The Informed Principal with Agent Moral Hazard*

Daniel Clark†

November 8, 2021

Abstract

We study principal-agent settings where the principal has private information, both the principal and agent take actions, and the agent's action is subject to moral hazard. Unlike past work focusing on explicit contracts, we allow the principal to propose contracts that give them flexibility in their choice of future actions. We develop an adaptation of sequential equilibrium called *contracting equilibrium* for our principal-agent games, and prove its existence. In environments where the principal's type and agent's action are complements, we also apply a refinement called *payoff-plausibility*. The *principal-optimal safe outcomes*, which are analogs of the least-cost separating outcomes of signaling games, are always contracting equilibrium outcomes. They also provide an important payoff benchmark: Every principal type must obtain a weakly higher payoff from any payoff-plausible equilibrium. Moreover, if there are complementarities between the principal's type and their action, payoff-plausibility selects the principal-optimal safe outcomes when the principal is restricted to offering *deterministic* mechanisms. Otherwise, pooling between principal types can survive payoff-plausibility, and is more prevalent than would be predicted with explicit contracts.

Keywords: informed principal, moral hazard, flexible contracts, contracting equilibrium, principal-optimal safe outcomes, payoff-plausibility.

*The author wishes to thank Drew Fudenberg and Alexander Wolitzky for their invaluable guidance and suggestions, as well as Ian Ball, Dirk Bergemann, Alessandro Bonatti, Roberto Corrao, Glenn Ellison, Joel Flynn, Parag Pathak, Giacomo Lanzani, Stephen Morris, and Michael Whinston for helpful comments and conversations.

†Department of Economics, MIT. Email: dgclark@mit.edu.

1 Introduction

In many economic settings, a principal with private information interacts with an agent whose actions are subject to moral hazard. For example, a firm that offers a performance-based incentive contract to a prospective employee may have a better sense than the employee of how the employee's effort will translate into profit. If the compensation specified by the contract depends on the firm's profit, then the potential employee's perception of what the firm knows will be important for their decisions of whether to accept the employment offer and, if they join the firm, how hard to work.[1]

This paper develops a framework for studying mechanism design by informed principals in these settings, and shows how the resulting predictions differ from those made when ignoring principal private information, as well as from the predictions in informed principal environments without agent moral hazard. With an uninformed principal, there are two natural benchmarks. One is the *complete-information benchmark*, which obtains in the situation where the principal's information is common knowledge before contracting occurs. The other is the *ex-ante mechanism design benchmark*. This is what would be implemented by a principal who does not possess any asymmetric information before contracting, but will learn their type after the contract is accepted. The complete-information benchmarks are frequently not equilibrium outcomes because of incentive compatibility failures where "worse" principal types prefer to mimic "better" ones. Even when the ex-ante mechanism design benchmarks are consistent with equilibrium, they may implausibly rely on the agent believing that certain off-path contract proposals were made by "bad" principal types when only "good" types could reasonably gain from making the proposal. When this is the case, the ex-ante mechanism design benchmarks can be ruled out by refinement.

The literature studying mechanism design by informed principals has focused on

[1] Other examples with informed principals and agent moral hazard include: (1) A publisher who is more informed than a prospective author about the likely sales of a future book, and (2) An insurance company offering a menu of policies when the company has better information about the likelihood of various outcomes than a potential insuree.

settings without agent moral hazard, with the most relevant for comparison with our setting being the "common values" environments first studied in Maskin and Tirole [1992].[2] There the principal's information is payoff-relevant for the agent, but the agent has no action other than accepting or rejecting the principal's contract proposal. The key outcomes are the "Rothschild-Stiglitz-Wilson (RSW) allocations." An allocation is a perfect Bayesian equilibrium outcome if and only if it is individually rational for the agent and gives every principal type a weakly higher payoff than the RSW allocations. Moreover, only the RSW allocations survive natural equilibrium refinements.

Several differences emerge with the presence of moral hazard. One difference is that there can be equilibria that give lower payoffs than the *principal-optimal safe outcomes*, which are the analogs of the least-cost separating equilibria of traditional signaling games, as well as the generalizations of the RSW allocations to environments with moral hazard. The reason for this discrepancy is that, with moral hazard, the agent's beliefs not only affect their decision of whether to accept a given contract proposal but also their choice of action after a contract has been accepted. So even when the agent accepts a proposed contract that would lead to a principal-optimal safe outcome under certain agent beliefs, other beliefs might lead them to take worse actions for the principal types. However, in a broad class of environments, a natural equilibrium refinement restores the prediction that equilibria give higher payoffs than the RSW allocations/principal-optimal safe outcomes. Still, outcomes other than those which are principal-optimal safe survive refinement, including more efficient outcomes that involve pooling.

We now preview our framework and results in more detail. In our model, mechanisms constrain the future actions of the principal but not those taken by the agent. The interaction between the principal and agent consists of (1) the proposal by the principal of a contract detailing a mechanism, followed by the acceptance/rejection of the offer by the agent, (2) the implementation of the mechanism in the contract

[2]We discuss this literature in more depth later, along with papers that consider restricted contracts rather than full-fledged mechanism design in settings with informed principals and agent moral hazard.

and the resulting determination of the principal's action, and (3) the agent's ultimate choice of action after observing the results of the mechanism. It is the final stage (3) and the fact that a mechanism cannot directly constrain the agent's action that mark the presence of agent moral hazard.

Our mechanism design approach allows the principal to avoid committing to a single pure action by proposing *flexible* contracts that allow them a non-trivial choice over their future actions. This contrasts with the standard approach to informed principal settings with agent moral hazard in which the principal is assumed to only propose *explicit* contracts that precisely pin down the action they will take should the agent accept the offer. Flexible contracts are an appropriate and realistic assumption for the informed principal setting. As noted by Segal and Whinston [2003], publishers often use contracts with multiple options concerning publication and copyrights of books. Similarly, a firm may offer a contract to a prospective employee that places some constraints on the possible tasks the firm could assign or the exact nature of how the firm will compensate the employee, but does not completely narrow down the firm's possible actions.

We consider two classes of flexible contracts: The unrestricted class which can commit to any mechanism, and a restricted class which can only commit to *deterministic* mechanisms. Deterministic mechanisms are essentially menus of pure actions from which the principal will later choose should the agent accept the associated contract, whereas general mechanisms allow the principal to choose between non-degenerate distributions over actions. The general-mechanism approach is consistent with the standard mechanism design literature, and often affords useful analytical tools, such as the Inscrutability Principle (Myerson [1983]). However, the deterministic-mechanism approach seems likely to be of applied interest in some settings and often leads to narrower predictions than the general-mechanism approach.

As in most of the principal-agent literature, we allow for infinite action spaces for both the principal and the agent. Owing to this and the fact that the space of possible mechanism proposals is infinite, the traditional solution concept of sequential

equilibrium cannot be directly applied to our game.[3] Instead, we develop and use an adaptation of sequential equilibrium called *contracting equilibrium*. These equilibria always exist, though this is not an immediate consequence of previous existence results. The proofs for the existence of contracting equilibria, as well as the solution concept itself, may be useful for analyses of other mechanism design games, such as those with multiple agents or agent adverse selection in addition to agent moral hazard.

Like sequential equilibrium, contracting equilibrium is often excessively permissive, so we also employ a refinement in the vein of the Intuitive Criterion called *payoff-plausibility*. Payoff-plausibility makes strong predictions, and it frequently rules out the ex-ante mechanism design benchmark. Where it is applied, payoff-plausibility is a consequence of existing refinements, appropriately adapted for our principal-agent games, such as robust neologism proofness (Clark [2021]) and strongly justified communication equilibrium (Clark and Fudenberg [2021]). We discuss these connections in more detail in Section 8.

We focus much of our analysis on *monotone-concave-supermodular (MCS)* environments. MCS environments satisfy supermodularity conditions that capture complementarities in the effects of the principal's type and the agent's action on payoffs. These conditions are satisfied in many settings of interest, and they lead to a tendency for higher principal types to separate from lower principal types. Indeed, the principal-optimal safe outcomes are always payoff-plausible contracting equilibrium outcomes. Moreover, payoff-plausibility selects only outcomes which give every type of the principal a weakly higher payoff than they obtain from the principal-optimal safe outcomes. If there are complementarities between the principal's type and their action on top of the standard complementarities in MCS environments, payoff-plausibility selects the principal-optimal safe outcomes when the principal is restricted to offering deterministic mechanisms. Otherwise, pooling between principal types can survive payoff-plausibility with flexible contracts. In contrast, when only explicit contracts can

[3] As we explain in Section 4.3.1, perfect Bayesian equilibrium is also inappropriate here, because the game where general mechanisms can be proposed does not have perfectly observed actions.

be proposed, payoff-plausibility typically predicts complete separation.

The paper features a running example: An informed-principal version of a canonical firm and worker problem, where the firm is more informed about the potential profitability of the task on which they seek to employ the worker. Here the complete-information benchmark is inconsistent with contracting equilibrium. Moreover, payoff-plausibility eliminates not only the ex-ante mechanism design benchmark, but all equilibrium outcomes that are Pareto-optimal from the perspective of the various firm types.

The remainder of the paper proceeds as follows. Section 2 discusses the related literature, while Section 3 formalizes and analyzes the firm and worker example. Section 4 then presents our general framework and establishes the existence of contracting equilibria for all payoff environments. In Section 5, we explore various properties of contracting equilibria, before applying the refinement of payoff-plausibility to MCS environments in Section 6. Section 7 compares informed principal environments with moral hazard to those without moral hazard. Section 8 then discusses the foundations payoff-plausibility has in robust neologism proofness and strongly justified communication equilibrium. Section 9 concludes.

2 Related Literature

Most analyses of principal-agent problems assume that the agent knows everything that the principal does. Myerson [1983] introduced the study of mechanism design by an informed principal. It analyzed a general setting in which the principal and agents can all posses asymmetric information and the agents' actions may be subject to moral hazard. Unlike our focus, most of the analysis in Myerson [1983] was from the perspective of cooperative rather than non-cooperative game theory. It established the existence of "expectational equilibria" under the assumption that all action spaces are finite. In contrast, throughout our analysis, we allow for infinite action spaces. Moreover, we give the agent the choice of whether to accept the principal's proposed

contract, whereas Myerson [1983] assumed that all agents were in a relationship with the principal even before mechanism proposal.

