost $c > 0$, which specifies her payoffs as a function of the outcome of the principals-only game. In turn, the delegate can either accept or reject the contract. In case of rejection, player 1 and 2 play the game themselves, receive the same payoffs as in the principals-only game, except that player 2 pays the contracting cost c, and the delegate receives her outside option $\delta \geq 0$. If, on the other hand, she accepts the contract, then the delegate plays the game in place of the delegating player, and at any given outcome, she receives whatever the contract specifies for her, the delegating player receives the principals-only game payoff minus the cost of hiring, and the other player receives the same payoff as in the principals-only game.[6] While our description

[6] Of course, the delegating player could himself earn an outside option by delegating. We, however, assume that the outside option of this player is zero to make the analysis interesting. Clearly, if this outside option was large relative to the potential payoffs in the game, delegation would obtain due to unstrategic reasons.

of it is not yet complete, we shall loosely refer to the resulting game as a *delegation game* in what follows.

Let us begin by observing that if it is player 1 who has the option of hiring an agent, then irrespective of which contracts are feasible and whether they are observable or not, the unique equilibrium of the delegation game would be characterized by player 1 not hiring and hence sustaining his subgame perfect equilibrium payoff. After all, \$2 is the largest possible payoff player 1 can hope for in this game, and he expects to receive this payoff if he plays the game himself. Consequently, the analysis becomes interesting only if we assume that it is rather player 2 who has the option of delegating. We thus posit that this is the case throughout this section.

Let us assume that there are only two contracts available to player 2, T (for tough) and W (for weak), which are specified as follows:

$$T = \begin{cases} \delta, & \text{if outcome is in } \{(l,n),(h,y)\} \\ 0, & \text{otherwise} \end{cases}$$

$$W = \begin{cases} \delta, & \text{if outcome is in } \{(l,y),(h,y)\} \\ 0, & \text{otherwise} \end{cases}$$

with $\delta < 1$. We will assume in this section that the delegate accepts any contract which yields her at least her outside option as expected payoff. (In later sections we will work with general contract spaces in which case we will not need this assumption.) Therefore, the delegate accepts any contract offer in the menu $\{T, W\}$. This is because irrespective of player 1's offer, there is always an action on her behalf which would earn her δ. Also notice that the contract W exactly replicates player 2's incentives in the principals-only game whereas contract T gives incentives to the agent to accept only the high offer. Therefore, owing to the simplicity of the principals-only game at hand, the contract space $\{T, W\}$ actually includes all interesting contracts.

It is easy to see that if the cost of hiring a delegate is too high, i.e., $\delta + c > 1$, then in any perfect Bayesian equilibrium of the delegation game (independent of contracts being observable or not) player 2 chooses to play the game himself. Therefore, we will analyze the case where $\delta + c < 1$.

If the contract signed between player 2 and his delegate were observable to player 1, then the unique subgame perfect equilibrium outcome of the delegation game would have player 2 offering the contract T, player 1 offering h and the delegate accepting the offer. This is of course nothing but a simple demonstration of the beneficial commitment effects of *observable* delegation. Things get a bit more complicated, however, if we (realistically) assume that only the decision to hire a delegate or not is observable by player 1, not the contract offered. The description of the game becomes complete under this informational assumption; we refer to this game as the *delegation game* and depict its basic structure in Figure 1.[7]

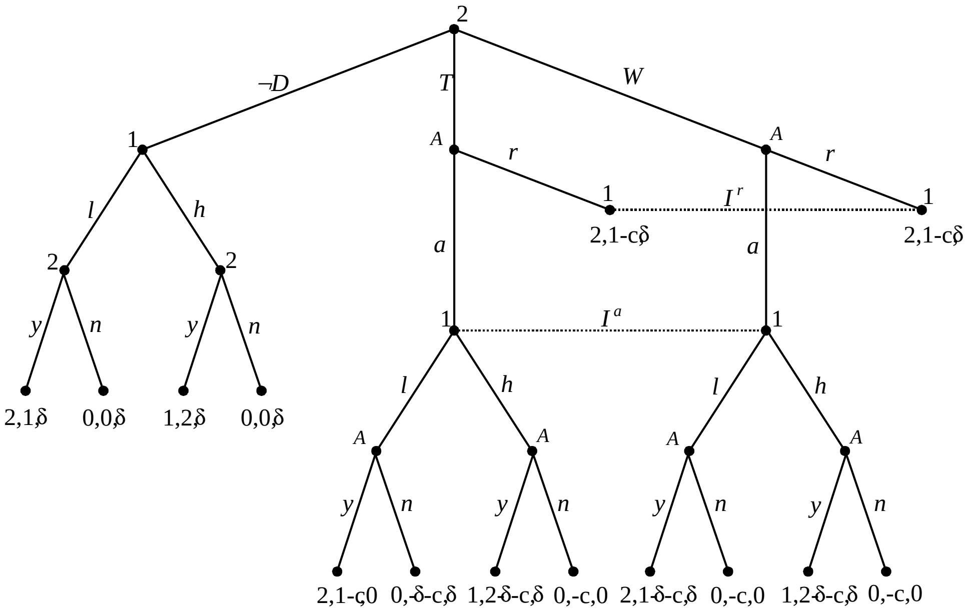

Figure 1: A Simple Delegation Game

There are two types of perfect Bayesian equilibria of this game. The first type is characterized by player 2 not delegating, and the second is characterized by player 2 choosing an action other than $\neg D$. In all equilibria of the first type, following player 2's action, player 1 offers l and player 2 accepts. At the out-of-equilibrium information set following an accepted contract (denoted I^a in Figure 1), player 1 believes that contract W has been offered with at least probability 1/2 and he plays l with at least probability $1-\delta-c$. In the second type of equilibrium, player 2 places at least probability 1/2 on contract T and, at the information set I^a, player 1 plays h.

While all of these equilibria are in fact trembling-hand perfect (Katz, 1991, Fersthman and Kalai, 1997), it is still possible to take issue with the plausibility of a first type of equilibrium. To make the associated argument as clearly as possible, let us first allow only for pure strategies. The point is that, in this case, we can apply a natural forward induction argument to "kill" any of the first type of equilibria. For instance, no such equilibrium survives the forward induction test proposed by van Damme (1989). In any forward induction-proof pure strategy equilibrium of the above delegation game, player 2 offers the contract T and player

[7] For simplicity, we truncated the branches of the game tree following the delegate's action r and replaced them by the equilibrium outcomes which would be realized if the play were ever to reach there.

1 plays h: forward induction ensures delegation even if the contracts are unobservable.[8]

Another observation which points out to what is unreasonable about the first type of equilibrium is that the contract W, which aligns the incentives of player 2 and the agent, is not offered in *any* pure strategy equilibria of the game since player 2 would be better off by playing the game himself rather than hiring an agent through the contract W. Yet, the only way one can support an equilibrium in which player 2 plays $\neg D$ is by assuming that player 1 believes at I^a that the agent he is facing has been offered nothing but the contract W!

In the final analysis, whether such beliefs are reasonable or not are formally captured in the equilibrium concept one adopts to "solve" the game. The out-of-equilibrium beliefs supporting an equilibrium of the first type (in which player 2 plays $\neg D$) are justified in a perfect Bayesian equilibrium simply because player 1 thinks that player 2 has made a mistake, without trying to make further inferences regarding player 2's possible play which caused the information set I^a to be reached. Suppose, in contrast, that player 1 rather reasons, upon facing a delegate unexpectedly, that it is actually he who made a mistake in assuming that a first type of equilibrium is accepted as the current norm. (Kreps, 1989, calls refinements based on this line of reasoning "mistaken theory" refinements; see Section 3.) He may then well conclude that player 2 is playing according to some other equilibrium in which a delegate is hired. But, in no such equilibrium player 2 offers the contract W, and therefore, so player 1 reasons, W cannot be the contract that is signed. Given that his beliefs put probability zero on contract W, player 1's optimal action is h and hence player 2 strictly prefers to delegate in any equilibrium which survives a "mistaken theory" refinement.

It is not clear which interpretation ("mistakes" or "mistaken theories") is more plausible in general. The answer is likely to depend on the situation being analyzed. In delegation games, however, there is reason to believe that the second interpretation is more convincing. These games depict situations in which individuals decide whether to hire someone to act on their behalf or not in a strategic interaction. It would not be reasonable to think that such a decision, which in reality requires time and effort, takes place without careful deliberation. Hiring someone is a costly action and it is unlikely to take place as a result of sheer irrationality or a simple mistake. Arguably, therefore, in all "reasonable" equilibria of the above delegation game, player 2 hires a delegate and offers a contract so that the equilibrium outcome is different from the subgame perfect equilibrium outcome of the principals-only game. Hence, we contend that delegation is likely to ensue even by means of *unobservable* contracts; this is the main thesis we shall defend formally in this paper.

[8] Strictly speaking, this conclusion requires us to add a superfluous move for player 2 so that he first decides whether to hire an agent or not, and only after this decision is made, he offers a contract. In this case, we may apply van Damme's forward induction condition since there is a unique pure strategy equilibrium of the subgame following the hire decision with a strictly higher payoff than the subgame perfect equilibrium payoff for player 2.

Of course, the above analysis is incomplete as it did not allow for mixed strategies. If one considers mixed strategies as well, van Damme's forward induction refinement looses its power (since in this case there are more than one equilibrium with delegation that gives to player 2 a payoff strictly higher than his subgame perfect equilibrium payoff). We could of course use a strengthening of van Damme's refinement, such as the forward induction refinement proposed by Al-Najjar (1995), to show that "no delegation" cannot be sustained in a mixed strategy forward induction equilibrium either. However, Al-Najjar's forward induction criterion is rather demanding and is known to eliminate certain reasonable equilibria in some games. One may thus view this refinement objectionable in the case of delegation environments.[9]

Fortunately, the off-the-equilibrium path reasoning we proposed above (that is, the mistaken theory approach) applies to the case of mixed strategies as well. Indeed, if one takes the position that players interpret deviations as a result of someone's confusion over which equilibrium is being played, it is possible to argue that the "no delegation" equilibria are implausible in general. To see this, suppose that, at information set I^a, player 1 believes that an equilibrium strategy by player 2 must have led the game to I^a, and he plays according to the precepts of that equilibrium. If player 2 "sees through" this reasoning, then he has all the incentives to delegate, because *all* equilibria in which he delegates have payoffs strictly higher than the equilibrium payoff she would obtain by not delegating.

In order to capture this intuition, it seems reasonable to require that player 1, upon finding himself at information set I^a, believes that previous play has conformed to an equilibrium which reaches that information set and follows the precepts of that equilibrium from then on, provided that such an equilibrium exists. If more than one such equilibrium exist, the player is free to contemplate any of those equilibria and if no such equilibrium exists then there is no further restriction on beliefs and actions. This is the basic premise behind the equilibrium refinement we posit in this paper and define formally in the next section. For the time being, let us loosely refer to any perfect Bayesian equilibrium whose out of equilibrium beliefs and actions satisfy the above mentioned restrictions as *well-supported.*

Endowed with the notion of well-supportedness, we may critically examine the no-delegation equilibria in mixed strategies as well. Indeed, it is not difficult to prove that none of the equilibria in which player 2 chooses $\neg D$ with positive probability (first type of

[9] Al-Najjar (1995) himself reports such an example. In a "burning money" game (Example 3 below), where the first mover either burns *zero* dollars or does not burn any money at all before playing a battle of the sexes game, his refinement chooses the equilibria with the highest payoff for this player. Moreover, this formulation of forward induction runs into a serious coordination problem, which is why Al-Najjar's refinement is most appealing in the case of repeated games. For the record, however, we note that all results of the present paper remain valid, if we adopted this refinement as opposed to the one we shall use in the sequel.

equilibria) is well-supported.[10] In all equilibria of this type player 1 offers l with positive probability, whereas in all equilibria that reach information set I^a (the second type of equilibria) player 1 offers l with zero probability. On the other hand, all the equilibria of the second type, i.e., delegation equilibria, are trivially well-supported.[11] Notice that if we have imposed the well-supportedness condition on the equilibria of the first type, i.e., player 1 plays h at the information set I^a, then player 2 will have an incentive to deviate from $\neg D$ and offer T or W, therefore this type of equilibria cannot be well-supported.

Even though the bargaining game we analyzed here is an extremely simple one, it provides important insights. The reason why delegation can be supported as a perfect Bayesian equilibrium is because (1) the cost of hiring a delegate is low enough, and (2) its outcome in the principals-only game, i.e., (h, y), is a Nash equilibrium outcome of the principals-only game which yields player 2 a higher payoff than his subgame perfect equilibrium outcome and in which player 1 plays sequentially rationally. Furthermore, if $\delta + c \neq 1$, then the delegation equilibria always yield strictly higher payoff than do the no delegation equilibria which is crucial in demonstrating that the latter cannot be well-supported. Indeed, the analysis leads naturally to the conjecture that if there exists a Nash equilibrium outcome of the principals-only game in which *(i)* the delegating principal receives a payoff strictly greater than her subgame perfect equilibrium payoff, and *(ii)* the outside party behaves sequentially rationally, then in any well-supported equilibrium, the principal will certainly choose to delegate rather than playing the game himself, provided that the cost of hiring an agent is low enough. In the rest of the paper we shall demonstrate that this conjecture holds true in the case of a very large class of finite extensive form principals-only games.

[10] Another refinement concept suggested by Hillas (1994), which is similar to the well-supportedness notion, would also lead to the same conclusion. However, we will continue to use the well-supportedness concept in this paper because of its simplicity, weakness, and intuitive appeal in delegation games.

[11] We should note that the second type of equilibria do not survive iterated elimination of weakly dominated strategies and hence do not form a strategically stable set. This should come as no surprise because well-supportedness, like other similar "mistaken theory" refinements, take the equilibrium concept more seriously, whereas elimination of weakly dominated strategies is based on the possibilities of simple mistakes by the other players. Yet, if one slightly perturbs the game so that there is a small probability that the offered contract will be observed, then delegation equilibrium survives the iterated elimination of weakly dominated strategies. It is also possible to show that all perfect Bayesian equilibria mentioned in the text pass the "never weak best response" test.

3 Preliminaries

3.1 Basic Nomenclature

In what follows we shall formally introduce several rudimentary notions of the theory of extensive games with imperfect information. While our treatment borrows from the formulation presented in Osborne and Rubinstein (1994), it also covers infinite games. However, since there is no general theory of perfect Bayesian equilibrium in the case of infinite games, certain aspects of our treatment is not standard.

A finite-horizon *extensive game with perfect recall* (and without exogenous chance moves) is defined to be a collection of the form

$$\Upsilon = [N, H, P, (\mathcal{I}_i, \pi_i)_{i \in N}].$$

Here N denotes a finite *set of players*, and H stands for a comprehensive set of finite sequences interpreted as the *set of all histories*.[12] A history h is said to be *terminal* (or a *pure outcome*) if $(h, a) \notin H$ for any $a \neq \emptyset$; we denote by Z the set of all terminal histories. While the function $\pi_i : Z \to \mathbb{R}$ is the *payoff function* of player i, the function $P : H \backslash Z \to N$ is the *player function*. If $P(h) = i$, we understand that i moves immediately after history h and chooses an action from the set $A(h) \equiv \{a \neq \emptyset : (h, a) \in H\}$. For each i, $\mathcal{I}_i$ is a partition of $H(i) \equiv \{h \in H : P(h) = i\}$ such that $A(h) = A(h')$ whenever $h, h' \in I \in \mathcal{I}_i$. Consequently, without ambiguity, we may write $A(I)$ ($P(I)$, resp.) instead of $A(h)$ ($P(h)$, resp.) for any $h \in I \in \mathcal{I}_i$. Any member of $\mathcal{I}_i$ is called an *information set* for player i. We assume that Υ is with *perfect recall*, that is, any two histories in $I \in \mathcal{I}_i$ prescribe the same sequence of information sets and actions taken at them for player i.

If all information sets of all players are singletons, we say that Υ is a game with *perfect information*, and omit information partitions in defining Υ. The subgames of Υ are defined in the usual way. Given any information set I and any history h, we adopt the following notation:

$$\mathcal{H}(h) \equiv \text{ the set of all histories that is consistent with } h$$

and

$$\mathcal{I}_i(I) \equiv \text{ the set of all information sets of } i \text{ that follows } I.$$[13]

Our definition so far is too general to be useful in applications, we need to introduce some technical structure to H. By far the most common way of doing this is to assume that

[12] By *comprehensiveness* of H, we mean that $\emptyset \in H$, and, for any integer $k \geq 1$, $(a^1, ..., a^k) \in H$ whenever $(a^1, ..., a^{k+1}) \in H$. The *length* of a history $h = (a^1, ..., a^k)$ is defined to be k and denoted by $|h|$. As a convention, we take $|\emptyset| = 0$ and let $(h, \emptyset) = h$ for any $h \in H$.

[13] Formally speaking, $h'' \in \mathcal{H}(h)$ iff $h'' = (h, h')$ for some h', and $\mathcal{I}_i(I) \equiv \{I' \in \mathcal{I}_i : h'' \in I'$ iff $h'' = (h, h')$ for some $h \in I\}$. Notice that $I \in \mathcal{I}_i(I)$.

H is finite. The class of games that is the subject of this paper, however, contain games with infinite action spaces by necessity, and hence we cannot safely assume that H is finite. Instead, we shall posit here that H is a Borel space with σ-field $\mathcal{F}_H$,[14] and refer to any game Υ which abide to this restriction simply as an *extensive game.* If H is finite, then Υ is called a finite extensive game. In what follows, we shall formulate the basic notions of extensive games in such a manner that they coincide with the standard formulation when H is finite and in a manner that is particularly suitable for our present purposes.

Fix an extensive game Υ, and note that any subset H' of H is also a Borel space (with σ-field $\mathcal{F}_{H'} \equiv \{A \cap H' : A \in \mathcal{F}_H\}$). We can thus consider any information set I as a Borel space, and denote the set of all Borel probability measures defined on $\mathcal{F}_I$ by $\mathcal{M}(I)$. Similarly, we consider $A(I)$ as a Borel space, and denote by $\mathcal{P}_s(A(I))$ the set of all (simple) probability measures on $A(I)$ with finite support. We define a *behavioral strategy* for player i as a set

$$\beta_i \equiv \{\beta_i[I] \in \mathcal{P}_s(A(I)) : I \in \mathcal{I}_i\}$$

such that $\{\beta_i[I'] : I' \in \mathcal{I}_i(I)\}$ is a set of independent probability measures for any I such that $|\mathcal{I}_i(I)| < \infty$. Notice that, to avoid measure-theoretic complications, we do not allow $\beta_i[I]$ to be any Borel measure on $A(I)$, but rather confine our attention to simple probability measures. Moreover, provided that there are finitely many of them, a player is assumed to randomize independently at her information sets that follow an information set I. While this condition is too restrictive to allow for a broad theory of a sequential equilibrium concept for infinite games, it will prove to be very useful in our present context. Obviously, it guarantees that our definition of a behavioral strategy coincides with the standard one in the case of finite games. It also allows us to use Kuhn's outcome-equivalence theorem in all finite subgames of a given extensive game with perfect recall.

One may write $\beta_i[h]$ for $\beta_i[I]$ for any $h \in I$ with the understanding that for any histories h and h' that belong to the same information set, we have $\beta_i[h] = \beta_i[h']$. The set of all behavioral strategies of a player is denoted $S_i(\Upsilon)$ and we let $S(\Upsilon) \equiv \times_{i\in N} S_i(\Upsilon)$. If Υ^* is a subgame of Υ, then the restriction of strategy profile $\beta \in S(\Upsilon)$ to this subgame is denoted $\beta|_{\Upsilon^*}$. Clearly, $\beta|_{\Upsilon^*} \in S(\Upsilon^*)$. If $\beta_i[I]$ has a single mass point, then we may identify $\beta_i[I]$ with its mass point. If $\beta_i[I]$ has a single mass point for all $I \in \mathcal{I}_i$, then we refer to β_i as a *pure strategy,* and denote the set of all pure strategies of player i by $S_i^{\text{pure}}(\Upsilon)$. Any member of the set

$$M_i(\Upsilon) \equiv \mathcal{P}_s(S_i^{\text{pure}}(\Upsilon))$$

is then called a *mixed strategy.* A profile $\sigma \in M(\Upsilon) \equiv \times_{i\in N} M_i(\Upsilon)$ is said to be a *Nash equilibrium* in mixed strategies if playing σ_i is a best response to σ_{-i} for all $i \in N$. We

[14] Formally speaking, $(H, \mathcal{F}_H)$ being a Borel space is equivalent to saying that there exists a measurable bijection $\phi : H \to [0,1]$ such that ϕ^{-1} is measurable. (That is, by definition, a Borel space is measurably isomorphic to $([0,1], \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-field.) It is easily checked that if H is finite, then $\mathcal{F}_H = 2^H$.

denote the set of all Nash equilibria of Υ by $NE(\Upsilon)$. Provided that the game Υ is with perfect information, the set of all *subgame perfect equilibria* of Υ in mixed strategies, denoted $SPE(\Upsilon)$, is defined in the usual way.

We define a *system of beliefs* as the set $\mu \equiv \{\mu[I] \in \mathcal{M}(I) : I \in \mathcal{I}_i \text{ for some } i\}$. Notice that, as opposed to the definition of behavioral strategies, we impose no restrictions on the definition of beliefs as probability measures. A finite-support restriction will be deduced rather from the structure of behavioral strategies. We denote the set of all systems of beliefs by $B(\Upsilon)$. If Υ^* is a subgame of Υ, then the restriction of μ to this subgame is denoted $\mu|_{\Upsilon^*}$. A 2-tuple $(\beta, \mu) \in S(\Upsilon) \times B(\Upsilon)$ is called an *assessment.*

Let $h \in H\backslash Z$ be any non-terminal history, and recall that $\mathcal{H}(h)$ stands for the set of all histories consistent with h. Given any strategy profile β, we define the function $p[\beta \,|\, h] : \mathcal{H}(h)\backslash\{h\} \to [0,1]$ as follows: $p[\beta \,|\, h](h, a^1) \equiv \beta_{P(h)}[h](a^1)$ and, for each $k \geq 2$,

$$p[\beta \,|\, h](h, a^1, ..., a^k) \equiv \beta_{P(h)}[h](a^1) \prod_{j=1}^{k-1} \beta_{P(h,a^1,...,a^j)}[h, a^1, ..., a^j](a^{j+1}).$$

Of course, for any $h' \neq \emptyset$, we interpret $p[\beta \,|\, h](h, h')$ as the probability of observing history (h, h'), conditional on h being reached and from there on the game being played according to β. Since $supp(\beta_i[I])$ is finite for all $I \in \mathcal{I}_i$ and $i \in N$,

$$supp(p[\beta \,|\, h]) \equiv \{(h, h') \in H : p[\beta \,|\, h](h, h') > 0\}$$

is a finite set. Consequently, we may say that an assessment (β, μ) is *consistent* if, for all $I \in \mathcal{I}_i$ and all $B \in \mathcal{F}_I$,

$$\mu[I](B) = \frac{\sum_{B \cap supp(p[\beta \,|\, \emptyset])} p[\beta \,|\, \emptyset](h)}{\sum_{I \cap supp(p[\beta \,|\, \emptyset])} p[\beta \,|\, \emptyset](h)},$$

whenever the denominator does not vanish (that is, I is reached via β with positive probability).

To be able to define sequential rationality, we need to deduce an outcome function from a given strategy profile. To this end, we use $p[\beta \,|\, h]$ to define the function $O[\beta \,|\, h] : Z \to [0,1]$ as

$$O[\beta \,|\, h](z) \equiv \begin{cases} p[\beta \,|\, h](z), & \text{if } z \in Z \cap \mathcal{H}(h) \\ 0, & \text{otherwise.} \end{cases}$$

We view $O[\beta \,|\, h]$ as the probability distribution over terminal nodes that will be reached if each player i plays according to β_i, conditional on h being reached. Given any strategy profile β, we define the *expected payoff of player i conditional on history h being reached* as

$$\Pi_i(\beta \,|\, h) \equiv \sum_{z \in supp(O[\beta \,|\, h])} O[\beta \,|\, h](z)\pi_i(z).$$

To simplify the notation, we write $\Pi_i(\beta)$ for $\Pi_i(\beta \,|\, \emptyset)$, the expected payoff of player i induced by β in the entire game.

Notice that $\Pi_i(\beta \,|\, \cdot)$ is a Borel measurable function on each information set I. We may thus define the *expected payoff of player i conditional on his/her information set I* being reached as

$$\Pi_i(\beta, \mu \,|\, I) \equiv \int_I \Pi_i(\beta \,|\, h)\, \mu[I](dh).$$

In turn, an assessment (β, μ) is said to be *sequentially rational* if, for all $i \in N$ and all $I \in \mathcal{I}_i$,

$$\Pi_i(\beta, \mu \,|\, I) \geq \Pi_i((\beta_i', \beta_{-i}), \mu \,|\, I) \quad \text{for all } \beta_i' \in S_i(\Upsilon).$$

An assessment which is both sequentially rational and consistent is called a *perfect Bayesian equilibrium*. We denote the set of all perfect Bayesian equilibrium of Υ by $PBE(\Upsilon)$. The set of all perfect Bayesian equilibria of Υ that reaches to the information set I with positive probability is denoted $PBE(\Upsilon; I)$.

3.2 Well-Supported Bayesian Equilibria

In this subsection we shall introduce a refinement of perfect Bayesian equilibria that will play a central role in this paper. This refinement is informally discussed in Section 2 and is based on imposing certain restrictions on the out-of-equilibrium beliefs and strategies of the players. It is thus in the same spirit with those proposed by McLennan (1985) and Hillas (1994). Informally put, this refinement leads us to those "well-supported" perfect Bayesian equilibria which envisage that at each off-the-equilibrium information set, the beliefs and behavior of the players are consistent with at least one perfect Bayesian equilibrium that admits this information set on its equilibrium path, provided that such an equilibrium exists.

In what follows, we shall give the formal definition of our refinement as it applies only to the class of all extensive form games which do not possess a proper imperfect information subgame. We denote this class of games by $\mathcal{G}$. Our formal treatment is distilled to its simplest form in the context of such games thereby making the intuition behind the refinement proposed here transparent. Moreover, the set of games that will be our focus in this paper is a subset of $\mathcal{G}$, so studying the notion of well-supported equilibrium on $\mathcal{G}$ is without loss of generality with respect to our present purposes. (The bargaining-delegation example studied in the previous section, for instance, belongs to $\mathcal{G}$.) Nevertheless, to be able to evaluate the potential and plausibility of an equilibrium refinement, one may need to understand the implications of it with respect to arbitrary games. For this reason, we shall provide the formal definition of our refinement concept as it applies to the class of all extensive form games in the appendix of this paper.[15]

[15] The generalization given in the appendix is inductive and admits the definition we present in this

Before stating our definition we need to introduce some more notation. Let Υ be any extensive form game which does not possess a proper imperfect information subgame, and consider a behavioral strategy profile $\beta \in S(\Upsilon)$. We define $\mathcal{I}_i(\beta)$ as the set of all information sets of player i which are reached by β with positive probability. On the other hand, $\mathcal{J}(\beta)$ stands for the set of all nonsingleton information sets which could be reached with the shortest sequence of actions after a deviation from $\beta_{P(\emptyset)}[\emptyset]$, while they are surely not reached when $\beta_{P(\emptyset)}[\emptyset]$ is played.[16]

We are now ready to define the well-supported equilibria on the class $\mathcal{G}$:

Definition. Let $\Upsilon \in \mathcal{G}$. A perfect Bayesian equilibrium $(\beta, \mu) \in PBE(\Upsilon)$ is said to be **well-supported** if, and only if, for each $I \in \mathcal{J}(\beta)$ one of the following statements holds:

(a) $PBE(\Upsilon; I) = \varnothing$,

(b) There exists a $(\beta', \mu') \in PBE(\Upsilon; I)$ such that

$$\mathcal{I}_j(\beta') \subseteq \mathcal{I}_j(\beta) \quad \text{for some } j \neq P(\emptyset), \tag{1}$$

(c) There exists a $(\beta', \mu') \in PBE(\Upsilon; I)$ such that (1) does not hold, and

$$\mu[I'] = \mu'[I'] \quad \text{and} \quad \beta_{P(I)}[I'] = \beta'_{P(I)}[I'] \tag{2}$$

for all $I' \in \mathcal{I}_{P(I)}(I)$.

We denote the set of all well-supported perfect Bayesian equilibria of Υ by $PBE_{\text{w-s}}(\Upsilon)$.

To understand this definition intuitively, let us take a two-player game $\Upsilon \in \mathcal{G}$. Suppose that player 1 moves first in this game and suppose that I is the first nonsingleton information set of player 2 which is on the out-of-equilibrium path (i.e., $I \in \mathcal{J}(\beta)$).[17] Player 1's move may indeed be a part of a perfect Bayesian equilibrium (β, μ), provided that it is suitably supported by beliefs and the continuation strategy of player 2 at the off-equilibrium information set I. Since in a perfect Bayesian equilibrium we may choose the beliefs at I arbitrarily (which essentially means that if I is ever reached, player 2 could interpret this as a "mistake"), justifying the move of player 1 is quite easy (except in certain trivial cases involving

subsection as the first step of the induction process. We shall clarify this connection by means of an example (Example A.1) in the appendix.

[16] To define $\mathcal{J}(\beta)$ formally, define $S(\beta_{P(\emptyset)}[\emptyset])$ as the set of all behavioral strategy profiles in which the first mover in the game plays $\beta_{P(\emptyset)}[\emptyset]$ at the initial node. Define next, $\mathcal{Q}(\beta_{P(\emptyset)}[\emptyset])$ as the set of all nonsingleton information sets that do not belong to $\cup_{i \in N} \cup_{\beta' \in S(\beta_{P(\emptyset)}[\emptyset])} \mathcal{I}_i(\beta')$. We have $I \in \mathcal{J}(\beta)$ if and only if $I \in \mathcal{Q}(\beta_{P(\emptyset)}[\emptyset])$ and for any $J \in \mathcal{Q}(\beta_{P(\emptyset)}[\emptyset])$ there do not exist h' and $h'' \neq \emptyset$ such that $h' \in J$ and $(h', h'') \in I$.

[17] By "first" information set we mean the information set that is reached with the shortest sequence of actions after a deviation. Of course, there may be more than one such information set if they are reached via action sequences which are not subsequences of each other (i.e., it is possible that $|\mathcal{J}(\beta)| > 1$).

domination). This arbitrariness is, however, unacceptable if, for example, one subscribes to the view that player 1's deviation should be interpreted by player 2 as a signal about player 1's future behavior (Kohlberg, 1990).[18] Like many other refinements, the concept of well-supported equilibrium postulates certain "reasonable" restrictions on out-of-equilibrium beliefs and actions.

How will player 2 reason when he finds himself at I which was not supposed to be reached in the equilibrium that is being played? It is quite plausible that she will think that player 1 is after coordinating on a different equilibrium (than what was supposed to be the "norm" before), provided that an equilibrium that would indeed lead the play to reach I with a positive probability exists (i.e., $PBE(\Upsilon; I) \neq \emptyset$).[19] If $PBE(\Upsilon; I) = \emptyset$, then there is no plausible explanation of 1's deviation, and hence no restriction is imposed on 2's beliefs and behavior at I. On the other hand, if there exists exactly one such equilibrium, then (β, μ) is well-supported only if the beliefs and the strategy of player 2 from this information set on accord with what is specified by (β, μ). (This is precisely the requirement embodied in (2).) In this case, as we shall illustrate below, the notion of well-supportedness comes close in spirit and in outcome to the forward induction criterion of van Damme (1989) which postulates that a deviation is a viable signal if there is a unique continuation that makes the deviation profitable for the deviating party. (See Example 1 below.) Alternatively, if there exist more than one such equilibrium, then (β, μ) is well-supported only if the continuation beliefs and the strategy of player 2 agrees with that of at least one such equilibrium.