The literature studying the design of general mechanisms by informed principals has focused on two settings: those with *private values* and those with *common values*. In both of these settings, there are no actions for the agent to take, so moral hazard is not present. The private values setting, first studied by Maskin and Tirole [1990], concerns situations in which both the principal and agent have asymmetric information, but each player's information is only relevant for their own payoff. Maskin and Tirole [1992] defined the common-value informed principal problem to mean that only the principal has private information. The common values setting is thus a special case of the general informed principal with agent moral hazard setting where the agent's action space is null. The principal-agent game we develop adapts the games in Maskin and Tirole [1990, 1992] to settings with moral hazard, and we study the differences that arise from moral hazard relative to settings with private values and especially those with common values.[4]

Unlike the general analysis presented here, Beaudry [1994], Inderst [2001], Chade and Silvers [2002], Bénabou and Tirole [2003], and Martimort and Sand-Zantman [2006] each studied specific settings with informed principals and agent moral hazard, and assumed that contracts commit the principal to a single (pure) action. Beaudry [1994] and Inderst [2001] in particular studied settings like the example presented in Section 3. Wagner et al. [2015], Bedard [2017], and Mekonnen [2021] allowed for unrestricted contracts, but limited attention to very special environments.[5]

Payoff-plausibility is a consequence of two signaling game refinements, *robust neologism proofness (RNP)* (Clark [2021]) and *strongly justified communication equilibrium*

[4]Cella [2008] and Mylovanov and Tröger [2012, 2014] also studied private values, and Inderst [2005], Severinov [2008], Balkenborg and Makris [2015], Koessler and Skreta [2016], DeMarzo and Frankel [2020], and DeMarzo et al. [2020] analyzed the common-values case.

[5]Wagner et al. [2015] and Mekonnen [2021] assumed the agent's first-best action is independent of the principal's type, and analyzed when the principal types could achieve the same payoff as if their information were common knowledge. Bedard [2017] gave a sufficient condition for (what we call) flexible contracts to enable outcomes that give both principal types higher payoffs than the least-cost separating outcome when there are two principal types and two actions for the agent.

(SJCE) (Clark and Fudenberg [2021]), when they are extended to MCS informed principal environments.[6] RNP and SJCE were intended to capture effects of communication from senders to receivers in signaling games. We discuss these refinements and their relationship to payoff-plausibility in more detail in Section 8.

3 Firm and Employee Example

3.1 Setup

Consider a firm (principal) attempting to hire a potential employee (agent) to work on a task. Both parties are risk neutral. The firm has private information $\theta \in \{2, 4\}$ about the profitability or quality of the task, where θ is equally likely to be 2 or 4. If the employee joins the firm, they will choose some effort level $e \in \mathbb{R}_+$, at cost $e^2/2$, that affects the probability of the task being successful. The firm will pay a transfer $t \in \mathbb{R}$ to the employee as well as a share $s \in [0, 1]$ of the profits. The expected profit given θ and e is θe, so the utility functions of the firm and employee are $U(\theta, s, t, e) = \theta(1 - s)e - t$ and $V(\theta, s, t, e) = \theta se - e^2/2 + t$, respectively. Both the firm and employee have an outside option that gives payoff 0.

To attempt to hire the employee, the firm offers them a contract that specifies how s and t will be determined. In this example, the principal's actions are simply the payment scheme (s, t); more generally, they can be things like task assignment or an investment. The contract cannot directly constrain the effort the employee exerts.

The standard approach of the literature to informed principals with agent moral hazard, seen for instance in Beaudry [1994] and Inderst [2001], requires that the firm's contract commit to a single action, which in this case is a payment scheme. The contracts can also contain a recommended action, so they correspond to (s, t, e) triples. With such an *explicit* contract, the agent knows precisely what share of profits and

[6] Maskin and Tirole [1990, 1992] and Mylovanov and Tröger [2012, 2014] applied notions of neologism-proofness (Farrell [1993]), which is a stronger refinement than RNP and is not guaranteed existence in general informed principal environments.

transfer the firm will implement should the agent accept.

This modeling approach does not allow *flexible* contracts, which are both plausible and observed in the real world. In the present example, the firm might want to leave themself some flexibility in the contract, e.g. about how much of the employee's compensation will be governed by profit sharing or transfers, rather than completely pinning down their future actions.[7]

We study two classes of flexible contracts. The first, in keeping with the mechanism design literature, allows the firm's contract to commit to any *mechanism* for determining their future actions. Formally, a mechanism corresponds to some pair (μ, M_F), where M_F is the message space of the firm, and $\mu : M_F \to \Delta([0,1] \times \mathbb{R} \times \mathbb{R}_+)$ maps messages into distributions over (s, t, e). Explicit contracts committing to a specific (s, t, e) are a trivial example of a mechanism, but there are many other ways of determining the firm's profit share and transfers that correspond to more complex mechanisms and less explicit contracts.

We also study the informed principal problem when contracts can only offer *deterministic* mechanisms. In the firm and employee setting, these are mechanisms (μ, M_F) in which $\mu(m_F)$ is a degenerate distribution putting probability 1 on some (s, t, e) for each $m_F \in M_F$.[8] They are contractually forbidden from implementing any other (s, t, e). Such mechanisms still allow for ambiguity in the contract, but do not require the commitment (or complexity) needed to implement a mechanism where the firm's message determines a non-degenerate probability distribution from which (s, t, e) is ultimately drawn.

[7]Note that a contract is flexible only if it gives the firm a non-trivial choice over their future actions.

[8]Equivalently, a deterministic mechanism here can be thought of as the menu of pairs of profit shares and transfers from which the firm can choose. Outside the specific context of this example, a deterministic mechanism is simply the menu of principal action and recommendation pairs that the principal could implement should the agent accept the contract.

3.2 Benchmarks

Before analyzing the equilibria of our contracting game, we first discuss two benchmark solutions for contracting with symmetric information, as well as the hypothetical situation in which the firm directly controls the employee's effort. These will enable us to compare the predictions that emerge with an informed firm with those when the firm is uninformed, and to compare the predictions made with and without moral hazard.

The *complete-information benchmark* is the outcome that would occur if the firm's type were commonly known to be θ. Here the standard solution is that the employee receives all of the profits ($s = 1$), the employee exerts first-best effort level ($e = \theta$), and the firm extracts all of the surplus ($t = -\theta^2/2$). This results in payoffs of 2 to the type 2 firm, 8 to the type 4 firm, and 0 to the employee regardless of the firm's type. This outcome is not possible with asymmetric information, because the type 2 firm would strictly prefer to mimic the type 4 firm, which would let them extract a higher fee from the employee.

The *ex-ante mechanism design benchmark* is the outcome that would occur if the firm could propose a contract ex-ante before learning their type. Here the solution maximizes the firm's ex-ante expected utility subject to incentive compatibility and individual rationality constraints for the worker as well as incentive compatibility of the interim firm types. Again, the solution has the employee receiving a full share of profits and exerting the first-best effort level regardless of the firm types. However, here both firm types extract the same fee of $t = -5$ and thus receive expected utilities of 5, equal to the total expected surplus. As we will see, while this outcome is consistent with our notion of contracting equilibrium, it is not a *plausible* equilibrium outcome.[9]

To see what happens in the absence of employee moral hazard, suppose that the payoffs of the firm and employee are as before, except now the firm has control over the effort level the employee exerts. (Alternatively, we could keep control of effort with the employee, but have the chosen effort level be directly observable and con-

[9]For a prior distribution with probability greater than 47/54 on $\theta = 2$, the ex-ante mechanism design benchmark is not even a contracting equilibrium outcome.

tractible.) In this case, the unique equilibrium payoffs coincide with those in the complete-information benchmark.[10] The reason is that each principal type can secure a payoff arbitrarily close to their complete-information benchmark by setting $s = 0$, assigning the employee the same effort as in the benchmark, and giving the employee a slightly greater total payment than in the benchmark. Moreover, given that every principal type is attaining a weakly higher payoff than their complete-information benchmark, no principal type could be attaining a strictly higher payoff. Otherwise, the firm would, in ex-ante expectation, be extracting more than the maximum total expected surplus, so the agent's total expected utility would be strictly negative.

3.3 Equilibria

We now return to the setting with an ex-ante informed firm and employee moral hazard. We first consider the possible equilibria when the firm can only propose explicit contracts. Essentially, this amounts to a standard signaling game with a slightly more convoluted timeline. First, the firm observes θ and then proposes a contract. Subsequently, the employee either accepts or rejects the offer. If the employee rejects, both parties get a payoff of 0. If instead the employee accepts, the employee will then exert some effort e, after which profits and payoffs are realized.

Under an adaptation of sequential equilibrium to games with infinite action spaces, the possible pairs of firm-type equilibrium payoffs, where $U(\theta)$ denotes the equilibrium payoff of type θ, are given in Figure 1.

To understand the possible equilibrium payoff pairs, observe that the type 2 firm can never get a lower payoff than 2, their complete-information benchmark. The reason is the firm can offer a contract corresponding to $(s, t) = (1, 2-\varepsilon)$ for some $\varepsilon > 0$, which amounts to a perturbation of their optimal contract with complete information. Such a proposal is guaranteed to be accepted and result in a payoff of $2-\varepsilon$ to the firm. This holds for all $\varepsilon > 0$, so the firm can always get arbitrarily close to a payoff of 2. Moreover,

[10] The "Rothschild-Stiglitz-Wilson (RSW) allocations" in this example are all incentive compatible outcomes that result in precisely these payoffs.

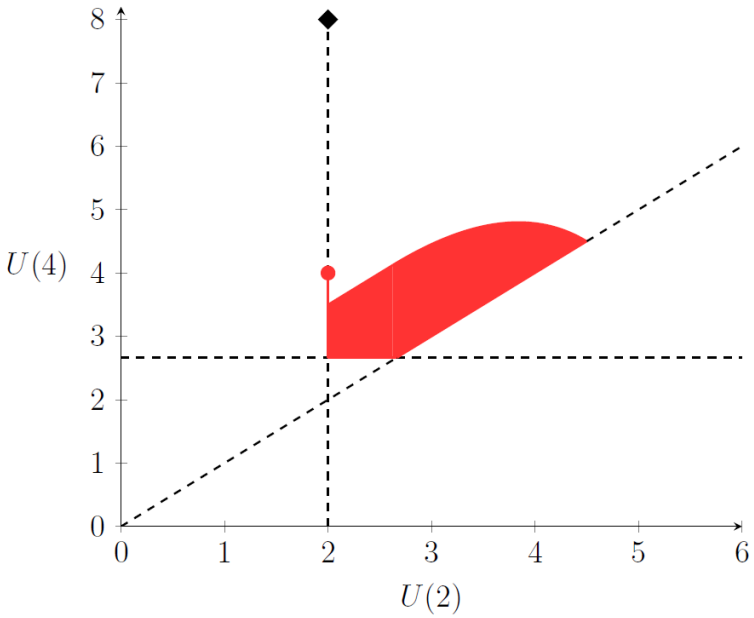

Figure 1: The red region depicts the possible equilibrium payoff pairs. The diamond at $(2, 8)$ denotes the payoffs of the firm types in the complete-information benchmark.

the lowest payoff that the type 4 firm can be held to is $8/3$, which comes from having the employee believe $\theta = 2$ following any off-path contract proposal. Additionally, the high type firm can never get a lower equilibrium payoff than the low type firm.

Having explained the various lower bounds on the set of equilibrium payoff pairs, we now turn to understanding its upper envelope. The dot at $(2, 4)$ corresponds to the least-cost separating outcome. In this outcome, the type 2 firm extracts the full surplus as with complete information, while the type 4 firm offers a higher transfer of $t = 0$ and a lower profit share of $s = 1/2$, leading the employee to exert effort $e = 2$. This is also the *principal-optimal safe outcome*, an object that will feature in much of our analysis. Here the principal-optimal safe outcome maximizes the payoff of both firm types across the outcomes in which the employee's decision of whether to join the firm and subsequent effort choices are always optimally calibrated to the firm's type.

All points to right of $U(2) = 2$ involve pooling. The reason is that the payoff of the type 2 firm in all separating equilibria is 2. Thus, in a pooling equilibrium where $U(2) > 2$, there must be some (s, t) played with positive probability by both firm types

where the employee's posterior puts at least probability $1/2$ on $\theta = 2$. This fact enables the formulation of a constrained optimization problem that maximizes the payoff of the type 4 firm subject to the type 2 firm's payoff equaling $U(2)$, employee incentive compatibility, and an individual rationality constraint that averages across both $\theta = 2$ and $\theta = 4$. The solution to this problem, the analysis of which is given in Section OA.1.2, characterizes the upper envelope in the $U(2) > 2$ region.

Now we consider the possible equilibria when the firm can propose flexible contracts.[11] The timing of the corresponding game is the same as when only explicit contracts can be proposed, with the following exception. Should the employee accept the firm's contract offer, the firm will choose some message $m_F \in M_F$. After this, (s, t) is drawn according to $\mu(m_F)$. The employee then observes the resulting (s, t) before exerting some effort level e, following which profits and payoffs are realized. Later we will develop *contracting equilibrium*, an adaptation of sequential equilibrium that applies to our principal-agent game with flexible contracts. Figure 2 depicts the contracting equilibrium payoffs with flexible contracts as well as those possible when only explicit contracts can be proposed.

Observe that, with flexible contracts, the type 4 firm cannot be held to same minimum payoff as with explicit contracts. The reason is that the type 4 firm can always get payoffs strictly higher than $8/3$ because of the richer space of deviations. In particular, there are contracts in which all the sequential continuation equilibria following their proposal give a higher payoff than $8/3$ to the type 4 firm. For example, consider a contract with two messages, where the first message results in $(s_1, t_1) = (1, -199/100)$ with probability 1 and the second results in $(s_2, t_2) = (2/3, -1)$ with probability 1. If the contract were proposed and accepted, then the type 4 firm would always select (s_2, t_2), and obtain a payoff of at least $25/9$. The type 2 firm would only select (s_2, t_2) when it induces the employee to exert effort at least $e = 297/200$. Given (s_2, t_2) and any belief that would induce the employee to exert effort higher than $e = 297/200$, the

[11] The results with either general mechanisms or deterministic mechanisms are the same in this example, but this is not in general true.

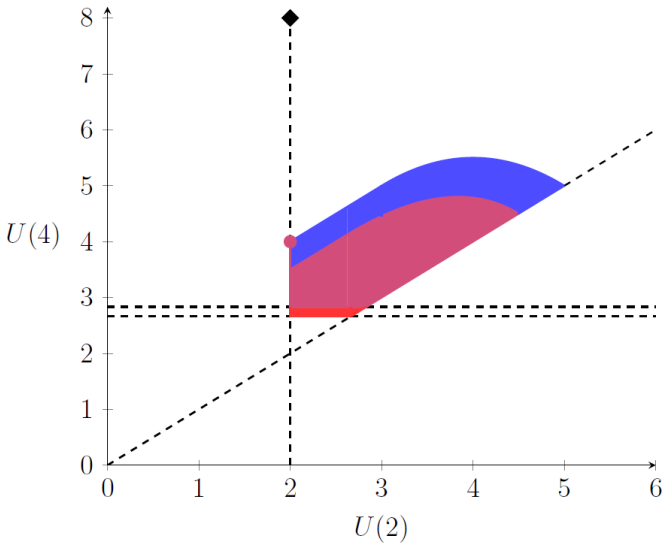

Figure 2: The blue region depicts the equilibrium payoffs that can only be sustained with flexible contracts, while the red region consists of equilibrium payoffs that can be sustained only with the restriction to explicit contracts. Equilibrium payoffs that can be sustained with both classes of contracts are purple.

employee's conditional expected utility must be at least $(297/200)^2/2 - 1 > 0$. Moreover, the employee's expected utility conditional on (s_1, t_1) is always strictly positive. Thus, the employee's expected utility from accepting the proposal is strictly positive in every sequential continuation equilibrium. So the type 4 firm's payoff is at least 25/9 in every sequential continuation equilibrium following the proposal of this contract.[12]

Additionally, with flexible contracts, the upper envelope is higher and smooth. It also can be found through a constrained optimization problem, details of which are in Section OA.1.1. However, unlike the case with explicit contracts, all the points on the upper envelope with flexible contracts correspond to outcomes where the agent correctly anticipates the principal's type when they choose their effort. In particular, any payoff on the upper envelope can be realized in an outcome where, conditional on the low type $\theta = 2$, the employee receives the full profit share $s = 1$ and exerts efficient effort

[12] All payoffs in the purple region weakly above $U(4) = 3$ can be sustained in contracting equilibrium with flexible contracts as well as with explicit contracts, but it is not known which of the payoffs in the purple region between $U(4) = 25/9$ and $U(4) = 3$ are consistent with contracting equilibrium when flexible contracts can be proposed. A similar qualification holds for the right panel of Figure 3 below.

level $e = 2$, and conditional on the high type $\theta = 4$, the employee exerts optimal effort $e = 4s$ for the corresponding profit share s. Intuitively, if this were violated, the payoffs of both the firm and the worker when $\theta = 2$ could be weakly increased by increasing the surplus to its maximum value of 2 and appropriately dividing it. Moreover, the payoff of the high type $\theta = 4$ could only improve from not being mistaken for the low type. The reason why these outcomes are possible with flexible contracts is that they can be achieved with both firm types proposing the same contract. This leads the employee to be willing to accept a relationship with a type 2 firm despite regretting it later.

There are many equilibria with both explicit and flexible contracts, but not all the equilibria are reasonable. Consider for instance equilibria with flexible contracts in which both firm types obtain a payoff of 5, as in the ex-ante mechanism design benchmark from Section 3.2. (Graphically, these equilibria correspond to the star in the right-hand plot of Figure 3.) We should expect the high type firm to obtain a strictly higher payoff than the low type firm, because the high type should be able to credibly signal their identity to the employee when the prevailing equilibrium has both types receiving the same payoff. For example, suppose the type 4 firm proposed a contract committing to $(s, t) = (1/2, -1.5)$. Every undominated response of the employee to such a contract would involve effort levels less than 2 and thus give the type 2 firm a strictly lower payoff than 5; however, the employee accepting and exerting effort 2, as they would if they knew $\theta = 4$, would give the type 4 firm a strictly higher payoff of 5.5. Because of this, *payoff-plausibility*, which is formally defined in Section 4.3, rules out the equilibria in which both firm types obtain a payoff of 5. More generally, payoff-plausibility eliminates equilibria when there is some type θ and a contract that, when the agent responds as if the type were θ, would give the type θ principal a strictly higher payoff than the equilibrium and all types below θ a strictly lower payoff.[13]

Payoff-plausibility selects precisely the green payoff pairs depicted in Figure 3.

[13]With two firm types, the Intuitive Criterion (Cho and Kreps [1987]) is equivalent to payoff-plausibility. With more types, the Intuitive Criterion is usually much weaker.

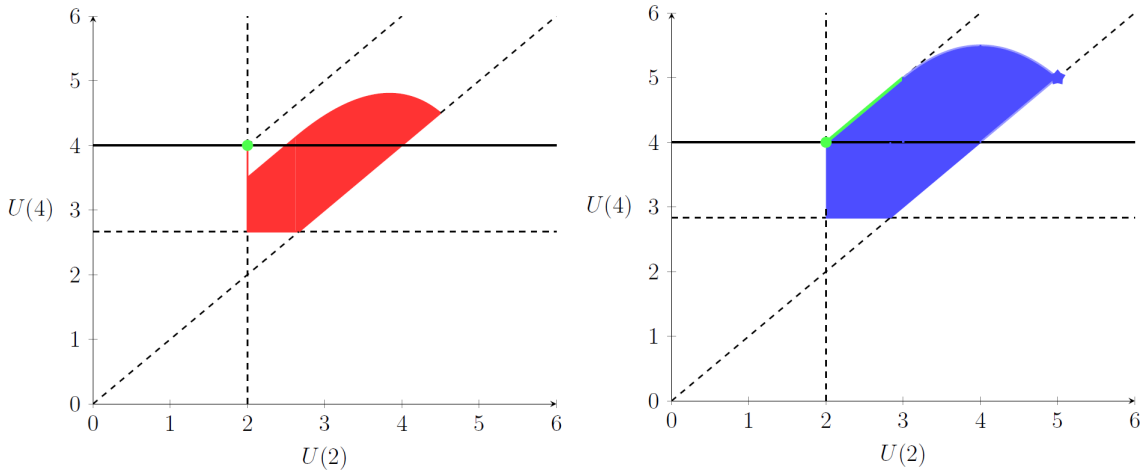

Figure 3: The left-hand figure depicts equilibrium payoffs with explicit contracts, with plausible payoffs in green and all other payoffs in red. The right-hand figure depicts equilibrium payoffs with flexible contracts, with plausible payoffs in green and all other payoffs in blue.

These are the payoffs that correspond to outcomes that can be obtained from the least-cost separating outcome by uniformly reducing the transfers paid by the firm types. With flexible contracts, there is a non-singleton line segment of such payoffs, as shown in Section OA.1.3, while there is only one such payoff with explicit contracts. As we will see in Section 6, in a broad class of environments nesting this example, payoff-plausibility selects the least-cost separating equilibria when only explicit contracts can be proposed, but frequently allows multiple equilibrium outcomes with flexible contracts.

Intuitively, payoff-plausibility eliminates any equilibrium whose payoffs are beneath the upper envelope with flexible contracts because, in any such equilibrium, the type 4 firm could propose a contract corresponding to a point on the upper envelope that is above and to the left of the equilibrium payoffs. (This holds in the example here for both when flexible contracts can be proposed and when only explicit contracts can be proposed, because the payoffs on the upper envelope with flexible contracts can be attained with deterministic contracts where the type 4 firm chooses a single payment scheme.) The type 2 firm would do worse by such a proposal, while the type 4 firm would do better if the employee were to respond under the belief that $\theta = 4$.

The requirement that plausible payoffs lie on the upper envelopes holds generally in a broad class of environments with two types. It is not clear that this always extends with more than two types. However, there are general thresholds that the payoffs in payoff-plausible equilibria must always meet. In particular, every principal type must always obtain a weakly higher payoff than they do in the principal-optimal safe outcomes (or least-cost separating outcomes if only explicit contracts can be proposed). In this example, this amounts to the requirement that the type 4 firm always obtain a weakly higher payoff than 4, which is the reason for the horizontal lines at $U(4) = 4$ in Figure 3.