When there are more than two players, however, one has to be more careful. Suppose, player 1 deviates from equilibrium A and the play reaches an out-of-equilibrium information set I of player 2. Also suppose that there is only one equilibrium, equilibrium B, which reaches I. It is possible that all the information sets of player 3 reached in equilibrium B are also reached in equilibrium A. In this case, player 3 will not be able to discern a deviation and hence we do not put any restrictions on her information sets. However, if player 2 has an incentive to play according to equilibrium B only if player 3 plays according to equilibrium B as well, then putting restrictions on player 2's information set I would be unwarranted as well and may indeed lead to suboptimal behavior on the part of player 2. (See Example 2 below.) Part (b) in the above definition attempts to capture this situation. This property

[18] This is in fact the basic idea behind *forward-induction* based refinements of sequential equilibria; see van Damme (1989), Kohlberg (1990), Ponnsard (1991) and Al-Najjar (1995), among others.

[19] As noted earlier, Kreps (1989) calls examining deviations by such a reasoning the "mistaken theory" approach. A version of this approach was first developed by McLennan (1985) who argued that "deviations from the equilibrium path are more probable if they can be explained in terms of some confusion over which sequential equilibrium is "in effect"." (McLennan, 1985, p. 891; also quoted in Kreps, 1989 and Hillas, 1994). The approach followed by Hillas (1994) is also motivated by this sort of a reasoning and thus relates closely to our refinement.

together with putting restrictions only on *some* information sets of *some* players makes well-supportedness a fairly weak refinement concept.

Remarks. ***(i)*** A unique perfect Bayesian equilibrium of a $\Upsilon \in \mathcal{G}$ is always well-supported. More generally, if all perfect Bayesian equilibria reach the same information sets of at least one player $j \neq P(\emptyset)$ with positive probability, then all equilibria of Υ are well-supported.

(ii) If Υ is with perfect information, then all perfect Bayesian equilibria (that is, all subgame perfect equilibria) are well-supported.

(iii) Any perfect Bayesian equilibria of a $\Upsilon \in \mathcal{G}$ that reaches to all information sets of at least one player $j \neq P(\emptyset)$ with positive probability is well-supported.

3.3 Illustrative Examples

Perhaps the best way of evaluating the refinement concept we introduce here is to observe it in action. Thus, we shall next investigate the performance of well-supported equilibria in a number of interesting games that are commonly studied in the equilibrium refinement literature and relate it to refinements which we think come closest in spirit to the notion of well-supportedness.

Example 1.[20] (*Battle of the Sexes with an Outside Option*) Consider the game $\Upsilon \in \mathcal{G}$ depicted in Figure 2. In this game there are two types of equilibria; one in which $\beta_1[\emptyset](O) = 1$ and one in which $\beta_1[\emptyset](T) = 1$. For the first type of equilibria, we must have either $\beta_2[I](R) = 1$ and $\mu[I](T) < 3/4$ or $\beta_2[I](L) \leq 2/3$ and $\mu[I](T) = 3/4$. The unique second type of equilibrium, however, has it that $\beta_2[I](L) = 1$ and $\mu[I](T) = 1$. For the first type of equilibria we have $\mathcal{J}(\beta) = \{I\}$, that is, I is the only out-of-equilibrium information set to check. Since only the second type of equilibrium reaches I, if a first type of equilibrium is well-supported, it must have $\beta_2[I](L) = 1$ and $\mu[I](T) = 1$, that is, part (c) of the definition must hold. Since this is not the case, we conclude that none of the first type of equilibria is well-supported. The unique well-supported equilibrium outcome of Υ is then (T, L) with the payoff profile $(3, 1)$. Notably, this is also the unique outcome that passes the forward induction test of van Damme (1989).[21] $\|$

[20] This example is due to Kohlberg and is probably the most commonly used game in motivating the basic idea behind the notion of forward induction. It is also experimentally analyzed by Cooper et al. (1993) who provide mixed evidence in support of the forward induction hypothesis.

[21] The main example analyzed by van Damme (1989. pp. 485-87) shows that strategically stable equilibria (Kohlberg and Mertens, 1986) does not satisfy his forward induction criterion. Since the requirement of well-supportedness chooses precisely the equilibrium chosen by van Damme's forward induction concept in his example (in which the moves of player 1 are coalesced), we may also conclude that strategic stability does not imply well-supportedness. Conversely, it has been shown in Section 2 that a well-supported equilibrium

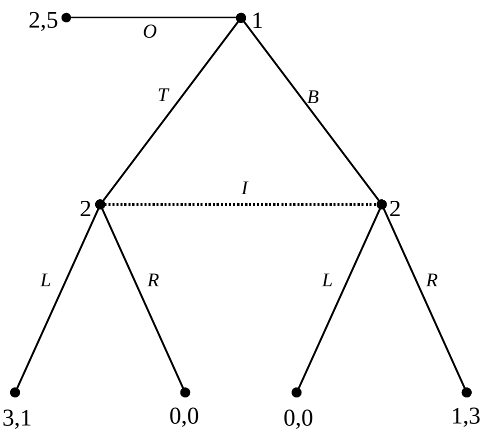

Figure 2: Battle of the Sexes with an Outside Option

Example 2.[22] (*A Variant of Selten's Horse*) Consider the game $\Upsilon \in \mathcal{G}$ given in Figure 3. There are three perfect Bayesian equilibria of this game:

(1) $\beta_1[\emptyset](D) = 1$, $\beta_2[I_2](a) = 1$, $\beta_3[I_3](L) = 1$;

(2) $\beta_1'[\emptyset](A) = 1$, $\beta_2'[I_2](d) = 1$, $\beta_3'[I_3](R) = 1$;

(3) $\beta_1''[\emptyset](D) = 1/3$, $\beta_1''[\emptyset](A) = 2/3$, $\beta_2''[I_2](d) = 1/2$, $\beta_3''[I_3](L) = 1/2$.

First, notice that, equilibria (2) and (3) reach every information set in the game so that they are well-supported. The only out-of-equilibrium information set for equilibrium (1) is I_2, and since the other two equilibria both reach that information set, for this equilibrium to be well-supported either part (b) or (c) of the definition has to be satisfied. While it can be checked that (c) is not satisfied, we have $\mathcal{I}_3(\beta) = \{I\} = \mathcal{I}_3(\beta') = \mathcal{I}_3(\beta'')$ so that (b) is satisfied, and therefore equilibrium (1) is well-supported as well. It must be noted that this equilibrium would be eliminated by means of certain other reasonable belief-based refinements (cf. Hillas, 1994). Our refinement, therefore, does not appear to be particularly demanding in this instance.[23] ||

need not be stable.

[22] This example is based on a simple modification of an example given in Hillas (1994).

[23] To see the intuition behind the well-supportedness of equilibrium (1), let us reason with player 2 who finds herself unexpectedly at information set I_2 following a deviation by player 1 from D to A: "It must be the case that I was mistaken in thinking that equilibrium (1) was going to be played and that player 1 must be after either equilibrium (2) or (3). If this is the case and if player 3 also realizes this and plays according to equilibrium (2) or (3) then I should also play d as this would give me either 2 or 1 whereas playing a gives me 1. But, is there a way that player 3 could actually realize that we are in a different equilibrium once I play d? No, for player 3's information set is reached under both equilibrium (1) and if I now play d. Therefore, player 3 has no way of realizing that we have deviated from equilibrium (1) and might continue playing according to that equilibrium which would give me a payoff of zero. Therefore, I am better of by playing a."

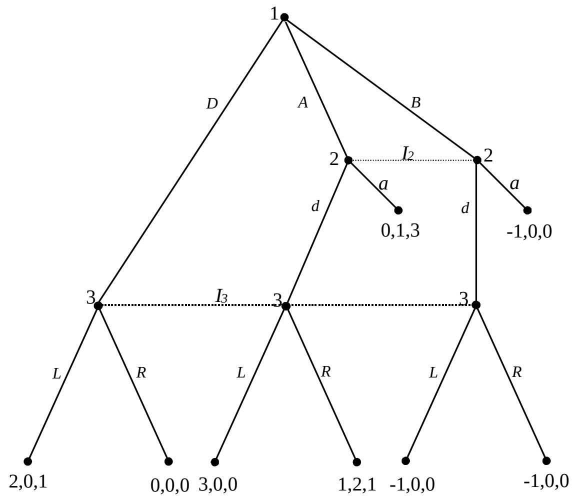

Figure 3: A Variant of Selten's Horse

Example 3. (*Burning Money*) Consider the game given in Figure 4 in which two players interact while playing a game of "Battle of the Sexes" after player 1 declares (in an observable manner) whether or not he has chosen to give up one dollar at the beginning of the game. This game provides a classical illustration of the power of forward induction and is first analyzed by van Damme (1989) and Ben-Porath and Dekel (1992). It is well known that the only equilibrium that survives an iterative application of forward induction (along with backward induction) is the one with $\beta_1[\emptyset](nb,T) = \beta_2[I_2](L) = 1$ (supported with out-of-equilibrium strategy $\beta_2[I_1](L) = 1$.) Our refinement concept is not strong enough to entail this conclusion. Indeed, along with this equilibrium, the equilibria in which $\beta_1[\emptyset](b,T) = \beta_2[I_1](L) = 1$ are also well-supported.[24] $\|$

These examples suggest that the notion of well-supportedness is a reasonable refinement of the perfect Bayesian equilibria. Its demand for rationality is not overwhelming, and it is quite a weak refinement concept. We believe that at least for games in $\mathcal{G}$, there is good reason to regard ill-supported equilibria as unlikely to be realized.

Of course, we must admit that well-supportedness is merely another way of refining equilibria, and the literature provides a disconcertingly large number of other ways of doing this. Rather than arguing for the superiority of our refinement over others, here we subscribe to the view that "the validity of a particular refinement for the analysis of a particular economic issue may depend on the setting of that issue in ways that go beyond the formal

[24] These last two equilibria are, however, only supported by off-the-equilibrium beliefs that constitute parts of *ill-supported* equilibria. Thus, if we ask for the supporting off-the-equilibrium beliefs to be taken from only the well-supported equilibria (which is equivalent to applying the well-supportedness test twice), then both of these equilibria would be eliminated. Yet, in this paper we shall not need to iterate our refinement.

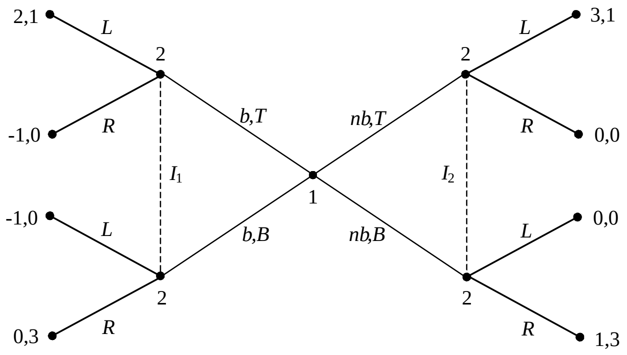

Figure 4: Burning Money

game-theoretic model that is adopted" (Kreps, 1989, p. 7.) The main objective of this paper is to use the concept of "well-supportedness" in an economic setting in which this refinement is particularly sensible and only mildly demanding. Moreover, in this context, we shall see that it allows one to obtain considerable insight with regard to the underlying economic problem. We turn next to describing the general economic framework that we will subsequently analyze.

4 One-Sided Delegation Environments

In this section we shall introduce a general environment in which we shall study the problem of delegation by unobservable incentive contracts. As one might expect, the framework we outline below admits the simple bargaining-delegation model studied in Section 2 as a special case.

We begin by fixing an arbitrary finite perfect information *principals-only* game

$$\Gamma = [\{1, 2\}, H, P, (\pi_1, \pi_2)].$$

We assume that this game has a unique subgame perfect equilibrium outcome, which we consider as the *status quo* outcome of the environment prior to delegation. (In fact, it is possible to generalize the main results of the paper to the case of arbitrary finite extensive games with a unique perfect Bayesian equilibrium outcome; see Section 6.) We denote the (status quo) expected payoff of player i in equilibrium by Π_i^{SPE}. The set of all Nash equilibrium payoffs of i is in turn denoted by $\mathbf{\Pi}_i^{\mathrm{NE}}(\Gamma)$; that is, $\mathbf{\Pi}_i^{\mathrm{NE}}(\Gamma) \equiv \{\Pi_i(\sigma) : \sigma \in NE(\Gamma)\}$. In what follows, we assume that $\mathbf{\Pi}_i^{\mathrm{NE}}(\Gamma)$ is a finite set for each $i = 1, 2$.[25]

[25] Finiteness of $\mathbf{\Pi}_i^{\mathrm{NE}}(\Gamma)$ is an assumption that helps us avoid some technical redundancies. It is a very weak assumption in that it is generically true that a finite game has finitely many Nash equilibria (see, for instance, Moulin, 1986).

We also need to define a subset of $NE(\Gamma)$ and this can be best done in terms of behavioral strategies. Let $NE_i^*(\Gamma)$ denote the set of Nash equilibria of Γ in which the behavioral strategy of player i is sequentially rational after any history. More formally, define $NE_i^*(\Gamma)$ as follows: a behavioral strategy profile $\beta^* \in NE_i^*(\Gamma)$ if, and only if, $\beta^* \in NE(\Gamma)$ and

$$\beta_i^* \in \arg \max_{\beta_i \in S_i(\Gamma)} \Pi_i(\beta_i, \beta_{-i}^* | h) \qquad \text{for all } h \in H(i).$$

It is easily seen that in any $\beta^* \in NE_i^*(\Gamma)$, β_{-i}^* is sequentially rational after any history reached with positive probability with β^*, whereas β_i^* is sequentially rational even after those histories which are not reached under β^*. The set $NE_i^*(\Gamma)$ in terms of mixed strategies is defined in the usual manner. Also let $\mathbf{\Pi}_j^{\mathrm{NE}_i^*}(\Gamma) \equiv \{\Pi_j(\beta) : \beta \in NE_i^*(\Gamma)\}$.

Let us now posit that player 2 is contemplating about hiring an agent to play the game Γ in place of him. The *outside option* of this agent, henceforth called player A, is denoted by δ. That is, player A receives δ dollars with certainty if she does not accept the contract offered by player 2. We assume that $\delta > 0$. Moreover, we assume that player 2 incurs a contracting cost of $c > 0$ dollars in case he decides to delegate.[26]

In the literature on strategic contract design, a contract is sometimes defined as a mapping that specifies a level of payment for each *strategy* of the agent in $S_2(\Gamma)$. Such a contract provides an extensive amount of control to the principal, and usually simplifies the analysis considerably. However, especially when mixed strategies are allowed, this definition would lead one to view a contract as an unrealistically complicated object (that lives in $\mathbb{R}^{S_2(\Gamma)}$). Moreover, it is not at all clear how a principal could in general "observe" the randomized strategy choice of an agent, which he must to be able to submit the agent's compensation. At the very least, this would necessitate to extend the model to account for private monitoring of the agent.

Consequently, in this paper, by a *contract*, we mean a function that maps the finite set of terminal nodes Z of the game Γ to the set of payments. Thus, the *contract space* in our model is the finite-dimensional space $\mathbb{R}^Z$, and we focus here on *incentive contracts* which can be conditioned only on the pure outcomes of the game rather than the delegate's strategy. While this introduces a number difficulties regarding the formal analysis, it brings us a step closer to realism and avoids worrying about issues related to the "monitoring" of the agents since the pure outcomes of the game Γ are observable.

The primitives of our model is thus the game Γ, the outside option $\delta > 0$ and the contracting cost $c > 0$. We thus refer to the 3-tuple $[\Gamma, \delta, c]$ as a *one-sided delegation environment.* Such an environment induces a delegation game

$$\Lambda(\Gamma, \delta, c) \equiv [\{1, 2, A\}, H^*, P^*, (\mathcal{I}_i^*, \pi_i^*)_{i \in \{1,2,A\}}]$$

[26] While the proofs of our results use the assumption that $\delta > 0$, they are easily modified to account for the case $\delta = 0$ as well.

which is a 3-person extensive form game. The game begins with player 2 deciding between taking the action of not delegating (denoted $\neg D$) and an action of attempting to delegate by offering a contract $f \in \mathbb{R}^Z$, which the agent A may accept (denoted a) or reject (denoted r). If player 2 chooses not to delegate, or he chooses to offer a contract but this contract is rejected by A, then Γ is played by players 1 and 2. But if player 2 offers a contract f which agent A accepts, then, agent A plays the game Γ with player 1. Consequently, we have

$$H^* \equiv (\{\neg D\} \times H) \cup \left(\mathbb{R}^Z \times \{a, r\} \times H\right).$$

and the player function P^* is defined as:

$$P^*(\varnothing) \equiv 2,\ P^*(f) \equiv A,\ P^*(\neg D, h) \equiv P^*(f, r, h) \equiv P(h) \text{ and } P^*(f, a, h) \equiv \begin{cases} A, & P(h) = 2 \\ 1, & P(h) = 1 \end{cases}$$

for all $f \in \mathbb{R}^Z$ and $h \in H$. We note that, for any terminal history $z^* \in Z^*$ in $\Lambda(\Gamma, \delta, c)$, there exists a unique terminal history $z \in Z$ in Γ such that $z^* \in \{(\neg D, z), (f, r, z), (f, a, z)\}$. In what follows, we shall refer to this terminal history z as a *pure outcome induced by* z^* in Γ.

Due to the infiniteness of the contract space, H^* is not a finite set. Yet, we can make H^* a Borel space by using the Borel σ-field of $\mathbb{R}^{|Z|}$ in the obvious way. $\Lambda(\Gamma, \delta, c)$ can thus be viewed as an extensive game as defined in Section 3.1. It is crucial to recognize that this game is with imperfect information precisely because the contracts signed between player 2 and agent A are unobservable to player 1. While player 1 observes whether or not a contract is accepted, that is, he always knows the *identity* of her opponent, he does not observe which contract is accepted (or rejected). Hence, once a contract is accepted, player 1 does not know the payoff function of his opponent (i.e. of player A). Thus, the information partition of player 1 is

$$\mathcal{I}_1^* \equiv \{\{(\neg D, h)\} : h \in H(1)\} \cup \{\{(f, \theta, h) : f \in \mathbb{R}^Z\} : \theta \in \{a, r\},\ h \in H(1)\}.$$

Players 2 and A, on the other hand, possess perfect information throughout the game so that all of their information sets are singletons. That is, we have

$$\mathcal{I}_2^* \equiv \{\emptyset\} \cup \{\{(\neg D, h)\} : h \in H(1)\} \cup \{\{(f, r, h)\} : f \in \mathbb{R}^Z,\ h \in H(2)\}.$$

and $\mathcal{I}_A^* \equiv \{\{(f, a, h)\} : f \in \mathbb{R}^Z, h \in H(2)\}$.

Next we need to specify the payoff functions of the players. Since player 1 is not involved with any sort of a delegation activity, we have $\pi_1^*(\neg D, z) \equiv \pi_1^*(f, \theta, z) \equiv \pi_1(z)$ for all $z \in Z$, $\theta \in \{a, r\}$, and $f \in \mathbb{R}^Z$. Similarly, the payoffs of player 2 would not be altered if he chooses not to delegate, that is, $\pi_2^*(\neg D, z) \equiv \pi_2(z)$ for all $z \in Z$. On the other hand, player 2 incurs the cost c if he chooses to offer a contract, and pays the promised compensation to the agent in case a contract is signed. Therefore,

$$\pi_2^*(f,\theta,z) \equiv \begin{cases} \pi_2(z) - f(z) - c, & \text{if } \theta = a \\ \pi_2(z) - c, & \text{if } \theta = r \end{cases}$$

for all $z \in Z$ and $f \in \mathbb{R}^Z$. Finally, the delegate's payoffs are determined as

$$\pi_A^*(\neg D, z) \equiv \pi_A^*(f,r,z) \equiv \delta \quad \text{and} \quad \pi_A^*(f,a,z) \equiv f(z)$$

for all $z \in Z$ and $f \in \mathbb{R}^Z$. This completes the description of the delegation game $\Lambda(\Gamma,\delta,c)$.

We conclude this section by noting that even though $\Lambda(\Gamma,\delta,c)$ is an infinite game, it is quite well-behaved in that it admits well-supported equilibria. The following proposition is proved in Section 7.

Proposition 1. *For any given one-sided delegation environment* $[\Gamma,\delta,c]$, *there exists a well-supported perfect Bayesian equilibrium of the delegation game* $\Lambda(\Gamma,\delta,c)$.

5 The Main Result

Fix a one-sided delegation environment $[\Gamma,\delta,c]$. As noted earlier, our main objective here is to understand the nature of the equilibria of the induced delegation game $\Lambda(\Gamma,\delta,c)$ as it pertains to the implications of the possibility of delegation. Thus, the first question we need to address is if delegation takes place in equilibrium at all, while the second question is if the *status quo* equilibrium outcome of the game Γ is altered, provided that at least some degree of delegation takes place in equilibrium. The literature on delegation since the influential contribution of Katz (1991) exhibits clearly the contention that neither of these questions have an affirmative answer. Indeed, the analysis of Katz (1991) culminates in showing that the Nash equilibrium outcomes of the principals-only game, and only these outcomes, can be reached via unobserved delegation in the Nash equilibria of the delegation game. Consequently, a sequentially rational equilibrium outcome of the delegation game has to be a Nash equilibrium outcome of the principals-only game. More formally, in the present setting, we have

Proposition 2. *If the pure outcome* z^* *is reached with positive probability in a perfect Bayesian equilibrium of the delegation game* $\Lambda(\Gamma,\delta,c)$, *then the pure outcome induced by* z^* *in the principals-only game* Γ *can be reached with positive probability in a Nash equilibrium of* Γ.

Moreover, it can be shown that the set of expected payoffs of the principal 2 (gross of the cost of hiring) obtained in a perfect Bayesian equilibrium of $\Lambda(\Gamma,\delta,c)$ lies in the convex hull of Nash equilibrium payoffs of principal 2 which are at least as large as his subgame perfect equilibrium payoff in game Γ (see Lemma 5 in Section 7). Thus, the possibility of delegation does not alter in a payoff-relevant way the set of *Nash equilibrium outcomes* of the mother game Γ which is played by the principals. It is undeniable that this observation creates severe difficulties for the well-known (perfect information) delegation results that concern simultaneous move principals-only games such as those of Vickers (1985), Fersthman and Judd (1987) and Sklivas (1987). However, its implications become limited when we shift our focus to games with sequential moves. Indeed, this is the main point raised by Fersthman and Kalai (1997) who took Γ to be a finite ultimatum game and showed that a perfect Bayesian equilibrium of $\Lambda(\Gamma, \delta, c)$ need not correspond to a subgame perfect equilibrium of Γ. We show that this observation holds much more generally. In particular, Lemma 5 in Section 7 demonstrates that any Nash equilibrium payoff of the delegating principal in Γ which is at least as large as his subgame perfect equilibrium payoff, and which can be obtained via a sequentially rational strategy of the outside party can also be obtained as a perfect Bayesian equilibrium payoff of the delegation game. That is, delegation *may* alter the outcome of an extensive form game that would obtain in the absence of delegation. In turn, the upshot of the present paper is that even a stronger version of this statement is true: we claim that reasonable requirements of rationality ensure that strategic delegation is bound to alter the outcome of an extensive principals-only game in a way that benefits the delegating principal.

To make things clear, let us formally state the main question we pose here.

Main Query. *Given a one-sided delegation environment* $[\Gamma, \delta, c]$, *are there any well-supported equilibria of* $\Lambda(\Gamma, \delta, c)$ *in which delegation does not take place with positive probability?*

Two comments about this query are in order. First, the requirement of well-supportedness is indeed needed in its statement, for we know from the analysis of Katz (1991) that there exist perfect Bayesian equilibria in which delegation does not obtain. (Recall Section 2.) As noted earlier, our main point in this paper is that such equilibria are unreasonable, and a suitable forward induction and/or out-of-equilibrium behavior restriction argument will eliminate all equilibria that envisage a neutral role for delegation. We propose the notion of "well-supportedness" in order to formalize this point, and examine the implications of the possibility of delegation with respect to well-supported equilibria. Second, the answer to our main inquiry would obviously be yes if $\delta + c$ was too high, that is, if it was simply too costly to hire an agent. This is, however, an uninteresting answer; one clearly needs to examine the issue for small $\delta + c$.

The main result of this paper provides an answer to the query stated above by characterizing the conditions under which delegation obtains in *any* well-supported equilibrium.

Theorem 1. *There exists an* $\ell > 0$ *such that* $PBE_{\text{w-s}}(\Lambda(\Gamma, \delta, c)) \neq \emptyset$ *and*

$$\beta_2[\emptyset](\neg D) = \begin{cases} 1, & \text{if } \Pi_2^{\text{SPE}} = \max \boldsymbol{\Pi}_2^{\text{NE}}(\Gamma) \\ 0, & \text{if } \Pi_2^{\text{SPE}} < \max \boldsymbol{\Pi}_2^{\text{NE}_1^*}(\Gamma) \end{cases}$$

for any well-supported equilibrium $(\beta, \mu) \in PBE_{\text{w-s}}(\Lambda(\Gamma, \delta, c))$ *and any* $\delta + c < \ell$.

But does the presence of delegation imply that the status quo outcome will be altered? The answer is yes. When delegation occurs, this always (strictly) benefits the delegating party, and hence the *status quo* outcome is bound to be altered in a payoff-relevant way in *any* well-supported equilibrium.

Corollary 1. *There exists an* $\ell > 0$ *such that*

$$\Pi_2^*(\beta) \begin{cases} = \Pi_2^{\text{SPE}}, & \text{if } \Pi_2^{\text{SPE}} = \max \boldsymbol{\Pi}_2^{\text{NE}}(\Gamma) \\ > \Pi_2^{\text{SPE}}, & \text{if } \Pi_2^{\text{SPE}} < \max \boldsymbol{\Pi}_2^{\text{NE}_1^*}(\Gamma) \end{cases}$$

for any well-supported equilibrium $(\beta, \mu) \in PBE_{\text{w-s}}(\Lambda(\Gamma, \delta, c))$ *and any* $\delta + c < \ell$.

Consequently, in a well-supported equilibrium, player 2 will choose not to delegate if he is already in an advantageous situation in the principals-only game Γ (in the sense that the *status quo* outcome is already the best that he can achieve in any Nash equilibrium) whereas he will delegate if there is a Nash equilibrium in which he obtains a payoff strictly greater than his subgame perfect equilibrium payoff and in which player 1 plays sequentially rationally. So, for instance, in any Stackelberg duopoly situation, the leader firm will not choose to delegate the decision-making power. On the other hand, by Theorem 1, the follower firm will (generically) choose to delegate the decision-making to an agent *even when the incentive contracts are fully unobservable.* Moreover, the delegation decision will certainly benefit the follower firm. Therefore, in sequential market games, it turns out that there is good reason to take Fersthman-Judd like delegation results seriously even in the presence of unobservable contracts.

While their formal demonstration is a bit tedious, the intuition behind the above results is for the most part identical to the one that is provided by the bargaining-delegation example of Section 2. To be a bit more concrete, we must provide a brief outline of the proof of Theorem 1. The first step is to notice that in any equilibrium of $\Lambda(\Gamma, \delta, c)$, the principal will choose a contract such that the agent A best responds to player 1 according to the preferences of the principal. This is a crucial ingredient in characterizing the perfect Bayesian equilibria of the delegation game. The next step is to observe that for $\delta + c < \ell$, where $\ell \equiv \min\{\alpha \in \boldsymbol{\Pi}_2^{\text{NE}}(\Gamma) : \alpha > \Pi_2^{\text{SPE}}\} - \Pi_2^{\text{SPE}}$, all equilibria involving delegation yields player 2 a net equilibrium payoff which is strictly greater than Π_2^{SPE}. Therefore, player 2 cannot be indifferent between delegating and not delegating in any perfect Bayesian equilibria. The third step is to prove that well-supportedness demands that player 2 chooses $\neg D$ (which must be with probability one by the previous step) only if there exists no equilibrium that involves delegation. To see this, suppose (β^*, μ^*) is a well-supported equilibrium in which player 2 chooses not to delegate with probability one. Therefore, under (β^*, μ^*) player 2's payoff is Π_2^{SPE}. Also suppose there is an equilibrium (β, μ) which involves delegation. From our earlier observations, (β, μ) must be yielding player 2 a payoff strictly greater than Π_2^{SPE}. Also, by definition of well-supportedness, the strategy of player 1 must be the same in β^* and β, following an information set, say I, that is reached by equilibrium β. Now, one can show that player 2 can deviate in strategy profile β^* by offering a contract which, under β_A^*, makes the agent reach information set I with probability 1, and then play in a way such that the net payoff of player 2 is the same with his payoff under strategy profile β. In other words, there is a deviation for player 2 in equilibrium (β^*, μ^*) which gives him a payoff strictly greater than Π_2^{SPE}. This contradicts that (β^*, μ^*) is a perfect Bayesian equilibrium. The fourth and final step is to show that for $\delta + c < \ell$, there exist no equilibria involving delegation if $\Pi_2^{\mathrm{SPE}} = \max \mathbf{\Pi}_2^{\mathrm{NE}}(\Gamma)$, and there exists an equilibrium involving delegation if $\Pi_2^{\mathrm{SPE}} < \max \mathbf{\Pi}_2^{\mathrm{NE}_1^*}(\Gamma)$. Theorem 1 is then established by using the third and fourth steps, and Proposition 1.

6 Extensions and Concluding Remarks

Larger Classes of Principals-Only Games. While we have studied in Section 5 only those perfect information principals-only games with a unique subgame perfect equilibrium outcome, it is easy to generalize the present findings to larger classes of games. For instance, let Γ be any finite extensive form game with a unique perfect Bayesian outcome (and with $\left|\mathbf{\Pi}_2^{\mathrm{NE}}(\Gamma)\right| < \infty$). The proof of Theorem 1 modifies in a trivial manner to show that, whenever $\Pi_2^{\mathrm{PBE}} < \max \Pi_2^{\mathrm{NE}_1^*}$ and

$$0 < \delta + c < \min\{\alpha \in \mathbf{\Pi}_2^{\mathrm{NE}}(\Gamma) : \alpha > \Pi_2^{\mathrm{PBE}}\} - \Pi_2^{\mathrm{PBE}}$$

in *all* well-supported equilibria of $\Lambda(\Gamma, \delta, c)$ player 2 chooses to delegate with probability one.

Two-Sided Delegation. We have assumed above that only one of the parties had the opportunity to delegate. A natural question, therefore, is if the findings reported here would still have a “bite” if both of the principals had an opportunity to delegate. Although their analysis is more complicated, it is easy to see that the following result holds: If there exists a Nash equilibrium of the principals-only game in which *(i)* one of the players obtains a payoff that is strictly greater than his status quo payoff, and *(ii)* the other player behaves sequentially rationally, then, for small $\delta + c$, at least one of the players delegates with positive probability in any well-supported equilibrium. Two-sided delegation games present some interesting possibilities and whether both players delegate with positive probability or not depends on the specifics of the underlying game. For preliminary results along these lines we refer the reader to Koçkesen (1999).