Further, note that no equilibrium that is Pareto-optimal for the firm types survives payoff-plausibility. This can be seen graphically by the fact that all the green payoffs are to the left of the peaks in the upper envelopes. The reason is that, to sustain relatively high equilibrium payoffs to the type 2 firm, the type 4 firm must give both a high transfer t and a high profit share s. (The increasing levels of s are reflected in the bending of the upper envelopes.) However, the high type would do better by offering a contract with a reduced profit share s and increased transfer t.

4 Framework and Existence Results

4.1 Primitives

The principal's type is $\theta \in \Theta$, where $\Theta = \{\theta_1, ..., \theta_N\}$ is a finite type space. There is a full-support prior distribution over Θ given by $\lambda \in \Delta(\Theta)$.[14] If a relationship is formed, the principal's action set is the compact metric space X, with $x \in X$ denoting a typical principal action, while the agent's action set is the compact metric space Y, with $y \in Y$ denoting a typical agent action. Here, a principal action x could represent an investment, task assignment, incentive scheme, or monitoring system, and an agent action y could represent effort level, type of work, or social behavior. In addition to

[14]Throughout the paper, we denote the set of probability distributions over a set Ω by $\Delta(\Omega)$, and, whenever Ω is a metric space, we endow $\Delta(\Omega)$ with the topology of weak convergence.

choosing an x, the principal makes an *action recommendation* to the agent, which the agent may or may not follow. An action recommendation r lies in a compact metric space R. We assume that $\Delta(\Theta) \times [0,1]^{|\Theta|+1} \times \Delta(Y) \subseteq R$. This allows the principal to recommend any mixed action to the agent, since there is a dimension of R that contains all of $\Delta(Y)$, as well as describe various possible beliefs over the principal type and mixture probabilities for the agent and the principal types. If a relationship is formed, $U(\theta, x, y)$ and $V(\theta, x, y)$ are the utilities of the principal and agent, respectively, when the principal's type is θ, the principal takes action x, and the agent takes action y. Both utility functions $U : \Theta \times X \times Y \to \mathbb{R}$ and $V : \Theta \times X \times Y \to \mathbb{R}$ are continuous. Note that neither the payoffs of the principal nor the agent depend on the action recommendation r.

If instead the principal and agent do not form a relationship, then both realize their outside options; the payoffs to all types of the principal and the agent from their outside options are normalized to 0. For convenience, we assume that there is some action $x_o \in X$ that the principal can take when a relationship is formed with the agent that automatically results in both parties realizing their outside option payoff: that is, $U(\theta, x_o, y) = 0$ and $V(\theta, x_o, y) = 0$ for all $\theta \in \Theta$ and $y \in Y$. Such an action is present, for instance, when the principal has the ability to end their relationship with the agent immediately upon its inception. This will let us appeal to the Inscrutability Principle of Myerson [1983], which states that, with unrestricted mechanisms, it is without loss of generality to assume that, on the path of play, all principal types propose the same mechanism.

In many, if not most, principal-agent relationships of interest, the principal can pay a transfer to the agent. To analyze such settings, we assume that the principal's action space is of the form $X \times T$, where, for some large $\bar{t} \in \mathbb{R}_+$, $T = [-\bar{t}, \bar{t}]$ represents the space of possible transfers and X represents the space of other actions the principal could take.[15]

[15] We bound the transfers to maintain compactness, but this can be relaxed without much difficulty.

Definition 1. *An environment has* **transfers** *if there are continuous functions* $u : \Theta \times X \times Y \to \mathbb{R}$*,* $v : \Theta \times X \times Y \to \mathbb{R}$*, and* $g : T \to \mathbb{R}$ *such that* $U(\theta, x, t, y) = u(\theta, x, y) - t$ *and* $V(\theta, x, t, y) = v(\theta, x, y) + g(t)$ *for all* $\theta \in \Theta$*,* $x \in X$*,* $t \in T$*, and* $y \in Y$*.*

We assume that g is weakly increasing and continuous, and that there is some $x_o \in X$ such that $u(\theta, x_o, y) = v(\theta, x_o, y) = 0$ for all $y \in Y$. This allows for the classic setting of quasilinear transfers, i.e. $g(t) = t$ for all $t \in T$, as well as the possibility e.g. that the agent is risk averse over income.

Moreover, as with the informed firm and employee example, there are often complementarities in the effects of the principal's type and the agent's action. For this, we assume that the principal's types are ordered so that $\theta_1 < ... < \theta_N$, and that the agent's action space is an interval of real numbers, $Y = [\underline{y}, \overline{y}]$.

Definition 2. *An environment with transfers is* **monotone-concave-supermodular (MCS)** *if*

1. *Monotone:*
   (a) $u(\theta, x, y)$ *is weakly increasing in* y *for all* $\theta \in \Theta$ *and* $x \in X$*.*
   (b) *For all* $\theta, \theta' \in \Theta$*,* $x \in X$*, and* $y, y' \in Y$*,* $u(\theta, x, y) \geq u(\theta, x, y')$ *if and only if* $u(\theta', x, y) \geq u(\theta', x, y')$*.*
   (c) $u(\theta, x, y)$ *and* $v(\theta, x, y)$ *are weakly increasing in* θ *for all* $x \in X$ *and* $y \in Y$*.*
2. *Concave:*
   (a) g *is weakly concave.*
   (b) $y^*(\tilde{\lambda}, x) \equiv \arg\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[v(\theta, x, y)]$ *is singleton for all* $\tilde{\lambda} \in \Delta(\Theta)$ *and* $x \neq x_o$*.*
3. *Supermodular:*
   (a) $y^*(\tilde{\lambda}, x)$ *is weakly increasing in* $\tilde{\lambda}$ *according to the FOSD partial ordering of* $\Delta(\Theta)$ *for all* $x \neq x_o$*.*
   (b) *For all* $\theta, \theta' \in \Theta$*,* $x \in X$*, and* $y, y' \in Y$ *such that* $\theta > \theta'$ *and* $y > y'$*,* $u(\theta, x, y) - u(\theta, x, y') \geq u(\theta', x, y) - u(\theta', x, y')$*, with the inequality holding strictly when* $u(\theta', x, y) - u(\theta', x, y') > 0$*.*

The monotonicity criteria state that (a) the principal always (weakly) prefers a higher agent action, (b) holding fixed the principal's action, the principal types share the same preference over the agent's action, and (c) both the agent and the principal gain (weakly) more by forming a relationship when the principal's type is higher. The concavity criteria require the agent's utility be (weakly) concave in the transfer they receive, and their best response be a singleton, which is necessarily the case for v strictly concave in y. The first supermodularity condition says that the agent's best response is weakly increasing in their posterior belief about the principal's type.[16] Finally, the second supermodularity criterion requires that the difference in principal utility from a higher agent action, holding fixed the principal's action, is higher for a higher principal type, and strictly so when the lower principal type strictly gains from the higher agent action.

4.2 Contracts, Mechanisms, and the Principal-Agent Game

At the beginning of their interaction, the principal offers the agent a contract that specifies the mechanism that will be used if the agent accepts. The principal can constrain their own action through the mechanism they design, but they are unable to impose any direct constraints on the action of the agent: If the principal and agent form a relationship, the agent will be free to take any action they desire upon observing the principal's action and recommendation pair. Formally, a mechanism consists of a message space for the principal of the form $M_P = \{1, ..., M\}$ for some $M \in \mathbb{N}$, and a mapping $\mu : M_P \to \Delta(X \times R)$ taking principal messages into distributions over principal action and recommendation pairs.[17] We denote this mechanism by (μ, M_P), and we let $\mathcal{M}$ denote the set of all possible mechanisms. (Note that even if the action

[16] A sufficient condition for this is that $v(\theta, x, y)$ be differentiable with the derivative $\frac{\partial v}{\partial y}(\theta, x, y)$ weakly increasing in θ for all $x \neq x_o$ and y.

[17] A mechanism could also contain a message space for the agent M_A and determine the distribution over the principal's action and recommendation pairs using the messages of both the principal and agent. This would be especially natural in situations where the agent possesses some hidden information. For notational simplicity, we ignore agent message spaces throughout; however, our results extend to this setup.

and recommendation sets were finite, there would be infinitely many mechanisms since the number of messages M can be any number in $\mathbb{N}$ and, as long as X or R were non-singleton, there would be infinitely many stochastic distributions over $X \times R$.)

Formally, the ***principal-agent game*** proceeds as follows. The principal observes their type θ, and proposes a mechanism (μ, M_P) to the agent. The agent observes the principal's choice of mechanism and then decides whether to accept the offer. If the agent rejects the offer, the game ends with the principal and agent each realizing their outside options. If instead the agent accepts the offer, the principal and agent form a relationship. Subsequently the principal chooses a message $m_P \in M_P$, and the principal's action and recommendation pair (x, r) is then drawn according to $\mu(m_P)$. The agent then observes the principal's action and recommendation pair and responds with an action y. After this the payoffs are realized.

Deterministic Mechanisms Up to this point, we have assumed that the principal can implement a mechanism that results in a stochastic determination of their action. This is consistent with the standard mechanism design literature, as well as the literature studying informed principals in settings without moral hazard. We will also consider a principal-agent game where the principal can only propose *deterministic* mechanisms. These are mechanisms (μ, M_P) in which $\mu(m_P)$ is a degenerate distribution for each $m_P \in M_P$, i.e. $\mu(m_P) = \delta_{(x,r)}$ for some $(x, r) \in X \times R$.

Unlike most papers that have studied informed principal settings with agent moral hazard, both the general-mechanism and deterministic-mechanism versions of the principal-agent game allow the principal to avoid committing to a single pure action by proposing flexible contracts. This strikes us as more realistic than insisting that the contract must specify exactly how the principal would act. Whether it is realistic that the principal can directly commit to non-degenerate distributions over their actions is less clear and likely depends on the application, which is why we consider both versions of the principal-agent game. Throughout the paper, there will frequently be a pair of definitions for a given concept: one for the principal-agent game with general mechanism

proposals and one for the game with the deterministic mechanism restriction.

4.3 Solution Concepts

4.3.1 Contracting Equilibrium

Because the space of mechanism proposals is infinite, we cannot apply the standard definition of sequential equilibrium to our principal-agent games. Perfect Bayesian equilibrium (Fudenberg and Tirole [1991]) cannot be applied to the general-mechanism game, because it does not have observed actions, as there are mechanisms in which the distributions over (x, r) pairs induced by distinct messages overlap.