Principal-Agent Bargaining. Another aspect of the present model which could be fruitfully generalized is the bargaining process between the principal and the delegate. We assumed here that principal makes a "take it or leave it" offer to the delegate within a symmetric and complete information context. In reality, of course, these assumptions are rarely valid. An interesting conjecture is that the existence of asymmetric information between the principal and the delegate might restrict the contracts that would be offered in equilibrium in such a manner that Theorem 1 holds true under even weaker refinements of perfect Bayesian equilibrium than what we have proposed. The analysis of this issue, while certainly a promising avenue of research, falls outside the scope of the present paper.

Renegotiation. As we have mentioned earlier, an important assumption in our model is that contracts cannot be renegotiated once the outside party starts taking actions. This could be either due to physical impossibility of renegotiation, as in closed-door negotiations, or because of the high costs associated with renegotiating a contract. It is clear that if renegotiation is costless and can take place at any point throughout the game, then delegation, with observed or unobserved contracts, would have no commitment power. In other words, in any perfect Bayesian equilibrium of the delegation game with costless and unlimited renegotiation, the delegate must behave sequentially rationally from the perspective of the delegating principal. Otherwise, there will exist Pareto improving renegotiation opportunities between the principal and the delegate. In many contexts, the truth probably lies somewhere in between these two extreme cases, i.e., renegotiation is costly and takes place only at certain points in the game as a result of a deterministic or a random process. Although, intuition suggests that in such an environment delegation will still have commitment effects, albeit limited, its full analysis awaits further research.

Experiments. The main findings reported in this paper are based on a particular equilibrium refinement the empirical validity of which must be tested against the data. An obvious way to conduct this test is of course via experiments in which agents play a delegation game such as the simple bargaining game presented in Section 2. Fershtman and Gneezy (1997) conducts an experimental test of strategic delegation under observable and unobservable contracts within the context of such an ultimatum bargaining game. They provide some results which are not really in line with any of the theoretical results in the literature including those in Fershtman and Kalai (1997). We should note, however, that the structure of what Fershtman and Gneezy define as a delegation game is not the same as what we used in this paper. In particular, in their setting contract space is defined differently and the player who delegates does not have the option of playing the game himself. Consequently, testing the empirical validity of the theory we proposed here remains as an integral part of our future research agenda.

7 Proofs

We begin by introducing some notation and preliminary observations. Consider a one-sided delegation environment $[\Gamma, \delta, c]$ and let $(\beta^*, \mu^*) \in PBE(\Lambda(\Gamma, \delta, c))$. Clearly, for each $f \in \mathbb{R}^Z$, the behavioral strategy β_A^* induces a behavioral strategy in the game Γ, which is defined as

$$b_{f,2}^*[h] \equiv \beta_A^*[f, a, h], \quad h \in H(2).$$

Notice that $\{b_{f,2}^*[h] : h \in H(2)\}$ is a set of independent probability measures by definition of a behavioral strategy. Thus, by Kuhn's theorem, there exists a mixed strategy that is outcome-equivalent to $b_{f,2}^*$; we denote this mixed strategy by $\sigma_{f,2}^*$. Similarly, σ_1^* denotes a mixed strategy in $M_1(\Gamma)$ that is outcome-equivalent to the behavioral strategy $b_1^* \in S_1(\Gamma)$, where

$$b_1^*[h] \equiv \beta_1^*[\{(f, a, h) : f \in \mathbb{R}^Z\}], \quad h \in H(1).$$

For any $h \in H$, we denote by $o[b_1^*, b_{f,2}^* \,|\, h]$ the probability distribution over terminal nodes that will be reached if each player plays the game Γ according to the strategy profile $(b_1^*, b_{f,2}^*) \in S(\Gamma)$, conditional on h being reached. The probability distribution $o[\sigma_1^*, \sigma_{f,2}^* \,|\, h]$ is defined analogously. By definition of outcome-equivalence, we have

$$o[\sigma_1^*, \sigma_{f,2}^* \,|\, h] = o[b_1^*, b_{f,2}^* \,|\, h] \equiv O[\beta^* \,|\, f, a, h], \quad h \in H.$$

If $h = \emptyset$, we simply write $o[\sigma_1^*, \sigma_{f,2}^*]$ for $o[\sigma_1^*, \sigma_{f,2}^* \,|\, \emptyset]$. Consequently, we write the expected payoff of the agent who takes on a contract f by

$$F(b_1^*, b_{f,2}^*) = F(\sigma_1^*, \sigma_{f,2}^*) \equiv \sum_{z \in Z} o[\sigma_1^*, \sigma_{f,2}^*](z) f(z).$$

The expected payoff of player 2 (gross of the payment to the agent and the contracting cost) is similarly written as $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) \equiv \sum_{z \in Z} o[\sigma_1^*, \sigma_{f,2}^*](z)\pi_i(z)$. Of course, we have

$$\Pi_2^*(\beta^* \,|\, f) = p\left(\Pi_2(\sigma_1^*, \sigma_{f,2}^*) - F(\sigma_1^*, \sigma_{f,2}^*) - c\right) + (1-p)\left(\Pi_2^{\mathrm{SPE}} - c\right)$$

where $p = \beta_A^*[f](a)$, so that

$$\Pi_2^*(\beta^* \,|\, f) = \Pi_2^*(\beta^* \,|\, f, a) = \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - F(\sigma_1^*, \sigma_{f,2}^*) - c \tag{3}$$

for any f with $\beta_A^*[f](a) = 1$.

We denote by $\sigma^*_{\neg D,1}$ and $\sigma^*_{r,1}$ the mixed strategies for player 1 in Γ that are outcome-equivalent to b'_1 and b''_1, respectively, where $b'_1[h] \equiv \beta^*_1[\{(\neg D, h)\}]$ and $b''_1[h] \equiv \beta^*_1[\{(f, r, h) : f \in \mathbb{R}^Z\}]$ for any $h \in H(1)$. The mixed strategies $\sigma^*_{\neg D,2}$ and $\sigma^*_{r,2}$ are defined analogously. Of course, we have $\{(\sigma^*_{\neg D,1}, \sigma^*_{\neg D,2})\} = \{(\sigma^*_{r,1}, \sigma^*_{r,2})\} = \mathit{SPE}(\Gamma)$.

Proof of Proposition 1. Take any $[\Gamma, \delta, c]$ and denote the game $\Lambda(\Gamma, \delta, c)$ by Λ. If there exists a unique $(\beta, \mu) \in \mathit{PBE}(\Lambda)$, then we have $\mathit{PBE}_{\text{w-s}}(\Lambda) = \mathit{PBE}(\Lambda)$. If, on the other hand, there exist more than one perfect Bayesian equilibria of Λ, then clearly the one with the maximum payoff for player 2 is well-supported by taking beliefs and actions at any out-of-equilibrium information set to be identical to those in one of the equilibria which visits that information set. Therefore, the proposition will be proved if we can show that $\mathit{PBE}(\Lambda) \neq \emptyset$. To this end, let b^{SPE} be the subgame perfect equilibrium behavioral strategy profile, and, for each $f \in \mathbb{R}^Z$, take any $b^f \in S_2(\Gamma)$ such that $b^f[h] = b^{f,h}[h]$ for all $h \in H(2)$, where

$$b^{f,h} \in \operatorname*{arg\,max}_{b_2 \in S_2(\Gamma)} \sum_{z \in Z} o[b_1^{\text{SPE}}, b_2 \,|\, h](z) f(z).$$

Define next the behavioral strategy profile $\beta \in S(\Lambda)$ as follows: For all $f \in \mathbb{R}^Z$,

$$\beta_1[\neg D, h] = \beta_1[f, a, h] = \beta_1[f, r, h] = b_1^{\text{SPE}}[h] \text{ for all } h \in H(1),$$

$$\beta_2[\emptyset](\neg D) = 1,\ \beta_2[\neg D, h] = \beta_2[f, r, h] = b_2^{\text{SPE}}[h] \text{ for all } h \in H(2),$$

$$\beta_A[f](a) = \begin{cases} 1, & \text{if } F(b_1^{\text{SPE}}, b^f) \geq \delta \\ 0, & \text{otherwise} \end{cases}, \text{ and } \quad \beta_A[f, a, h] = b^f[h] \text{ for all } h \in H(2).$$

On the other hand, we define $\mu \in B(\Lambda)$ by specifying that $\mu[I](\pi_2, \theta, h) = 1$ for all $I = \{(f, \theta, h) : f \in \mathbb{R}^Z\} \in \mathcal{I}^*_1$, $\theta \in \{a, r\}$. Since it can be checked that $(\beta, \mu) \in \mathit{PBE}(\Lambda)$, the proof is complete.[27] $\square$

In what follows, we fix a one-sided delegation environment $[\Gamma, \delta, c]$, and denote the delegation game $\Lambda(\Gamma, \delta, c)$ by Λ for brevity. For any equilibrium strategy profile β^* in Λ, we let

$$C(\beta^*_2) \equiv \mathit{supp}(\beta^*_2[\emptyset]).$$

Since $\beta^*_2[\emptyset]$ is a simple probability measure over $\{\neg D\} \cup \mathbb{R}^Z$, we have $|C(\beta^*_2)| < \infty$.

Lemma 1. Let $(\beta^*, \mu^*) \in \mathit{PBE}(\Lambda)$. For any $f \in C(\beta^*_2)$, we have

(a) $\Pi^*_2(\beta^* \,|\, f) \geq \Pi^{\text{SPE}}_2$,

(b) $\Pi^*_2(\beta^* \,|\, f) \geq \Pi^*_2(\beta^* \,|\, g)$ for all $g \in \mathbb{R}^Z$ with equality for $g \in C(\beta^*_2)$,

[27] What makes (β, μ) an equilibrium is the fact that in this assessment player 1 believes that player 2 has aligned the incentives of A with himself perfectly at all out-of-equilibrium information sets.

(c) $\Pi_2^*(\beta^* \mid f) = \Pi_2^{\mathrm{SPE}}$ if $\beta_2^*[\emptyset](\neg D) \in (0,1)$.

Proof. To prove part (a), take any $f \in C(\beta_2^*)$ and notice that

$$\Pi_2^*(\beta^* \mid \emptyset) = \beta_2^*[\emptyset](\neg D)\Pi_2^{\mathrm{SPE}} + \sum_{g \in C(\beta_2^*)} \beta_2^*[\emptyset](g)\Pi_2^*(\beta^* \mid g)$$

since $\Pi_2^*(\beta^* \mid \neg D) = \Pi_2^{\mathrm{SPE}}$ by sequential rationality. But then, if $\Pi_2^*(\beta^* \mid f) < \Pi_2^{\mathrm{SPE}}$ was the case, we would have $\Pi_2^*(\beta_2, \beta_{-2}^* \mid \emptyset) > \Pi_2^*(\beta^* \mid \emptyset)$ where β_2 is any strategy in $S_2(\Lambda)$ with

$$\beta_2[\emptyset](s) = \begin{cases} \beta_2^*[\emptyset](\neg D) + \beta_2^*[\emptyset](f), & s = \neg D \\ 0, & s = f \\ \beta_2^*[\emptyset](s), & \text{otherwise.} \end{cases}$$

Since this would contradict the sequential rationality of (β^*, μ^*), we may conclude that $\Pi_2^*(\beta^* \mid f) \geq \Pi_2^{\mathrm{SPE}}$. Parts (b) and (c) are proved similarly. □

Lemma 2. Let $(\beta^*, \mu^*) \in PBE(\Lambda)$. For any $f \in C(\beta_2^*)$, we have $\sigma_{f,2}^* \in BR_2(\sigma_1^*)$.[28]

Proof. Fix an arbitrary $f \in C(\beta_2^*)$ and take any pure strategy $s_2 \in BR_2(\sigma_1^*)$. Consider now the function $g_\varepsilon : Z \longrightarrow \mathbb{R}$ defined as

$$g_\varepsilon(z) \equiv \begin{cases} \delta + \varepsilon, & \text{if } z \in \mathit{supp}(o(\sigma_1^*, s_2)) \\ \varepsilon, & \text{otherwise,} \end{cases} \tag{4}$$

where $\varepsilon > 0$. Notice that, given μ^* and β_1^*, the behavioral strategy $b_{g_\varepsilon,2}^*$ (and hence $\sigma_{g_\varepsilon,2}^*$) must be chosen to guarantee that $z \in \mathit{supp}(o[\sigma_1^*, s_2])$ holds for all $z \in \mathit{supp}(o[\sigma_1^*, \sigma_{g_\varepsilon,2}^*])$. So, we must have

$$\mathit{supp}(o[\sigma_1^*, \sigma_{g_\varepsilon,2}^*]) \subseteq \mathit{supp}(o[\sigma_1^*, s_2]) \tag{5}$$

while $\beta_A^*[g_\varepsilon](a) = 1$ and $G(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) \equiv \sum_{z \in Z} o[\sigma_1^*, \sigma_{g_\varepsilon,2}^*](z) g_\varepsilon(z) = \delta + \varepsilon$. Moreover, we have

Claim 1. $\sigma_{g_\varepsilon,2}^* \in BR_2(\sigma_1^*)$.

Proof of Claim 1. Take any $s_2' \in \mathit{supp}(\sigma_{g_\varepsilon,2}^*)$. By (5), we must have

$$\mathit{supp}(o[\sigma_1^*, s_2']) \subseteq \mathit{supp}(o[\sigma_1^*, s_2]). \tag{6}$$

Suppose that the converse containment does not hold. This means that there exists a history $h \in H(2)$ that is reached with positive probability (when Γ is played according to either (σ_1^*, s_2) or (σ_1^*, s_2')) such that $s_2(h) \neq s_2'(h)$. But then, for any $z \in \mathit{supp}(o[\sigma_1^*, s_2 \mid h])$ and any $z' \in \mathit{supp}(o[\sigma_1^*, s_2' \mid h])$, we have $z \neq z'$ since the $(|h|+1)$th element of these terminal histories are distinct. This entails that $\mathit{supp}(o[\sigma_1^*, s_2' \mid h]) \cap \mathit{supp}(o[\sigma_1^*, s_2 \mid h]) = \emptyset$ which contradicts

[28] Throughout this section BR_i stands for the *best response* correspondence of player i in the principals-only game Γ.

(6) given that h is reached with positive probability. Consequently, we may conclude that $supp(o[\sigma_1^*, s_2']) = supp(o[\sigma_1^*, s_2])$. Since both s_2 and s_2' are pure strategies, this implies that

$$o[\sigma_1^*, s_2'](z) = o[\sigma_1^*, s_2](z) \quad \text{for all } z \in Z.$$

Therefore, since $s_2 \in BR_2(\sigma_1^*)$, we must also have $s_2' \in BR_2(\sigma_1^*)$. Given that s_2' was arbitrary in this discussion, we have established that $s_2' \in BR_2(\sigma_1^*)$ for all $s_2' \in supp(\sigma_{g_\varepsilon,2}^*)$. This proves the claim. ‖

We are now ready to prove the following

Claim 2. $\beta_A^*[f](a) = 1$.

Proof of Claim 2. Notice first that if $\beta_A^*[f](a) = 0$ was the case, we would have $\Pi_2^*(\beta^* \,|\, f) = \Pi_2^{\text{SPE}} - c < \Pi_2^{\text{SPE}}$, which contradicts Lemma 1(a). Assume next then that $0 < p \equiv \beta_A^*[f](a) < 1$. In this case the agent A is indifferent between accepting and rejecting the contract f, and hence it must be the case that $F(\sigma_1^*, \sigma_{f,2}^*) = \delta$. But then by Lemma 1(a)

$$\Pi_2^{\text{SPE}} \leq \Pi_2^*(\beta^* \,|\, f) = p(\Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \delta - c) + (1-p)\left(\Pi_2^{\text{SPE}} - c\right)$$

so that $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) \geq \Pi_2^{\text{SPE}} + \delta + c/p > \Pi_2^{\text{SPE}} + \delta$. Now choose any

$$0 < \varepsilon < (1-p)\left(\Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \Pi_2^{\text{SPE}} - \delta\right),$$

and consider the contract g_ε defined in (4). Since $\beta_A^*[g_\varepsilon](a) = 1$ and $G(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) = \delta + \varepsilon$, by using Claim 1 and the choice of ε we find

$$\begin{aligned}
\Pi_2^*(\beta^* \,|\, g_\varepsilon) &= \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) - \delta - \varepsilon - c \\
&\geq \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \delta - \varepsilon - c \\
&> \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \delta - (1-p)\left(\Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \Pi_2^{\text{SPE}} - \delta\right) - c \\
&= p(\Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \delta - c) + (1-p)\left(\Pi_2^{\text{SPE}} - c\right) \\
&= \Pi_2^*(\beta^* \,|\, f),
\end{aligned}$$

which contradicts Lemma 1(b). ‖

Since Claim 2 shows that the agent accepts the contract f with positive probability (in fact with probability one), her expected payoff conditional on accepting this contract must be at least as large as her outside option. We state this observation explicitly for future reference.

Claim 3. $F(\sigma_1^*, \sigma_{f,2}^*) \geq \delta$.

Our objective now is to establish that

$$\Pi_2(\sigma_1^*, \sigma_{f,2}^*) \geq \Pi_2(\sigma_1^*, \sigma_2) \qquad \text{for all } \sigma_2 \in M_2(\Gamma).$$

To derive a contradiction, we assume that $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) < \Pi_2(\sigma_1^*, \sigma_2)$ for some $\sigma_2 \in M_2(\Gamma)$. Then, by Claim 1, we have $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) < \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*)$ for any $\varepsilon > 0$. But notice that $\Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*)$ is in fact independent of $\varepsilon > 0$ since, for any $\varepsilon, \varepsilon' > 0$, g_ε and $g_{\varepsilon'}$ are positive affine transformations of one another. Thus, we may unambiguously write $K \equiv \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*)$ and conclude that $K > \Pi_2(\sigma_1^*, \sigma_{f,2}^*)$. We then choose $0 < \varepsilon < K - \Pi_2(\sigma_1^*, \sigma_{f,2}^*)$ and observe that

$$\begin{aligned}
\Pi_2^*(\beta^* \,|\, f) &= \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - F(\sigma_1^*, \sigma_{f,2}^*) - c && \text{(by Claim 2)} \\
&< K - \varepsilon - \delta - c && \text{(by choice of } \varepsilon \text{ and Claim 3)} \\
&= \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) - \varepsilon - \delta - c && \\
&= \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) - G(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) - c && \text{(by definition of } g_\varepsilon) \\
&= \Pi_2^*(\beta^* \,|\, g_\varepsilon),
\end{aligned}$$

which contradicts Lemma 1(b). The proof of Lemma 2 is then complete. □

Lemma 3. Let $(\beta^*, \mu^*) \in PBE(\Lambda)$. For any $f \in C(\beta_2^*)$, we have

$$\beta_A^*[f](a) = 1 \quad \text{and} \quad F(\sigma_1^*, \sigma_{f,2}^*) = \delta.$$

Proof. In view of Claims 2 and 3 above, all we need to show is that $F(\sigma_1^*, \sigma_{f,2}^*) > \delta$ cannot hold for any $f \in C(\beta_2^*)$. Suppose it does, and consider the contract $g_\varepsilon : Z \to \mathbb{R}$ defined by (4) for any $s_2 \in supp(\sigma_{f,2}^*)$ and any $0 < \varepsilon < F(\sigma_1^*, \sigma_{f,2}^*) - \delta$. Clearly, $\beta_A^*[g_\varepsilon] = a$ and $G(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) = \delta + \varepsilon$ and, by Claim 1 above, we have $\sigma_{g_\varepsilon,2}^* \in BR_2(\sigma_1^*)$. But, by Lemma 2, we also have $\sigma_{f,2}^* \in BR_2(\sigma_1^*)$. Therefore, $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) = \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*)$ so that

$$\begin{aligned}
\Pi_2^*(\beta^* \,|\, f) &= \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - F(\sigma_1^*, \sigma_{f,2}^*) - c \\
&< \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \delta - \varepsilon - c \\
&= \Pi_2(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) - G(\sigma_1^*, \sigma_{g_\varepsilon,2}^*) - c \\
&= \Pi_2^*(\beta^* \,|\, g_\varepsilon),
\end{aligned}$$

which contradicts Lemma 1(b). □

Lemma 4. Let $(\beta^*, \mu^*) \in PBE(\Lambda)$. If $(f', a, z) \in supp(o(\beta^*))$, then there exists a Nash equilibrium $\hat{\sigma} \in NE(\Gamma)$ such that $z \in supp(o(\hat{\sigma}))$.

Proof. If $C(\beta_2^*) = \{\neg D\}$ then the claim is trivially established, so throughout we assume that $C(\beta_2^*) \neq \{\neg D\}$. Let $H_o(1)$ stand for the set of all shortest histories in $H(1)$; that is,

$H_{\mathrm{o}}(1) \equiv \{h \in H(1) : \text{there does not exist any } h' \in H(1) \text{ and } h'' \neq \emptyset \text{ such that } h = (h', h'')\}$. We let $I^h \equiv \{(f, a, h) : f \in \mathbb{R}^Z\}$ for any $h \in H(1)$, and define

$$T^h \equiv \sum_{(f,a,h)\in I^h \cap supp(p[\beta^*|\emptyset])} p[\beta^*|\emptyset](f, a, h) \quad \text{for all } h \in H_{\mathrm{o}}(1),$$

where $p[\cdot|\cdot]$ is defined for Λ as in Section 3.1. It is important to note that the definition of T^h is independent of β_1^* since $p[\beta^*|\emptyset](f, a, h)$ is independent of β_1^* for any shortest history h in $H(1)$. Therefore, by consistency and sequential rationality, we have

$$\beta_1^* \in \arg\max_{\beta_1 \in S_1(\Lambda)} \sum_{f\in C(\beta_2^*)} \sum_{z\in supp(O[\beta_1, \beta_{-1}^* \,|\, f,a,h])} \left(\mu^*[I^h](f, a, h)\right) O[\beta_1, \beta_{-1}^* \,|\, f, a, h](z)\pi_1(z)$$

for all $h \in H_{\mathrm{o}}(1)$ with $p[\beta^*|\emptyset](f, a, h) > 0$ for some $f \in \mathbb{R}^Z$. Recalling that $b_1^* \in S_1(\Gamma)$ is the behavioral strategy in Γ induced by β_1^* (that is, $b_1^*[h] \equiv \beta_1^*[I^h]$ for all $h \in H(1)$), we thus have

$$\begin{aligned} b_1^* \;\; &\in \;\; \arg\max_{b_1 \in S_1(\Gamma)} \sum_{f\in C(\beta_2^*)} \sum_{z\in Z} \left(\mu^*[I^h](f, a, h)\right) o[b_1, b_{f,2}^* \,|\, h](z)\pi_1(z) \\ &= \;\; \arg\max_{b_1 \in S_1(\Gamma)} \sum_{f\in C(\beta_2^*)} \sum_{z\in Z} \left(\frac{p[\beta^*|\emptyset](f, a, h)}{T^h}\right) o[b_1, b_{f,2}^* \,|\, h](z)\pi_1(z) \\ &= \;\; \arg\max_{b_1 \in S_1(\Gamma)} \sum_{f\in C(\beta_2^*)} \sum_{z\in Z} \beta_2^*[\emptyset](f) p[\beta^*|f](f, a, h) o[b_1, b_{f,2}^* \,|\, h](z)\pi_1(z) \qquad (7) \end{aligned}$$

for all $h \in H_{\mathrm{o}}(1)$ with $T^h > 0$. (This is because for all such h consistency assures that $\mu^*[I^h](f, a, h) = p[\beta^*|\emptyset](f, a, h)/T^h$ and $\beta_A^*[f](a) = 1$ for all $f \in C(\beta_2^*)$ by Lemma 3.) But notice that (7) holds trivially for any $h \in H_{\mathrm{o}}(1)$ such that $p[\beta^*|\emptyset](f, a, h) = 0$ for all $f \in \mathbb{R}^Z$, since in this case the maximand of the associated optimization problem is identically zero. Consequently, (7) holds for all $h \in H_{\mathrm{o}}(1)$, and we thus have

$$b_1^* \in \arg\max_{b_1 \in S_1(\Gamma)} \sum_{f\in C(\beta_2^*)} \sum_{z\in Z} \beta_2^*[\emptyset](f) \left(\sum_{h\in H_{\mathrm{o}}(1)} p[\beta^*|f](f, a, h) o[b_1, b_{f,2}^* \,|\, h](z) \right) \pi_1(z) \qquad (8)$$

Now let $\mathcal{H}(h)$ stand for the set of all histories in H that is consistent with $h \in H$. Our next task is to prove the following

Claim. For all $z \in Z \cap \bigcup_{h\in H_{\mathrm{o}}(1)} \mathcal{H}(h)$,

$$\sum_{h\in H_{\mathrm{o}}(1)} p[\beta^*|f](f, a, h) o[b_1, b_{f,2}^* \,|\, h](z) = o[b_1, b_{f,2}^*](z).$$

Proof of Claim. Take any $z \in Z \cap \bigcup_{h\in H_{\mathrm{o}}(1)} \mathcal{H}(h)$. By definition of $H_{\mathrm{o}}(1)$, there can be at most one h in $H_{\mathrm{o}}(1)$ that is consistent with z. By the choice of z, therefore, there exists a unique $h \in H_{\rm o}(1)$ that is consistent with z, denote this history by h_z. But then, $o[b_1, b^*_{f,2} \,|\, h](z) = 0$ for all $h \neq h_z$, and hence we have

$$\sum_{h \in H_{\rm o}(1)} p[\beta^*|f](f, a, h) o[b_1, b^*_{f,2} \,|\, h](z) = p[\beta^*|f](f, a, h_z) o[b_1, b^*_{f,2} \,|\, h_z](z) = o[b_1, b^*_{f,2}](z)$$

by definition of $p[\beta^*|f]$. $\|$

Using this claim and (8), we find

$$b^*_1 \in \arg\max_{b_1 \in S_1(\Gamma)} \sum_{f \in C(\beta^*_2)} \sum_{z \in Z} \beta^*_2[\emptyset](f) o[b_1, b^*_{f,2}](z) \pi_1(z)$$

since it is readily observed that, for any f, the probability $o[b_1, b^*_{f,2}](z)$ is independent of b_1 for any terminal history $z \notin \bigcup_{h \in H_{\rm o}(1)} \mathcal{H}(h)$. Now define the function $\lambda : C(\beta^*_2) \to [0, 1]$ such that

$$\lambda(f) = \frac{\beta_2[\emptyset](f)}{1 - \beta_2[\emptyset](\neg D)}$$

and notice that $\sum_{f \in C(\beta^*_2)} \lambda(f) = 1$ and

$$b^*_1 \in \arg\max_{b_1 \in S_1(\Gamma)} \sum_{f \in C(\beta^*_2)} \sum_{z \in Z} \lambda(f) o[b_1, b^*_{f,2}](z) \pi_1(z) \tag{9}$$

Moreover, for any $b_1 \in S_1(\Gamma)$, we have

$$\begin{aligned} \sum_{f \in C(\beta^*_2)} \sum_{z \in Z} \lambda(f) o[b_1, b^*_{f,2}](z) \pi_1(z) &= \sum_{z \in Z} o\left[b_1, \textstyle\sum_{f \in C(\beta^*_2)} \lambda(f) b^*_{f,2}\right](z) \pi_1(z) \\ &= \Pi_1\left(b_1, \textstyle\sum_{f \in C(\beta^*_2)} \lambda(f) b^*_{f,2}\right) \end{aligned}$$

so that, by (9), we have $b^*_1 \in \arg\max_{b_1 \in S_1(\Gamma)} \Pi_1(b_1, \hat{b}_2)$, where $\hat{b}_2 \in S_2(\Gamma)$ is defined as $\hat{b}_2 \equiv \sum_{f \in C(\beta^*_2)} \lambda(f) b^*_{f,2}$. So, letting

$$\hat{\sigma}_2 \equiv \sum_{f \in C(\beta^*_2)} \lambda(f) \sigma^*_{f,2} \in M_2(\Gamma)$$

which is obviously outcome-equivalent to $\hat{b}_2$, we find $\sigma^*_1 \in BR_1(\hat{\sigma}_2)$. But since, by Lemma 2, $\sigma^*_{f,2} \in BR_2(\sigma^*_1)$ for all $f \in C(\beta^*_2)$, we also have $\hat{\sigma}_2 \in BR_2(\sigma^*_1)$. Therefore, $(\sigma^*_1, \hat{\sigma}_2) \in NE(\Gamma)$. Moreover, if $(f', a, z) \in supp(o(\beta^*))$, then it must be the case that $f' \in C(\beta^*_2)$ and $z \in supp(o(\sigma^*_1, \sigma^*_{f',2}))$. Therefore, we have $z \in supp(o(\sigma^*_1, \hat{\sigma}_2))$ and the proof is complete. □

Proof of Proposition 2. Apply Lemma 4. □

Define

$$\mathbf{\Pi}^{\rm PBE}_2(\Lambda) \equiv \{\tilde{\Pi}_2(\beta) : (\beta, \mu) \in PBE(\Lambda) \text{ for some system of beliefs } \mu\},$$

where

$$\tilde{\Pi}_2(\beta) \equiv \beta_2[\emptyset](\neg D)\Pi_2^{\mathrm{SPE}} + \sum_{f \in C(\beta_2)} \beta_2[\emptyset](f)\Pi_2(\sigma_1, \sigma_{f,2}) \tag{10}$$

for any behavioral strategy profile $\beta \in S(\Lambda)$. Since, by Lemma 3, no contract $f \in C(\beta_2^*)$ is rejected with a positive probability in equilibrium, $\tilde{\Pi}_2(\beta)$ is the expected payoff of player 2 (gross of the compensation he pays to the delegate in case of a hire and the contracting cost) in the equilibrium (β, μ). The following important lemma points to the close connection between the sets $\boldsymbol{\Pi}_2^{\mathrm{NE}}$, $\boldsymbol{\Pi}_2^{\mathrm{NE}_1^*}(\Gamma)$ and $\boldsymbol{\Pi}_2^{\mathrm{PBE}}(\Lambda)$, and constitutes a crucial step towards proving Theorem 1.

Lemma 5. **(*a*)** $\boldsymbol{\Pi}_2^{\mathrm{PBE}}(\Lambda(\Gamma, \delta, c)) \subseteq \{\Pi_2 \in co\,\boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma) : \Pi_2 \geq \Pi_2^{\mathrm{SPE}}\}$.

(*b*) There exists an $\ell > 0$ such that

$$\boldsymbol{\Pi}_2^{\mathrm{PBE}}(\Lambda(\Gamma, \delta, c)) \supseteq \{\Pi_2 \in \boldsymbol{\Pi}_2^{\mathrm{NE}_1^*}(\Gamma) : \Pi_2 \geq \Pi_2^{\mathrm{SPE}}\}$$

for all $\delta + c < \ell$.