Instead, we develop an adaptation of sequential equilibrium called *contracting equilibrium*. We defer the formal definitions to Appendix A, but here we discuss some of the important aspects of contracting equilibrium. As in PBE and sequential equilibrium, each player has a strategy, and the uninformed agent has a belief about the principal's type at each of their information sets. Each principal type plays optimally: Their expected payoff must be no less than the payoff they could get by playing an arbitrary mechanism and subsequent message given the play of the agent. Moreover, the agent plays optimally: For each mechanism, their acceptance decision and their subsequent choice of actions conditional on the various principal action-message pairs maximize their expected utility given their posterior belief about the principal's type. Additionally, the agent's posterior belief at the mechanism proposal stage must come from a regular conditional probability distribution obtained from their prior and the mechanism proposal rules of the principal types. Likewise, the agent's posterior belief upon observing a given principal action-message pair must be consistent with the agent's interim belief about the principal's type when the corresponding mechanism is proposed and the distributions over messages used by the various principal types.[18]

[18]The consistency requirement applied at this stage is essentially that of sequential equilibrium in the subgame after the mechanism has been accepted. We can invoke it here because the message space in any mechanism is finite.

Theorem 1. *Contracting equilibria exist in both the general-mechanism and deterministic-mechanism principal-agent games.*

Appendix B presents the proof of Theorem 1, which takes sequences of games, with finite approximations of X and Y and finite sets of mechanisms that can be proposed. It shows that the limits of the contracting equilibrium outcomes in these games are contracting equilibrium outcomes in the true games. Broadly this involves showing that the limits of the contracting equilibrium outcomes are consistent with both the "off-path" and "on-path" requirements of contracting equilibrium. It seems likely that the proof techniques involved may be useful for establishing equilibrium existence in similar games, such as other informed principal settings.

The off-path requirement that must be satisfied is that, for every mechanism (μ, M_P), there is a sequential continuation equilibrium that deters every principal type from proposing it. The difficulty here this is that the correspondence mapping mechanisms into sequential continuation equilibria is not upper hemicontinuous, so we cannot simply take an arbitrary limit of sequential continuation equilibria that follow the proposals of a sequence of mechanisms in the approximating games that converges to a given (μ, M_P). However, we show that, for every (μ, M_P), there is a sequence of well-calibrated mechanisms in the sequence of finite games whose sequential continuation equilibria converge to sequential continuation equilibria after (μ, M_P) is proposed.

The on-path requirement is that the outcome be consistent with a valid profile of mechanism proposal distributions and rule governing sequential continuation equilibria following the proposal of each mechanism. An obstacle here is that the space of mechanisms is not compact. With general mechanisms, this poses little difficulty because of the Inscrutability Principle, which enables the construction of a single mechanism that is proposed by all principal types and induces the outcome of interest. With deterministic mechanisms the Inscrutability Principle does not hold. Instead, we show that every equilibrium in the approximation games can be realized with principal types proposing *binary* and *obedient* mechanisms on-path. Binary mechanisms have precisely

two messages for the principal, while the recommendations in an obedient mechanism must be consistent with a sequential continuation equilibrium following its proposal. The set of binary and obedient mechanisms is compact, so standard theorems regarding the convergence of probability distributions on compact metric spaces apply, and the obedience property simplifies the construction of the rule governing sequential continuation equilibria consistent with the outcome of interest.

4.3.2 Payoff-Plausibility

Contracting equilibrium is often excessively permissive in the principal-agent game, so in our analysis of MCS environments, we will frequently apply the criterion of *payoff-plausibility* to refine the set of contracting equilibria.

Definition 3. *Suppose the environment is MCS. In the general-mechanism game, the profile of principal-type expected utilities* $(U^*(\theta_1), ..., U^*(\theta_N))$ *is* **plausible** *if*

$$\begin{aligned} U^*(\theta_n) \geq \max_{\chi \in \Delta(X \times T)} & \mathbb{E}_\chi[u(\theta_n, x, y^*(\theta_n, x)) - t] \\ s.t.\ \ & \textit{AIR: } \mathbb{E}_\chi[v(\theta_n, x, y^*(\theta_n, x)) + g(t)] \geq 0, \\ & \textit{PIC: } \mathbb{E}_\chi[u(\theta_{n'}, x, y^*(\theta_n, x)) - t] \leq U^*(\theta_{n'})\ \forall n' < n. \end{aligned} \tag{1}$$

An equilibrium or outcome is **payoff-plausible** *if the associated profile of principal-type expected utilities is plausible.*

Payoff-plausibility requires that each principal type θ get a payoff at least that from proposing any distribution χ that satisfies the agent IR and principal IC constraints when the agent responds under the belief that the type is θ. In particular, the agent IR constraint guarantees that the agent obtains a weakly positive expected utility from χ under type θ. The principal IC constraint says that every principal type smaller than θ must obtain a weakly lower payoff from proposing χ and having the agent respond under the belief that the type is θ than they obtain in equilibrium.

We adapt payoff-plausibility for the deterministic-mechanism game as follows.

Definition 4. *Suppose the environment is MCS. In the deterministic-mechanism game, the profile of expected utilities* $(U(\theta_1), ..., U(\theta_N))$ *is* **plausible** *if*

$$\begin{aligned} U(\theta_n) \geq & \max_{(x,t)\in X\times T} u(\theta_n, x, y^*(\theta_n, x)) - t \\ & s.t.\ v(\theta_n, x, y^*(\theta_n, x)) + g(t) \geq 0, \\ & \quad u(\theta_{n'}, x, y^*(\theta_n, x)) - t \leq U(\theta_{n'})\ \forall n' < n. \end{aligned} \tag{2}$$

The difference between this and Definition 3 is that here the domain of optimization is the set of (x, t) pairs, rather than the full set of distributions over them.

Section 8 discusses the relationship of payoff-plausibility to various adaptations of signaling game refinements to the principal-agent game. In particular, payoff-plausibility characterizes both the set of robust neologism proof (Clark [2021]) contracting equilibria and the set of strongly justified communication equilibria (Clark and Fudenberg [2021]) in MCS environments.

5 Properties of Contracting Equilibrium Outcomes

Here we focus on the possible outcomes that can emerge in contracting equilibria. An *outcome* $p \in \Delta(\Theta \times \mathcal{M} \times [0,1] \times X \times Y)$ is a probability distribution over tuples $(\theta, \mu, M_P, \alpha, x, y)$, where $(\theta, \mu, M_P, \alpha, x, y) \in \Theta \times \mathcal{M} \times [0,1] \times X \times Y$ represents the principal's type being θ, mechanism (μ, M_P) being proposed and accepted with probability α, and the action pair (x, y) occurring subsequent to acceptance. Because of the Inscrutability Principle, in the general-mechanism game it will prove convenient to identify an outcome $p \in \Delta(\Theta \times \mathcal{M} \times [0,1] \times X \times Y)$ with the corresponding distribution $p' \in \Delta(\Theta \times X \times Y)$ that is obtained from identifying each tuple $(\theta, \mu, M_P, \alpha, x, y)$ with the binary distribution $\alpha\delta_{(\theta,x,y)} + (1-\alpha)\delta_{(\theta,x_0,y')}$ for some $y' \in Y$ (the value of which is irrelevant).

We first establish some necessary conditions that must hold in all contracting equi-

librium outcomes. We then define a class of *safe* outcomes satisfying stronger versions of these necessary conditions, and show that, in MCS environments, the *principal-optimal* safe outcomes are always contracting equilibria. This then enables us to give a partial characterization of the contracting equilibrium set in MCS environments.

5.1 Necessary Conditions

Here we give some conditions that outcomes must satisfy if they occur in equilibrium in the general-mechanism game. For an arbitrary outcome p, we let $U(\theta, p)$ denote the expected utility of type θ and $\hat{U}(\theta, p, \theta')$ denote the expected utility type θ would obtain by mimicking type θ'.

Definition 5. *In the general-mechanism game, outcome p satisfies* **principal incentive compatibility** *if $U(\theta, p) \geq \hat{U}(\theta, p, \theta')$ for all $\theta, \theta' \in \Theta$. Moreover, p satisfies* **principal individual rationality** *if $U(\theta, p) \geq 0$ for all $\theta \in \Theta$.*

In the general-mechanism game, principal incentive compatibility requires that every principal type weakly prefers the conditional outcome given their type to the conditional outcome given any other type, while individual rationality says that every principal type weakly prefers their conditional outcome to their outside option. Since every principal type can always mimic any other type or simply take their outside option, both incentive compatibility and individual rationality are necessary conditions for principal optimization in equilibrium.

Even stronger incentive compatibility conditions must hold for equilibria in the deterministic-mechanism game, because of the inability of the principal to commit to non-degenerate distributions over actions. In particular, each principal type must weakly prefer their conditional outcome to the conditional outcome given any type, mechanism, and principal action triple in the outcome's support. The general statement of this condition is somewhat messy; however, it is considerably simpler for the case of an *always-accepting* outcome p where $\mathbb{P}[\alpha = 1] = 1$, that is, there is probability 1 that the mechanism proposed is accepted. We can identify such an outcome with

the $p \in \Delta(\Theta \times X \times Y)$ formed by mapping each tuple $(\theta, \mu, M_P, 1, x, y)$ to (θ, x, y). For $p \in \Delta(\Theta \times X \times Y)$, let $\hat{U}(\theta, p, \theta', x') = \mathbb{E}[U(\theta, x', y)|(\theta', x')]$ be the expected utility of type θ from the conditional distribution given (θ', x'). We then have the following requirement.

Definition 6. *In the deterministic-mechanism game, an always-accepting outcome satisfies* **principal incentive compatibility (PIC)** *if* $\mathbb{P}[U(\theta, p) \geq \hat{U}(\theta, p, \theta', x')] = 1$ *for all* $\theta \in \Theta$. *Moreover,* $p \in \Delta(\Theta \times X \times Y)$ *satisfies* **principal individual rationality (PIR)** *if* $U(\theta, p) \geq 0$ *for all* $\theta \in \Theta$.

5.2 Safe Outcomes

Safe outcomes satisfy both the principal incentive compatibility and individual rationality conditions, and they additionally require that a relationship be formed with probability 1 and that the prescribed play of the agent is optimal regardless of the probability distribution over the principal's type. For an arbitrary outcome p, let $V(\theta, p)$ denote the expected utility of type the agent conditional on type θ.

Definition 7. *In the general-mechanism game, outcome* $p \in \Delta(\Theta \times X \times Y)$ *is* **safe** *if it satisfies both the principal's incentive compatibility and individual rationality constraints, and it further satisfies:*

1. *Agent-safe IC (ASIC):* $\mathbb{P}[y \in \arg\max_{y' \in Y} V(\theta, x, y')] = 1$.
2. *Agent-safe IR (ASIR):* $V(\theta, p) \geq 0$ *for all* $\theta \in \Theta$.

The same holds in the deterministic-mechanism game for always-accepting outcomes.

ASIC ensures that, whenever the agent takes a given action y, it is a best response to the principal's action and every principal type that has positive probability when the agent is supposed to play y. ASIR says that the agent gets a weakly higher payoff conditional on each principal type than from their outside option.