Proof. **(*a*)** Let $\Pi_2 \in \boldsymbol{\Pi}_2^{\mathrm{PBE}}(\Lambda)$ and take any $(\beta^*, \mu^*) \in PBE(\Lambda)$ such that $\tilde{\Pi}_2(\beta^*) = \Pi_2$. By Lemma 1(a), we have $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) > \Pi_2^*(\beta^* \mid f) \geq \Pi_2^{\mathrm{SPE}}$ so that $\tilde{\Pi}_2(\beta^*) \geq \Pi_2^{\mathrm{SPE}}$, and it remains to show that $\Pi_2 \in co\,\boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma)$. Notice first that if $\beta_2^*[\emptyset](\neg D) = 1$, then $\Pi_2 = \tilde{\Pi}_2(\beta^*) = \Pi_2^{\mathrm{SPE}} \in \boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma)$, so we are done. If, on the other hand, $\beta_2^*[\emptyset](\neg D) = 0$, we define $\hat{\sigma}$ as in the proof of Lemma 4 (so that $\hat{\sigma}_2 \equiv \sum_{f \in C(\beta_2^*)} \beta_2^*[\emptyset](f)\sigma_{f,2}^*$), and recall that $(\sigma_1^*, \hat{\sigma}_2) \in NE(\Gamma)$. But, since we have $\lambda(f) = \beta_2^*[\emptyset](f)$ for all $f \in C(\beta_2^*)$ in this case, we have

$$\tilde{\Pi}_2(\beta^*) = \sum_{f \in C(\beta_2^*)} \lambda(f)\Pi_2(\sigma_1^*, \sigma_{f,2}^*) = \Pi_2\left(\sigma_1^*, \sum_{f \in C(\beta_2^*)} \lambda(f)\sigma_{f,2}^*\right) = \Pi_2(\sigma_1^*, \hat{\sigma}_2) \in \boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma)$$

and the claim follows.

To complete the proof we need to consider the case in which $\beta_2^*[\emptyset](\neg D) \in (0, 1)$. But this case is easily settled by noticing that

$$\begin{aligned} \tilde{\Pi}_2(\beta^*) &= \beta_2[\emptyset](\neg D)\Pi_2^{\mathrm{SPE}} + (1 - \beta_2[\emptyset](\neg D))\Pi_2\left(\sigma_1^*, \textstyle\sum_{f \in C(\beta_2^*)} \lambda(f)\sigma_{f,2}^*\right) \\ &= \beta_2[\emptyset](\neg D)\Pi_2^{\mathrm{SPE}} + (1 - \beta_2[\emptyset](\neg D))\Pi_2(\sigma_1^*, \hat{\sigma}_2) \\ &\in co\,\boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma), \end{aligned}$$

where we again used the bilinearity of Π_2.

(*b*) Let $\Pi_2 \in \boldsymbol{\Pi}_2^{\mathrm{NE}_1^*}(\Gamma)$ and $\Pi_2 \geq \Pi_2^{\mathrm{SPE}}$. We need to show that there exists a $(\beta, \mu) \in PBE(\Lambda)$ such that $\tilde{\Pi}_2(\beta) = \Pi_2$. We construct such an equilibrium by distinguishing between two cases.

Case 1. $\Pi_2 = \Pi_2^{\mathrm{SPE}}$.

In this case we define the assessment (β, μ) exactly as in the proof of Proposition 1. For this assessment we have $(\beta, \mu) \in \mathit{PBE}(\Lambda)$ and $\tilde{\Pi}_2(\beta) = \Pi_2^{\mathrm{SPE}}$.

Case 2. $\Pi_2 > \Pi_2^{\mathrm{SPE}}$.

Since $\Pi_2 \in \boldsymbol{\Pi}_2^{\mathrm{NE}_1^*}(\Gamma)$, there exists a $\bar{\sigma} \in \mathit{NE}(\Gamma)$ such that $\Pi_2(\bar{\sigma}) = \Pi_2$, and $\bar{\sigma}_1$ is sequentially rational. Let $\bar{s}_2 \in \mathit{supp}(\bar{\sigma}_2)$ and let $\bar{b}_1$ be an outcome-equivalent strategy to $\bar{\sigma}_1$. We define the contract

$$g(z) \equiv \delta \mathbf{1}_{\mathit{supp}(o(\bar{\sigma}_1, \bar{s}_2))}.$$

Define next $\beta \in S(\Lambda)$ as follows: For all $f \in \mathbb{R}^Z$,

$$\beta_1[f, r, h] = \beta_1[\neg D, h] = b_1^{\mathrm{SPE}}[h] \text{ and } \beta_1[f, a, h] = \bar{b}_1[h] \text{ for all } h \in H(1),$$

$$\beta_2[\emptyset](g) = 1,\ \beta_2[\neg D, h] = \beta_2[f, r, h] = b_2^{\mathrm{SPE}}[h] \text{ for all } h \in H(2),$$

$$\beta_A[f](a) = \begin{cases} 1, & \text{if } F(\bar{b}_1, b^f) \geq \delta \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \beta_A[f, a, h] = \begin{cases} \bar{s}_2[h], & \text{if } f = g \\ b^f[h], & \text{if } f \neq g \end{cases}$$

for all $h \in H(2)$, with $b^f \in S_2(\Gamma)$ being such that $b^f[h] = b^{f,h}[h]$ for all $h \in H(2)$, where

$$b^{f,h} \in \operatorname*{arg\,max}_{b_2 \in S_2(\Gamma)} \sum_{z \in Z} o[\bar{b}_1, b_2 \,|\, h](z) f(z).$$

On the other hand, we define $\mu \in B(\Lambda)$ by specifying that $\mu[I](g, \theta, h) = 1$ for all $I = \{(f, \theta, h) : f \in \mathbb{R}^Z\} \in \mathcal{I}_1^*$. It can be checked that, for small enough $\delta + c$, we have $(\beta, \mu) \in \mathit{PBE}(\Lambda)$ while $\tilde{\Pi}_2(\beta) = \Pi_2(\bar{\sigma}_1, \bar{s}_2) = \Pi_2$. □

Lemma 6. There exists an $\ell > 0$ such that

$$\beta_2[\emptyset](\neg D) \in \{0, 1\}$$

for any $(\beta, \mu) \in \mathit{PBE}(\Lambda)$ and any $\delta + c < \ell$.

Proof. Fix any $(\beta^*, \mu^*) \in \mathit{PBE}(\Lambda)$ and notice that Lemma 1(b) and (3) imply that $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) = \Pi_2(\sigma_1^*, \sigma_{g,2}^*)$ for all $f, g \in C(\beta_2^*)$. Therefore, by (10),

$$\tilde{\Pi}_2(\beta^*) = \beta_2^*[\emptyset](\neg D)\Pi_2^{\mathrm{SPE}} + (1 - \beta_2^*[\emptyset](\neg D))\Pi_2(\sigma_1^*, \sigma_{f,2}^*) \tag{11}$$

for any $f \in C(\beta_2^*)$. Now if $\Pi_2^{\mathrm{SPE}} = \max \boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma)$, then by Lemma 5(a), $\tilde{\Pi}_2(\beta^*) = \Pi_2^{\mathrm{SPE}}$ and (11) yields $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) = \Pi_2^{\mathrm{SPE}}$ for any $f \in C(\beta_2^*)$. But then, by (3) and Lemma 3, $\Pi_2^*(\beta^* \,|\, f) = \Pi_2(\sigma_1^*, \sigma_{f,2}^*) - \delta - c < \Pi_2^{\mathrm{SPE}}$ for any $f \in C(\beta_2^*)$. In view of Lemma 1(a), this can hold only if $C(\beta_2^*) = \{\neg D\}$, which implies that $\beta_2^*(\emptyset)(\neg D) = 1$.

Assume next that $\Pi_2 > \Pi_2^{\mathrm{SPE}}$ for some $\Pi_2 \in \boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma)$, and define

$$\ell \equiv \min \left\{ \Pi_2 - \Pi_2^{\mathrm{SPE}} : \Pi_2^{\mathrm{SPE}} < \Pi_2 \in \boldsymbol{\Pi}_2^{\mathrm{NE}}(\Gamma) \right\}.$$

(That ℓ is well-defined follows from the hypothesis that $\left|\mathbf{\Pi}_2^{\mathrm{NE}}\right| < \infty$.) To derive a contradiction, assume next that $\ell > \delta + c$ and $\beta_2^*[\emptyset](\neg D) \in (0,1)$. Then, by Lemma 1(c), $\Pi_2^*(\beta^* \,|\, f) = \Pi_2^{\mathrm{SPE}}$ so that, by (3) and Lemma 3, $\Pi_2(\sigma_1^*, \sigma_{f,2}^*) = \Pi_2^{\mathrm{SPE}} + \delta + c$ for all $f \in C(\beta_2^*)$. Defining λ and $\hat{\sigma}_2$ as in the proof of Lemma 4, we thus find

$$\Pi_2(\sigma_1^*, \hat{\sigma}_2) = \Pi_2\left(\sigma_1^*, \sum_{f \in C(\beta_2^*)} \lambda(f)\sigma_{f,2}^*\right) = \sum_{f \in C(\beta_2^*)} \lambda(f)\Pi_2(\sigma_1^*, \sigma_{f,2}^*) = \Pi_2^{\mathrm{SPE}} + \delta + c$$

whereas $(\sigma_1^*, \hat{\sigma}_2) \in \mathbf{\Pi}_2^{\mathrm{NE}}(\Gamma)$. But then $\Pi_2^{\mathrm{SPE}} + \delta + c \in \{\Pi_2 \in \mathbf{\Pi}_2^{\mathrm{NE}}(\Gamma) : \Pi_2 > \Pi_2^{\mathrm{SPE}}\}$, and thus by definition of ℓ, we obtain $\ell \leq \Pi_2^{\mathrm{SPE}} + \delta + c - \Pi_2^{\mathrm{SPE}} = \delta + c$. This completes the proof. □

In what follows we adopt the notation introduced in proving Lemma 4. Thus H_{o} stands for the set of all shortest histories in $H(1)$, and we let $I^h \equiv \{(f,a,h) : f \in \mathbb{R}^Z\} \in \mathcal{I}_1$ for any $h \in H_{\mathrm{o}}(1)$.

Lemma 7. There exists an $\ell > 0$ such that

$$\beta_2^*[\emptyset](\neg D) = \begin{cases} 1, & \text{if } \bigcup_{h \in H_{\mathrm{o}}(1)} PBE(\Lambda; I^h) = \emptyset \\ 0, & \text{otherwise} \end{cases}$$

for any well-supported equilibrium $(\beta^*, \mu^*) \in PBE(\Lambda(\Gamma, \delta, c))$, provided that $\delta + c < \ell$.

Proof. Pick an arbitrary $(\beta^*, \mu^*) \in PBE_{\text{w-s}}(\Lambda)$, choose ℓ as in Lemma 6, and assume that $\delta + c < \ell$. By Lemma 6, either $\beta_2^*[\emptyset](\neg D) = 0$ or $\beta_2^*[\emptyset](\neg D) = 1$. If $PBE(\Lambda; I^h) = \emptyset$ for all $h \in H_{\mathrm{o}}(1)$, then $\beta_2^*[\emptyset](\neg D) = 0$ cannot hold since otherwise we have $(\beta^*, \mu^*) \in \bigcup_{h \in H_{\mathrm{o}}(1)} PBE(\Lambda; I^h)$. Assume next that there exists an $h \in H_{\mathrm{o}}(1)$ such that $PBE(\Lambda; I^h) \neq \emptyset$. Let us denote the information set I^h by I, and, to derive a contradiction, suppose that $\beta_2^*[\emptyset](\neg D) = 1$. Consequently, $I \in \mathcal{J}(\beta^*)$, and by well-supportedness of (β^*, μ^*), there exists an assessment $(\beta, \mu) \in PBE(\Lambda; I)$ such that $\mu^*[I'] = \mu[I']$ and

$$\beta_1^*[I'] = \beta_1[I'] \quad \text{for all } I' \in \mathcal{I}_1(I), \tag{12}$$

where $\mathcal{I}_1(I)$ is the set of all information sets of player 1 that follows I.

Let us now pick any $f \in C(\beta_2)$ such that $b_{f,2}$ visits I with positive probability (where, of course, $b_{f,2}[h] \equiv \beta_A[f,a,h]$ for all $h \in H(2)$). We have the following

Claim 1. $\Pi_2^*(\beta \,|\, f) > \Pi_2^{\mathrm{SPE}}$.

Proof of Claim 1. If $\Pi_2^*(\beta \,|\, f) = \Pi_2^{\mathrm{SPE}}$ held, then we would have $((\beta_1, \beta_2', \beta_A), \mu) \in PBE(\Lambda)$ where $\beta_2' \in S_2(\Lambda)$ is defined as $\beta_2'[\emptyset](\neg D) \equiv 1/2$, $\beta_2'[\emptyset](g) \equiv 2^{-1}\beta_2[\emptyset](g)$ for all $g \in \mathbb{R}^Z$ and $\beta_2'[h] \equiv \beta_2[h]$ for all $h \in H^*(2)\backslash\{\emptyset\}$. This, however, contradicts Lemma 6. We must then have $\Pi_2^*(\beta \,|\, f) \neq \Pi_2^{\mathrm{SPE}}$, and the claim follows from Lemma 1(a). ‖

Let $\sigma_{f,2}$ be an outcome-equivalent mixed strategy to $b_{f,2}$ and let $\sigma_1 \in M_1(\Gamma)$ be an outcome equivalent mixed strategy to $b_1 \in S_1(\Gamma)$ (which is induced by $\beta_1[(f,a,h) : f \in \mathbb{R}^Z]$, $h \in H(1)$, in the usual way). By the choice of f, $\sigma_{f,2}$ visits I with positive probability. (Recall that player 1 does not play a role in I being reached since $I = I^h$ with h being a shortest history in $H(1)$). We choose any pure strategy $s_2 \in supp(\sigma_{f,2})$ such that s_2 reaches I.[29] Define next the contract $g \in \mathbb{R}^Z$ by

$$g(z) \equiv \begin{cases} \pi_2(z) - \Pi_2(\sigma_1, s_2) + \delta + \varepsilon, & \text{if } z \in \ supp(o[\sigma_1, s_2]) \\ 0, & \text{otherwise} \end{cases}$$

for any $\varepsilon > 0$.

Claim 2. $G(\sigma_1, s_2) = G(\sigma_1^*, \sigma_{g,2}^*)$.

Proof of Claim 2. By definition of g and sequential rationality, any $s_2' \in supp(\sigma_{g,2}^*)$ must reach I regardless of the strategy choice of player 1. Therefore, all information sets that are reached when $\sigma_{g,2}^*$ is played follow I regardless of the play of 1. Then, by (12), we must have

$$o[\sigma_1, \sigma_{g,2}^*] = o[\sigma_1^*, \sigma_{g,2}^*]. \tag{13}$$

Since s_2 reaches I as well, by the same reasoning we also have $o[\sigma_1, s_2] = o[\sigma_1^*, s_2]$. Thus by sequential rationality, we find $G(\sigma_1^*, \sigma_{g,2}^*) \geq G(\sigma_1^*, s_2) = G(\sigma_1, s_2)$. To establish the converse inequality, notice that by the choice of g we have

$$BR_2(\sigma_1) = \underset{\sigma_2' \in M_2(\Gamma)}{\arg\max}\, G(\sigma_1, \sigma_2').$$

But by Lemma 2, $\sigma_{f,2} \in BR_2(\sigma_1)$, and hence since $s_2 \in supp(\sigma_{f,2})$, we also have $s_2 \in BR_2(\sigma_1)$. By the previous observation and (13), therefore, $G(\sigma_1, s_2) \geq G(\sigma_1, \sigma_{g,2}^*) = G(\sigma_1^*, \sigma_{g,2}^*)$. This proves the claim. $\|$

By using Claim 2 we observe that

$$G(\sigma_1^*, \sigma_{g,2}^*) = \sum_{z \in Z} o[\sigma_1, s_2](z) g(z) = \sum_{z \in Z} o[\sigma_1, s_2](z) \pi_2(z) - \Pi_2(\sigma_1, s_2) + \delta + \varepsilon = \delta + \varepsilon \tag{14}$$

which, in turn, establishes that $\beta_A^*[g](a) = 1$. Moreover, by Claim 2 and the definition of g, we similarly obtain

$$\Pi_2(\sigma_1, s_2) = \Pi_2(\sigma_1^*, \sigma_{g,2}^*). \tag{15}$$

[29] Formally speaking, by "s_2 reaches I" we mean that I is visited with positive probability via any strategy profile in Λ such that f is played with positive probability and the agent plays according to s_2 upon acceptance of f (recall Lemma 3). The same formalism applies throughout the proof. Also notice that since I is the "first" information set, s_2 reaches I with probability 1.

We now consider the behavioral strategy $\beta_2^{'} \in S_2(\Lambda)$ which is defined as follows:

$$\beta_2'[\emptyset](t) \equiv \begin{cases} 1, & \text{if } t = g \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \qquad \beta_2'[h] = \beta_2^*[h] \text{ for all } h \in H^*(2)\backslash\{\emptyset\}.$$

By (14) and (15),

$$\begin{aligned} \Pi_2^*(\beta_1^*, \beta_2', \beta_A^*) &= \Pi_2(\sigma_1^*, \sigma_{g,2}^*) - G(\sigma_1^*, \sigma_{g,2}^*) - c \\ &= \Pi_2(\sigma_1, s_2) - \delta - \varepsilon - c. \end{aligned} \tag{16}$$

On the other hand, by Lemma 3 and the fact that $s_2 \in BR_2(\sigma_1)$,

$$\Pi_2^*(\beta \,|\, f) = \Pi_2(\sigma_1, \sigma_{f,2}) - F(\sigma_1, s_2) - c = \Pi_2(\sigma_1, s_2) - \delta - c. \tag{17}$$

Thus, by choosing $0 < \varepsilon < \Pi_2^*(\beta \,|\, f) - \Pi_2^{\text{SPE}}$ (recall Claim 1), and by using (16), (17) and the hypothesis that $\beta_2^*[\emptyset](\neg D) = 1$, we get

$$\Pi_2^*(\beta_1^*, \beta_2', \beta_A^*) = \Pi_2^*(\beta \,|\, f) - \varepsilon > \Pi_2^{\text{SPE}} = \Pi_2^*(\beta^*),$$

which contradicts the sequential rationality of (β^*, μ^*). □

Lemma 8. There exists an $\ell > 0$ such that, for all $\delta, c > 0$ such that $\ell > \delta + c$,

$$\bigcup\nolimits_{h \in H_\text{o}(1)} PBE(\Lambda; I^h) \begin{cases} = \emptyset, & \text{if } \Pi_2^{\text{SPE}} = \max \mathbf{\Pi}_2^{\text{NE}}(\Gamma) \\ \neq \emptyset, & \text{if } \Pi_2^{\text{SPE}} < \max \mathbf{\Pi}_2^{\text{NE}_1^*}(\Gamma) \end{cases}.$$

Proof. If $\Pi_2^{\text{SPE}} = \max \mathbf{\Pi}_2^{\text{NE}}(\Gamma)$, then by repeating the argument given in the first paragraph of the proof of Lemma 6, we find $\beta_2[\emptyset](\neg D) = 1$ for all $(\beta, \mu) \in PBE(\Lambda)$ so that $PBE(\Lambda; I^h) = \emptyset$ for all $h \in H_\text{o}(1)$. If $\Pi_2^{\text{SPE}} < \max \mathbf{\Pi}_2^{\text{NE}_1^*}(\Gamma)$, then define the assessment (β, μ) as in Lemma 5(b) Case 2. Clearly, there exists an $h \in H_\text{o}(1)$ such that $(\beta, \mu) \in PBE(\Lambda; I^h)$. □

Proof of Theorem 1. Apply Lemmas 7 and 8, and Proposition 1. □

Proof of Corollary 1. Apply Theorem 1 and Lemma 8. □

Appendix: Well-Supported Equilibria in General Games

In this appendix we shall provide an inductive definition of well-supported equilibria as it applies to the class of all finite extensive form games. Towards this end, take an arbitrary finite extensive form game (with perfect recall) Υ. We define $U(\Upsilon)$ as the set of all histories

h such that there exists an imperfect information subgame (denoted $\Upsilon(h)$) of Υ that follows h.[30] Define H^1 as the set of all h in $U(\Upsilon)$ that is of maximal length. For any integer $t \geq 2$, we let H^t be the set of all h in $U(\Upsilon) \backslash \cup_{r=1}^{t-1} H^r$ that is of maximal length. By finiteness of H, there must exist a smallest integer $\langle \Upsilon \rangle$ such that $H^t = \emptyset$ for all $t \geq \langle \Upsilon \rangle$. Our refinement will be defined by induction on $\langle \Upsilon \rangle$.

Let $h \in U(\Upsilon)$ and define $\mathcal{I}_i(h \,|\, \beta)$ as the set of all information sets of player i which follow h and is reached by strategy profile β with positive probability. Also define $S(\beta_{P(h)}[h])$ as the set of all behavioral strategy profiles in which player $P(h)$ plays $\beta_{P(h)}[h]$ at the information set containing h, and $\mathcal{Q}(\beta_{P(h)}[h])$ as the set of all non-singleton information sets that do not belong to

$$\cup_{i \in N} \cup_{\beta' \in S(\beta_{P(h)}[h])} \mathcal{I}_i(h \,|\, \beta').$$

Define next $\mathcal{J}(h \,|\, \beta)$ as the set of all nonsingleton information sets which could be reached with the shortest sequence of actions after a deviation from $\beta_{P(h)}[h]$, while they are surely not reached when $\beta_{P(h)}[h]$ is played. Thus, $I \in \mathcal{J}(h \,|\, \beta)$ if, and only if, $I \in \mathcal{Q}(\beta_{P(h)}[h])$ and for any $J \in \mathcal{Q}(\beta_{P(h)}[h])$ there do not exist h' and $h'' \neq \emptyset$ such that $h' \in J$ and $(h', h'') \in I$.

We are now ready to define the well-supported equilibria:

Definition. Let Υ be any finite extensive form game with perfect recall. The set of all **well-supported equilibria** of Υ, denoted $PBE_{\text{w-s}}(\Upsilon)$, is defined inductively as follows.

(i) If $\langle \Upsilon \rangle = 1$, then $PBE_{\text{w-s}}(\Upsilon) \equiv PBE(\Upsilon)$.

(ii) Let $\langle \Upsilon \rangle \geq 2$. A perfect Bayesian equilibrium (β, μ) belongs to $PBE_{\text{w-s}}(\Upsilon)$ if, and only if, for all $h \in H^t$, $t \in \{1, ..., \langle \Upsilon \rangle - 1\}$

(a) If, for any $h \in \cup_{r=1}^{t-1} H^r$, there exists a well-supported equilibrium of $\Upsilon(h)$, then

$$(\beta|_{\Upsilon(h)}, \mu|_{\Upsilon(h)}) \in PBE_{\text{w-s}}(\Upsilon(h)), \tag{18}$$

and for all $I \in \mathcal{J}(h \,|\, \beta)$ one of the following holds:

(b) There does not exist a $(\beta', \mu') \in PBE(\Upsilon; I)$ such that

$$(\beta'|_{\Upsilon(h)}, \mu'|_{\Upsilon(h)}) \in PBE_{\text{w-s}}(\Upsilon(h)) \quad \text{for all } h \in \cup_{r=1}^{t-1} H^r, \tag{19}$$

(c) There exists a $(\beta', \mu') \in PBE(\Upsilon; I)$ such that (19) holds and

$$\mathcal{I}_j(h \,|\, \beta') \subseteq \mathcal{I}_j(h \,|\, \beta) \quad \text{for some } j \neq P(h), \tag{20}$$

(d) There exists a $(\beta', \mu') \in PBE(\Upsilon; I)$ such that (19) holds, (20) does not hold and

$$\mu[I'] = \mu'[I'] \quad \text{and} \quad \beta_{P(I)}[I'] = \beta'_{P(I)}[I'] \tag{21}$$

[30] We require $\Upsilon(h)$ to be an imperfect information (sub)game just to prevent dealing with trivial steps in induction.

for all $I' \in \mathcal{I}_{P(I)}(I)$.

That $PBE_{\text{w-s}}(\Upsilon)$ is well-defined for any extensive form game Υ, follows from the fact that if $\langle\Upsilon\rangle \geq 2$, then we have $\langle\Upsilon(h)\rangle \leq \langle\Upsilon\rangle - 1$ for any $h \in H^{\langle\Upsilon\rangle-1}$ with $H^0 = \emptyset$. Moreover, it is easily checked that this definition reduces to the one presented in Section 3. Indeed, for any $\Upsilon \in \mathcal{G}$, we have $\langle\Upsilon\rangle \geq 2$ and hence induction stops after the first step and part (a) of the definition and (19) is automatically satisfied.

We conclude with an example that illustrates the superiority of the above definition over the one considered in the text.

Example A.1. (*Battle of the Sexes with Two Outside Options*; van Damme, 1989) Consider the game Υ depicted in Figure 4, where $H^1 = \{A\}$, $H^2 = \{\emptyset\}$, and $\langle\Upsilon\rangle = 3$. Here there are two types of equilibria. In the first type of equilibria we have $\beta_2[\emptyset](A) = 1$ while $(\beta|_{\Upsilon(A)}, \mu|_{\Upsilon(A)})$ is a first type of equilibrium in $\Upsilon(A)$ as computed in Example 1 of Subsection 3.3. The unique second type of equilibrium is $\beta_2[\emptyset](O_1) = \beta_1[A](T) = \beta_2[I](L) = \mu[I](A,T) = 1$. As we have established in Example 1, the restriction of the equilibria of the first type to the subgame $\Upsilon(A)$ does not belong to $PBE_{\text{w-s}}(\Upsilon(A))$ and hence these equilibria are not well-supported. (Notice that without the requirement (a) in the definition, these equilibria would actually be well-supported.) On the other hand, the restriction of the second type of equilibrium to the subgame $\Upsilon(A)$ is well-supported in that subgame and hence part (a) of the definition is satisfied. When we take $h = \emptyset$, we have $\mathcal{J}(\emptyset \mid \beta) = \{I\}$. Since there exists no $(\beta', \mu') \in PBE(\Upsilon; I)$ part (b) of the definition is satisfied. For $h = A$, we have $\mathcal{J}(A \mid \beta) = \{I\}$ and since there exists no $(\beta', \mu') \in PBE(\Upsilon; I)$, part (b) is again satisfied. Therefore, the unique well-supported equilibrium is of the second type. Precisely the same result would have obtained, had we refined the equilibria of Υ by using the forward induction requirements of Kohlberg (1990) and Al-Najjar (1995). ||

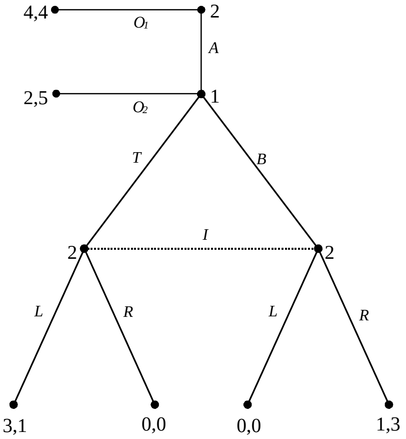

Figure 5: Battle of the Sexes with Two Outside Options

References

Al-Najjar, N. (1995), “A Theory of Forward Induction in Finitely Repeated Games,” *Theory and Decision*, 38, 173-193.

Ben-Porath, E. and E. Dekel (1992), “Signalling Future Actions and the Potential for Sacrifice,” *Journal of Economic Theory*, 57, 36-51.

Bolton, P. and D. S. Scharfstein (1990), “A Theory of Predation Based on Problems in Financial Contracting,” *American Economic Review*, 80, 93-106.

Brander, J. A. and T. R. Lewis (1986), “Oligopoly and Financial Structure: The Limited Liability Effect,” *American Economic Review*, 76, 956-970.

Brander, J. A. and B. J. Spencer (1983), “Strategic Commitment with R&D: The Symmetric Case,” *Bell Journal of Economics*, 14, 225-235.

Brander, J. A. and B. J. Spencer (1985), “Export Subsidies and International Market Share Rivalry,” *Journal of International Economics*, 18, 83-100.

Caillaud, B., B. Jullien, and P. Picard (1995), “Competing Vertical Structures: Precommitment and Renegotiation,” *Econometrica*, 63, 621-646.

Cho, I-K. (1987), “A Refinement of Sequential Equilibrium,” *Econometrica*, 55, 1367-1389.

Cooper, R., D. DeJong, R. Forsythe and T. Ross (1993), “Forward Induction in the Battle-of-the-Sexes Games,” *American Economic Review*, 83, 1303-1316.

Das, S. P (1997), “Strategic Managerial Delegation and Trade Policy,” *Journal of International Economics,* 43, 173-188.

Dewatripont, M. (1988), “Commitment Through Renegotiation-Proof Contracts with Third Parties,” *Review of Economic Studies*, 55, 377-390.

Eaton, J. and G. Grossman (1986), “Optimal Trade and Industrial Policy under Oligopoly,” *Quarterly Journal of Economics*, 101, 383-406.

Fershtman, C. and U. Gneezy (1997), “Stragetic Delegation: An Experiment,” mimeo.

Fershtman, C. and K. Judd (1987), “Equilibrium Incentives in Oligopoly,” *American Economic Review*, 77, 927-940.

Fershtman, C. and E. Kalai (1997), “Unobserved Delegation,” *International Economic Review*, 38, 763-774.

Gatsios, K. and L. Karp (1991), “Delegation Games in Custom Unions,” *Review of Economic Studies*, 58, 391-398.

Hadfield, G. K (1991), “Credible Spatial Preemption Through Franchising,” *Rand Journal of Economics*, 22, 531-543.

Hillas, J. (1994), “Sequential Equilibria and Stable Sets of Beliefs,” *Journal of Economic Theory*, 64, 78-103.

Katz, M. (1991), “Game-Playing Agents: Unobservable Contracts as Precommitments,” *Rand Journal of Economics*, 22, 307-328.

Jensen, H. (1997), “Credibility of Optimal Monetary Delegation,” *American Economic Review*, 87, 911-920.

Koçkesen, L. (1999), “Two-Sided Delegation and Unobservability,” mimeo, NYU.

Kohlberg, E. and J.-F. Mertens (1986), “On the Strategic Stability of Equilibria,” *Econometrica*, 54, 1003-1038.

Kohlberg, E. (1990), “Refinement of Nash Equilibrium: The Main Ideas,” pp. 3-45 in *Game Theory and Applications* (T. Ichiishi, A. Neyman and Y. Tauman, eds.), San Diego: Academic Press.

Kreps, D. (1989), “Out-of-equilibrium Beliefs and Out-of-equilibrium Behavior,” pp. 7-45 in *The Economics of Missing Markets, Information and Games* (F. Hahn, ed.), Oxford: Oxford University Press.

Kreps, D. and R. Wilson (1982), “Sequential Equilibria,” *Econometrica*, 50, 863-894.

McLennan, A. (1985), “Justifiable Beliefs in Sequential Equilibrium,” *Econometrica*, 53, 889-904.

Moulin, H., *Game Theory for the Social Sciences*, New York University Press, 1986.

Osborne, M. and A. Rubinstein, *A Course in Game Theory*, MIT Press, 1994.

Persson, T. and G. Tabellini (1993), “Designing Institutions for Monetary Stability,” *Carnegie-Rochester Conference Series on Public Policy*, 39, 53-84.

Ponssard, J-P. (1991), “Forward Induction and Sunk Costs Give Average Cost Pricing,” *Games and Economic Behavior*, 3, 221-236.

Schelling, T. (1960), *The Strategy of Conflict*, Harvard University Press.

Segendorff, B. (1998), “Delegation and Threat in Bargaining,” *Games and Economic Behavior*, 23, 266-283.

Sklivas, S. (1987), “The Strategic Choice of Managerial Incentives,” *Rand Journal of Economics*, 452-458.

Van Damme, E. (1989), “Stable Equilibria and Forward Induction,” *Journal of Economic Theory*, 48, 476-496.

Vickers, J. (1985), “Delegation and the Theory of the Firm,” *Economic Journal*, (supplement) 95, 138-147.

Walsh, C. E. (1995), “Optimal Contracts for Independent Central Bankers,” *American Economic Review*, 85, 150-167.