5.3 Principal-Optimal Safe Outcomes and Contracting Equilibria in MCS Environments

Definition 8. *Safe outcome $p \in \Delta(\Theta \times X \times Y)$ is a* **principal-optimal** *safe outcome if it gives every type of the principal a weakly higher payoff than every other safe outcome p': $U(\theta, p) \geq U(\theta, p')$ for all $\theta \in \Theta$ and safe $p' \in \Delta(\Theta \times X \times Y)$.*[19]

Proposition 1. *In both the general-mechanism and deterministic-mechanism game, principal-optimal safe outcomes exist.*

There is always at least one safe outcome, the degenerate outcome in which x_o occurs with probability 1, i.e. players obtain their outside option payoffs. The proof of Proposition 1, given in Section OA.2, further shows that the set of safe outcomes is sequentially compact. This guarantees that, for every principal type, there is a safe outcome giving the type a higher payoff than any other safe outcome. Moreover, the outcome which is constructed by assigning to each type the same conditional outcome given their type as their most preferred safe outcome is itself safe, since no principal type would want to mimic any other. Consequently, this outcome is a principal-optimal safe outcome.

Since, in MCS environments, the principal prefers higher agent actions and the agent's optimal action increases with the principal's type, higher principal types would like to separate from lower principal types. Supermodularity between the principal's type and the agent's action allows the higher principal types to credibly do so by paying higher transfers to the agent. The following proposition characterizes the general-mechanism game principal-optimal safe outcomes in MCS environments by determining the corresponding conditional outcome distributions for each type. (For outcome $q \in \Delta(\Theta \times X \times Y)$, we let $q(\theta) \in \Delta(X \times Y)$ denote the conditional outcome distribution given type θ, and let $y^*(\theta, x) \equiv \arg\max_{y \in Y} v(\theta, x, y)$ denote the agent's best response to x when the principal is type θ.)

[19] In environments without moral hazard, the principal-optimal safe outcomes are frequently referred to as the "Rothschild-Stiglitz-Wilson (RSW) allocations."

Proposition 2. *In MCS environments, the conditional distributions of the principal-optimal safe outcomes* $\{q^*(\theta)\}_{\theta\in\Theta}$ *in the general-mechanism game are characterized inductively by*

$$\begin{aligned} q^*(\theta_n) \in \underset{q\in\Delta(X\times T\times Y)}{\arg\max}\ & \mathbb{E}_q[u(\theta_n,x,y)-t] \\ s.t.\ \ & \textit{AIC: } \mathbb{P}_q[y=y^*(\theta_n,x)|x\neq x_o]=1, \\ & \textit{AIR: } \mathbb{E}_q[v(\theta_n,x,y)+g(t)]\geq 0, \\ & \textit{PIC: } \mathbb{E}_q[u(\theta_{n'},x,y)-t]\leq \mathbb{E}_{q^*(\theta_{n'})}[u(\theta_{n'},x,y)-t]\ \forall n'<n, \end{aligned}$$

for all $n\in\{1,...,N\}$. *Moreover, the same inductive characterization holds for the deterministic-mechanism game when the PIC constraint is strengthened to* $\mathbb{P}_q[u(\theta_{n'},x,y)-t\leq U(\theta'_n,q^*(\theta_{n'}))]=1$ *for all* $n'<n$.[20]

The first and second constraints are simply the agent's incentive compatibility and individual rationality conditions for a safe outcome in MCS environments. The third constraint is a principal incentive compatibility condition guaranteeing that lower types than θ_n weakly prefer their outcome to mimicking θ_n. The proof of Proposition 2 follows standard lines and is given in Section OA.3. The strengthened principal incentive compatibility constraint for the deterministic-mechanism game ensures that no lower type would every want to deviate to a (x,t,y) in the support of the type θ_n distribution.

Theorem 2. *In both the general-mechanism and deterministic-mechanism games, any principal-optimal safe outcome is a contracting equilibrium outcome in MCS environments.*

The proof, given in Appendix C.1 and Section OA.9 of the Online Appendix, constructs sequences of modified principal-agent games and shows that the payoffs to the principal types in any limit of equilibrium outcomes in these games satisfy two conditions. The first is that each principal type's payoff is below their principal-optimal

[20] While the principal-optimal safe outcomes can differ between the two games because of the relaxed constraints for safety in the general-mechanism game, they are the same in all of our examples.

safe payoff. The second is that, for each mechanism, there is a sequential continuation equilibrium after the mechanism is proposed in the true principal-agent game that gives each principal type a lower payoff than they obtain from the limit of the equilibrium outcomes. We then show that these two conditions together ensure that the principal-optimal safe outcome is a contracting equilibrium outcome.

Here we describe the modifications made for the general-mechanism game. (The modifications, and overall argument, for the deterministic-mechanism game are similar.) One modification is that for each $\theta \in \Theta$ and $\chi \in \Delta(X \times T)$, the mechanism $(\mu_{\chi,\theta}, \{0\})$ in which the principal commits to the distribution that draws (x, t) according to χ and always gives θ as the recommendation to the agent is assumed to induce the outcome in which the agent accepts the proposal and then plays $y^*(\theta, x)$ after observing any $x \neq x_o$. (For the deterministic-mechanism game, this modification only applies to mechanisms of the form $(\delta_{((x,t),\theta)}, \{0\})$, which commit to some pure action-transfer pair (x, t).) Effectively, for such mechanisms, the agent's individual rationality constraint is discarded and the agent chooses their action under a belief that the type is certain to be the same as given in the recommendation. Aside from this modification, the mechanisms and play of the principal and agent are as in the true principal-agent game. (Thus, the possible sequential continuation equilibria after any mechanism outside the modified class are the same as in the real game.) The point of the modification is to prevent pooling between the principal types in equilibrium. It accomplishes this since, if there were pooling, the highest type involved in pooling would be strictly better off playing the mechanism that (a) commits to the same distribution over action-transfer pairs as they are realizing in equilibrium and (b) always gives their type as the recommendation.[21]

On its own, this modification could allow equilibrium outcomes that do not satisfy incentive compatibility or agent individual rationality in the unmodified game. We

[21]It is possible that the highest type involved in pooling could be indifferent between their equilibrium outcome and playing this mechanism. We prevent this by giving a small additional benefit to each type θ from proposing mechanisms of the form $(\mu_{\chi,\theta}, \{0\})$.

avoid these problems by modifying the utility functions of the principal types with costs to using mechanisms of the form $(\mu_{\chi,\theta}, \{0\})$. In particular, we make it prohibitively costly for a type θ principal to propose a $(\mu_{\chi,\theta}, \{0\})$ mechanism whenever (1) there is some other principal type who would get a higher payoff from proposing $(\mu_{\chi,\theta}, \{0\})$ (if the agent were to accept and respond according to $y^*(\theta, x)$) than they do from the prevailing outcome, or (2) the agent's total expected utility in the prevailing outcome conditional on θ is too low. We are careful to ensure that all the modifications together still ensure that no pooling can occur in equilibrium. These modifications also ensure that, conditional on any principal type, every equilibrium outcome satisfies the agent's individual rationality constraint.

Thus, every principal type gets a lower payoff in any equilibrium than in the principal-optimal safe outcomes. Moreover, there are sequential continuation equilibria in the true principal-agent game following the proposal of any mechanism that give the principal types even lower payoffs because the principal optimizes in equilibrium, and every mechanism has the same sequential continuation equilibria as some mechanism for which there are no modifications.

Before proceeding, we develop a sufficient condition for contracting equilibria in MCS environments in the general-mechanism game. For an arbitrary outcome p, let $V(p)$ denote the expected utility of the agent.

Definition 9. *In the general-mechanism game, an outcome $p \in \Delta(\Theta \times X \times Y)$ is* **incentive compatible** *if it satisfies the principal incentive compatibility and individual rationality constraints, and it further satisfies:*

1. *Agent IC (AIC):* $\mathbb{P}[y \in arg\max_{y' \in Y} \mathbb{E}[V(\theta, x, y')|(x, y)]] = 1$.
2. *Agent IR (AIR):* $V(p) \geq 0$.

These agent incentive compatibility and individual rationality constraints weaken those in the definition of safe outcomes. AIC ensures that the agent is only asked to play actions y that are best responses to the conditional distribution of the principal's type given the principal's action and the fact that the agent is supposed to play y.

AIR says that the agent gets a weakly higher payoff from the outcome than from their outside option. Both of these are necessary for an outcome to be consistent with the agent playing optimally, so incentive compatibility is required of any contracting equilibrium outcome.

Say that an outcome p ***principal-payoff-dominates*** the principal-optimal safe outcome p' if every principal type obtains a weakly higher payoff from p than p'.[22] When this is the case, incentive compatibility is not only necessary for p to be a contracting equilibrium outcome but also sufficient.

Proposition 3. *In the general-mechanism game in an MCS environment, an outcome that principal-payoff-dominates the principal-optimal safe outcome is a contracting equilibrium outcome if and only if it is incentive compatible.*

Intuitively, if the principal types would be (weakly) deterred from proposing a given mechanism when they are receiving their principal-optimal safe payoff, then they would also be deterred should they receive a higher payoff. Moreover, due to the Inscrutability Principle, it is easy to construct a single mechanism and subsequent optimal play that results in any given incentive compatible outcome. These two facts enable us to construct the desired contracting equilibrium. The formal proof is given in Appendix C.2.

The boundary of the set of incentive compatible payoffs that principal-payoff-dominate the principal-optimal safe outcome can be found using familiar design techniques from (uninformed) principal-agent problems. Since, with general mechanisms, the set of incentive compatible payoffs is convex, the full set of contracting equilibrium payoffs that principal-payoff-dominate the principal-optimal safe payoffs then emerges as the convex hull of this boundary.

[22] Formally, $p \in \Delta(\Theta \times X \times Y)$ principal-payoff-dominates $p' \in \Delta(\Theta \times X \times Y)$ if $\mathbb{E}_p[U(\theta, x, y)|\theta] \geq \mathbb{E}_{p'}[U(\theta, x, y)|\theta]$ for all $\theta \in \Theta$.

6 Payoff-Plausibility in MCS Environments

We now apply payoff-plausibility to refine the set of contracting equilibria in general MCS environments. We show that the principal-optimal safe outcome provides a payoff benchmark that every payoff-plausible contracting equilibrium must meet, and we show that payoff-plausibility often eliminates the ex-ante mechanism design benchmark. We also discuss how, with flexible contracts, payoff-plausibility can permit outcomes with higher principal payoffs than the principal-optimal safe outcome, while payoff-plausibility typically selects the least-cost separating equilibria when only flexible contracts can be proposed. However, in a special subclass of MCS environments, payoff-plausibility selects precisely the principal-optimal safe outcome when only deterministic flexible contracts can be proposed.

6.1 The Principal-Optimal Safe Benchmark

Section 5.3 showed that, in MCS environments, principal-optimal safe outcomes are always contracting equilibrium outcomes. They are additionally always payoff-plausible, and they provide payoff benchmarks that every payoff-plausible equilibrium must meet.

Theorem 3. *Suppose the environment is MCS. In both the general-mechanism and deterministic-mechanism games:*

1. *Every payoff-plausible equilibrium principal-payoff-dominates the principal-optimal safe outcomes.*
2. *The principal-optimal safe outcomes are payoff-plausible.*

Theorem 3 follows from combining the characterizations of the principal-optimal safe outcomes in Proposition 2 and the requirements of payoff-plausibility. In particular, the proof of Theorem 3.1 shows that, for any equilibrium that does not principal-payoff-dominate the principal-optimal safe outcome, there must be a lowest principal type θ whose expected utility violates payoff plausibility. Theorem 3.2 is an immediate consequence of the observation that, in the principal-optimal safe outcomes, each principal

type's payoff precisely equals their plausibility threshold.

6.2 Ruling Out the Ex-Ante Mechanism Design Benchmark

Here we generalize the finding from the firm and employee example that payoff-plausibility eliminates the ex-ante mechanism design benchmark. In general environments, the ***ex-ante mechanism design benchmarks*** are the outcomes that maximize the principal's ex-ante expected utility subject to incentive compatibility. Recall that in the firm-employee example, the ex-ante mechanism design benchmark involves the same profit shares and employee efforts as in the complete-information benchmark, but has different transfers.[23] In many other environments, the ex-ante mechanism design benchmarks also use the same actions but different expected transfers than the complete-information benchmark. Moreover, when this is the case and each type gains more from slight perturbations of their complete-information benchmark action than any lower type, payoff-plausibility rules out the ex-ante mechanism design benchmark.

Proposition 4. *For each $\theta \in \Theta$, let $x_\theta^{CI} \in X$ be the principal action in the complete-information benchmark when the principal's type is known to be θ. Suppose the environment is MCS and that the ex-ante mechanism design benchmarks have the same actions as the complete-information benchmark but different expected transfers for at least one principal type. If, for each $\theta \in \Theta$, there is a sequence $\{x_i\}$ converging to x_θ^{CI} such that $u(\theta, x_i, y^*(\theta, x_i)) - u(\theta, x_\theta^{CI}, y^*(\theta, x_\theta^{CI})) > u(\theta', x_i, y^*(\theta, x_i)) - u(\theta', x_\theta^{CI}, y^*(\theta, x_\theta^{CI}))$ for all $\theta' < \theta$ and i, then the ex-ante mechanism design benchmarks are not payoff-plausible.*

Since the agent's expected utility in an ex-ante mechanism design benchmark must exactly equal 0, and the transfers are different than in the complete-information benchmark, which also gives the agent expected utility 0, some type $\overline{\theta}$ must give the agent a strictly positive expected utility. By the inequality on payoff differences, we can find

[23] As with the firm-employee example, in general environments, the ***complete-information benchmark*** are the outcomes that could occur if the agent were to always learn the principal's type before contracting.

a slight perturbation of the action-transfer pair played by $\overline{\theta}$ such that, when the agent responds with the belief that the type is $\overline{\theta}$, (1) the type $\overline{\theta}$ principal would be strictly better off than in the ex-ante mechanism design benchmark, (2) all lower types would be strictly worse off, and (3) the agent would attain a positive utility. These three conditions mean that the payoff of the type $\overline{\theta}$ does not meet their plausibility threshold. The formal proof is given in Section OA.4 of the Online Appendix.

Not only are the conditions of Proposition 4 satisfied in the firm-employee example, but they are satisfied more generally in environments that feature profit-sharing between the principal to the agent. In particular, suppose the principal's action space is of the form $X \times [0,1] \times T$. Suppose further that the utilities are defined by $u(\theta, x, s, y) = (1-s)\pi(\theta, x, y) - \kappa(x)$ and $v(\theta, x, s, y) = s\pi(\theta, x, y) - c(y)$, where (1) $\pi(\theta, x, y) - \kappa(x) - c(y)$ is strictly concave in (x, y), (2) $\pi(\theta, x, y)$ is strictly increasing in θ for all $x \in X$ and $y > \underline{y}$, and (3) $\pi(\theta, x, y)$ and $c(y)$ are differentiable in y and satisfy $\frac{\partial \pi}{\partial y}(\theta, x, \underline{y}) > c'(\underline{y})$ as well as $\frac{\partial \pi}{\partial y}(\theta, x, y) > 0$ for all $y \in Y$. Then the ex-ante mechanism design benchmark will maximize the surplus conditional on each type, which requires setting a profit-share of $s = 1$. Principal incentive compatibility then requires that all the principal types obtain the same payoff, which necessarily involves different transfers than in the complete-information benchmarks, since these would give higher principal types strictly higher payoffs. Moreover, the payoff difference inequalities are satisfied since $\pi(\theta, x, y)$ is increasing in the principal's type, and payoff-plausibility intuitively precludes the ex-ante mechanism design benchmark since some type would be better off slightly decreasing the profit share below 1 while lower types would be worse off mimicking them.

However, environments where one of the principal's actions is the choice of a profit-share level are by no means the only ones in which Proposition 4 applies. The following is a modified example of the firm and employee in which the profit share is fixed. The firm instead has a costly investment action that affects the profitability of the task along with the firm's type and the employee's effort. (This example belongs to a class of *quasi-strict* and *doubly supermodular* environments, that always satisfy the difference

in payoffs inequality in the statement of Proposition 4. Quasi-strict is defined in Section 6.3, while doubly supermodular is defined in Section 6.4.)

Example with Firm Investment and Restricted Profit Sharing. As before, the firm has private information $\theta \in \{2, 4\}$ about the profitability or quality of a task for which they seek to hire an employee, a hired employee will choose an effort level $e \in \mathbb{R}_+$ that affects the probability of the task being successful, and the firm will pay a transfer t to the agent. However, unlike before, the profit share is fixed at $s = 1/2$ and the firm makes a costly investment $i \in \mathbb{R}_+$ that increases the productivity of the employee's effort. The utility functions of the firm and employee are $U(\theta, i, s, t, e) = \theta \ln(1 + i)e/2 - i^2/2 - t$ and $V(\theta, i, s, t, e) = \theta \ln(1 + i)e/2 - e^2/2 + t$, respectively. □

6.3 Flexible Versus Explicit Contracts

We show here that the implications of payoff-plausibility are very different with flexible contracts than with explicit ones: With flexible contracts, payoff-plausibility does not typically require separation between principal types, while when only explicit contracts can be proposed, payoff-plausibility selects the least-cost separating outcomes under broad conditions.

For an example where payoff-plausibility allows pooling under flexible contracts, consider again the firm and employee of Section 3, except now suppose that $\Theta = \{1, 2, 4\}$ and $\lambda(1) = \lambda(2) = \lambda(4) = 1/3$. Here there is an additional low type $\theta = 1$, and all three types are equally likely. One payoff-plausible pooling outcome is for the low type and medium type to pool and give all profit residuals to the employee ($s(1) = s(2) = 1$) along with the same base transfer of $t(1) = t(2) = -2.05$. The corresponding level of effort exerted by the employee is $e = 3/2$. The high type separates by giving half of the profit to the employee ($s(4) = 1/2$) along with a base transfer of $t(4) = -.05$; the corresponding level of effort exerted by the employee is $e = 2$. This outcome, which gives each principal type a strictly higher payoff than the principal-optimal safe outcome, is payoff-plausible because both the low and medium

types get at least their first-best payoff, while the high type's payoff precisely equals their plausibility threshold.

In contrast, if only explicit contracts can be proposed, payoff-plausibility selects the least-cost separating outcome.[24] Moreover, this selection holds in a broad set of MCS environments.

Definition 10. *An MCS environment is* **quasi-strict at** $x \in X$ *if*

1. *Strict monotonicity:* $u(\theta, x, y^*(\tilde{\lambda}, x))$ *and* $v(\theta, x, y^*(\tilde{\lambda}, x))$ *are strictly increasing in* θ *for all* $\tilde{\lambda} \in \Delta(\Theta)$.
2. *Strict supermodularity:*
   (a) $y^*(\tilde{\lambda}, x)$ *is strictly increasing in* $\tilde{\lambda}$ *according to the FOSD partial ordering of* $\Delta(\Theta)$.
   (b) *For all* $\theta, \theta' \in \Theta$ *and* $y, y' \in Y$ *such that* $\theta > \theta'$ *and* $y > y'$, $u(\theta, x, y) - u(\theta, x, y') > u(\theta', x, y) - u(\theta', x, y')$.

An MCS environment is **quasi-strict** *if it is quasi-strict at every* $x \neq x_o$.

Quasi-strictness strengthens some of the MCS conditions to hold strictly.

Definition 11. *An MCS environment has* **definite gains** *if the first-best payoff of type* θ_1 *under complete information is strictly positive.*[25]

Definite gains means that the lowest principal type is assured a strictly positive payoff in the complete information environment, which also ensures that each principal type must obtain a strictly positive payoff in any contracting equilibrium.

Proposition 5. *In quasi-strict MCS environments with definite gains, payoff-plausibility selects the least-cost separating outcomes when contracts must be explicit.*

Payoff-plausibility precludes pooling in quasi-strict MCS environments, because the highest type $\overline{\theta}$ would gain strictly more than the lower types from being recognized as $\overline{\theta}$,

[24] The criterion for payoff-plausibility when only explicit contracts can be proposed is the same criterion as in the deterministic-mechanisms game.

[25] Formally, $\underline{U}(\theta_1) > 0$ where $\underline{U}(\theta_1) = \max_{(x,t)} u(\theta_1, x, y^*(\theta_1, x)) - t$ s.t. $v(\theta_1, x, y^*(\theta_1, x)) + g(t) \geq 0$.

and the agent's expected utility conditional on the highest pooling type must be weakly positive.[26] Moreover, when only explicit contracts can be proposed, payoff-plausibility requires that every principal type obtain at least their least-cost separating payoff, just as they must obtain at least their principal-optimal safe payoff when flexible contracts can be proposed.[27]

6.4 Doubly Supermodular Environments

With deterministic flexible contracts, payoff-plausibility does select the principal-optimal safe outcomes in a class of MCS environments where there are complementarities between the principal's action and the principal's type and agent's action.[28] In these environments, $X = (X_1 \times X_2 \times ... \times X_K) \cup \{x_o\}$, and that $X_1 = [\underline{x}_1, \overline{x}_1] \subset \mathbb{R}$. To avoid boundary issues, we assume that $\max_{y \in Y} u(\theta, \overline{x}_1, y) + \overline{t} < 0$ for all $\theta \in \Theta$, which ensures that the highest value of x_1 is prohibitively costly.

Definition 12. *An environment with transfers is* **doubly supermodular** *if it is MCS and additionally satisfies:*

1. $y^*(\tilde{\lambda}, x_1, x_{-1})$ *is weakly increasing in* x_1 *for all* $\tilde{\lambda} \in \Delta(\Theta)$ *and* $x_{-1} \in X_{-1}$.
2. *For all* $\theta, \theta' \in \Theta$, $x_1, x_1' \in X_1$, $x_{-1} \in X_{-1}$, *and* $y \in Y$ *such that* $\theta > \theta'$ *and* $x_1 > x_1'$, $u(\theta, x_1, x_{-1}, y) - u(\theta, x_1', x_{-1}, y) \geq u(\theta', x_1, x_{-1}, y) - u(\theta', x_1', x_{-1}, y)$, *with the inequality holding strictly when* $u(\theta, x_1', x_{-1}, y) > u(\theta', x_1', x_{-1}, y)$.

The first condition says the agent's best response is weakly increasing in the x_1 component of the principal's action. The second condition requires that the difference in principal utility from a higher x_1, holding fixed the remaining components of the principal's action as well as the agent's action, is higher for a higher principal type,

[26] As seen in the earlier three-type firm and employee example, it can be that, with flexible contracts, the agent's expected utility conditional on each pooling type is strictly negative.

[27] Quasi-strict MCS environments do not contain the firm-employee example, because the strict supermodularity conditions fail at $s = 0$, and the strict monotonicity condition and second strict supermodularity condition fail at $s = 1$. Section OA.10 states and proves a more general version of Proposition 5 that does cover the example. Intuitively, neither the issues at $s = 0$ nor $s = 1$ prevent the conclusion of Proposition 5, because quasi-strictness holds at arbitrarily close values of s.

[28] The result we develop does not hold with general flexible contracts.

and strictly so at points where when the higher principal type gets a strictly higher utility than the lower type.

These requirements are satisfied in many economic applications, including the informed firm and employee example with firm investment from Section 6.2. (Additionally, while profit sharing was restricted in the original presentation of that example, the example would continue to be doubly supermodular if unrestricted profit sharing were allowed.) The conditions of Definition 12 can be readily verified when taking i to be the first component of the firm's action.

Proposition 6. *In an environment with definite gains that is doubly supermodular and quasi-strict, in the deterministic-mechanism game, the payoff-plausible contracting equilibrium outcomes are the principal-optimal safe outcomes.*

The proof, which is in Section OA.5 of the Online Appendix, shows that every payoff-plausible contracting equilibrium outcome is always-accepting. Intuitively, for any contracting equilibrium outcome that is not always-accepting, there is a mechanism that is accepted with some probability $\alpha \in (0, 1)$, and an (x, t) allowed by the mechanism such that some type $\overline{\theta}$ is willing to propose the mechanism and play (x, t) and the agent gets a conditionally positive expected utility when this occurs. Without loss, we can take $\overline{\theta}$ to be the highest such type. Then $\overline{\theta}$ could propose an action x' with a slightly increased first component relative to x, and adjust their transfer so that if the agent accepts and plays $y^*(\theta, x')$, the agent obtains a strictly higher payoff than 0, while θ is strictly better off, and every lower type is strictly worse off than in equilibrium. But this violates payoff-plausibility. The proof then uses a similar argument to show that no payoff-plausible contracting equilibrium has a principal type θ playing an action x that gives the agent an expected utility strictly above 0 when the agent plays $y^*(\theta, x)$. It follows that the expected utility of the agent conditional on any principal type must be weakly less than 0. Since the agent's unconditional expected utility must be no less than 0, the expected utility of the agent conditional on any principal type must exactly equal 0. These facts together imply that the lowest type

involved in pooling would have to give the agent a strictly negative expected utility, which we have established is not possible, so it follows that there can be no pooling of types. Thus, every payoff-plausible outcome must be safe. Since, by Theorem 3.1, every payoff-plausible outcome principal-payoff-dominates the principal-optimal safe outcome, it follows that every payoff-plausible outcome must be a principal-optimal safe outcome.

As previously noted, the firm and employee example is not quasi-strict, so Proposition 6 does not apply. In Section OA.5, we state and prove a stronger version of the proposition that covers the doubly supermodular firm and employee example.

7 Effects of Moral Hazard

We now discuss general differences between informed principal environments with and those without moral hazard. A key difference between the sets of contracting equilibria is that, without moral hazard, all contracting equilibria must principal-payoff-dominate the RSW allocations/principal-optimal safe outcomes, whereas there can be contracting equilibria that do not principal-payoff-dominate the principal-optimal safe outcomes when moral hazard is present. Another difference is that, when payoff-plausibility is applied, typically the RSW allocations are selected in the absence of moral hazard. In contrast, when moral hazard is present, outcomes that are more efficient than the principal-optimal safe outcomes and involve pooling can survive payoff-plausibility. Also, while there are always safe equilibria in environments without moral hazard, this is not the case in all environments with moral hazard outside of the MCS class.

7.1 Effects on the Set of Contracting Equilibria

Maskin and Tirole [1992] showed that, in the common values setting, contracting equilibrium outcomes must principal-payoff-dominate the RSW allocations whenever there is a sequence of *strictly* safe outcomes that converges to an RSW allocation, a condi-

tion that always holds in environments with transfers.[29] Additionally, in the private values setting studied by Maskin and Tirole [1990], every contracting equilibrium must give each principal type a weakly higher payoff than they could secure when their information is publicly known, which is the payoff-benchmark corresponding to the principal-optimal safe outcomes in that setting. However, with moral hazard, there can be contracting equilibria where some principal types get less than their principal-optimal safe payoff, as seen in the example in Section 3.3.

The reason why every contracting equilibrium principal-payoff-dominates the principal-optimal safe outcome when there is no moral hazard is that the principal can always propose a direct mechanism that induces a strictly safe outcome. Because the agent does not act, each principal type strictly prefers to report their type truthfully if the mechanism is accepted. Consequently, the agent must accept the proposal of any such mechanism, so each principal type can obtain a payoff no less than their principal-optimal safe payoff in any contracting equilibrium.

The Maskin and Tirole [1992] result does not extend to settings with agent moral hazard, because the agent's beliefs about the principal's type can influence their play should they accept a contract. In the context of the firm-worker example, it is possible that the employee believes that the firm type is low after the proposal of any off-path mechanism. If so, then when a strictly safe direct mechanism is proposed the employee responds as if the firm is the low type, which deters the high type firm from proposing it. If the firm could directly control the effort of the employee so that there were no moral hazard, then the employee beliefs would only be relevant for the decision of whether to accept a given contract, and every strictly safe direct mechanism would necessarily be accepted. However, in MCS environments, payoff-plausibility restores the qualitative prediction that equilibria principal-payoff-dominate

[29]Maskin and Tirole [1992] only analyzed general-mechanism proposal games. Formally, when general mechanisms can be proposed in a common values setting, outcome $p \in \Delta(\Theta \times X)$ is ***strictly*** safe if the principal incentive compatibility condition is strengthened to $U(\theta, p) > \mathbb{E}_p[U(\theta, x)|\theta']$ for all $\theta, \theta' \in \Theta$ and the agent individual rationality constraint is strengthened to $V(\theta, p) > 0$ for all $\theta \in \Theta$. A similar condition for strictly safe outcomes applies when only deterministic mechanism can be proposed.

the principal-optimal safe outcomes.

7.2 Effects on the Set of Payoff-Plausible Equilibria

We saw earlier in Section 6.3 that payoff-plausibility can allow outcomes that give the principal types strictly higher payoffs than the principal-optimal safe outcomes. In contrast, Maskin and Tirole [1992] showed that in a class of environments without moral hazard satisfying a "sorting condition," the Intuitive Criterion selects precisely the RSW allocations. (Payoff-plausibility makes precisely the same prediction in these environments.) Moreover, unlike the RSW allocations, some of the payoff-plausible outcomes with moral hazard do involve pooling, including the outcome in the example of Section 6.3.

7.3 Effects on the Existence of Safe Equilibria

Outside of MCS environments, safe equilibria do not necessarily exist when moral hazard is present. In contrast, safe equilibria always exist without moral hazard.

Proposition 7. *In both the general-mechanism and deterministic-mechanism games:*

1. *Without moral hazard, the principal-optimal safe outcomes are always contracting equilibrium outcomes.*
2. *With moral hazard, there may be no safe contracting equilibrium outcomes.*

Maskin and Tirole [1992] originally proved Proposition 7.1 under two additional assumptions: (1) that the principal-optimal safe outcome is "interim efficient" for some full-support probability distribution over the principal's type, and (2) that the principal and agent have access to a public randomization device, ensuring the convexity of the set of sequential continuation equilibria following any mechanism proposal.[30]

[30]Section OA.11 gives a proof of Proposition 7.1 that relaxes these assumptions and allows for mechanisms with agent message spaces, as in Maskin and Tirole [1992]. DeMarzo and Frankel [2020] proved an analog of the Maskin and Tirole [1992] result for a dynamic version of correlated equilibrium.

The following example demonstrates Proposition 7.2. There the example, the principal-optimal safe outcome is not a contracting equilibrium outcome. Since the principal-optimal safe outcome is a contracting equilibrium outcome whenever any safe outcome is, it follows that there are no safe contracting equilibrium outcomes in the example.[31]

Example 1. Suppose that $\Theta = \{\theta_1, \theta_2\}$, $X = \{x_1, x_2\}$, $Y = \{y_1, y_2\}$, and that $U(\theta, x, y)$ and $V(\theta, x, y)$ are as shown below. (The first number in each pair is the principal's payoff, while the second is the agent's.)

θ_1	y_1	y_2
x_1	$4, -1$	$-1, 2$
x_2	$1, 1$	$1, 1$

θ_2	y_1	y_2
x_1	$-1, 2$	$4, -1$
x_2	$1, 1$	$1, 1$

In the principal-optimal safe outcome, both principal types play x_2 and the agent accepts the corresponding mechanism, resulting in a payoff of 1 to all parties. However, there is no equilibrium with this outcome, since at least one type of the principal would be strictly better off by instead proposing a mechanism that commits to x_1. To see this, note that the agent's expected utility from accepting such a mechanism, regardless of the agent's belief over the principal's type, is no less than 1, so the agent would accept such a mechanism. Moreover, the sum of the expected utilities of the principal types equals 3 for all agent responses to x_1, so at least one of the principal types must obtain a strictly higher expected utility than 1. □

8 Foundations for Payoff-Plausible Equilibrium

Payoff-plausibility is not an ad hoc requirement imposed without justification; rather, it precisely captures the predictions made by two communication-based signaling refinements when they are adapted to MCS informed principal environments. Such

[31] However, in Section OA.12, we show that, in any environment, there are always contracting equilibria that principal-payoff-dominate the principal-optimal safe outcomes.

refinements are natural here since communication is prevalent in many principal-agent settings, as with a firm and employee, and can play an important role in determining the resulting outcomes.[32] Here we describe the two refinements, along with their motivations and foundations, and discuss why they are characterized by payoff-plausibility in MCS environments.

Robust Neologism Proofness: *Robust neologism proofness (RNP)* was developed in Clark [2021] for traditional signaling games. Loosely, RNP is a refinement of contracting equilibrium that that has a similar motivation to the "informal speech" motivation of the Intuitive Criterion (Cho and Kreps [1987]). However, it allows the sender (principal) to convince the agent they belong to a specific subset of types, rather than just convincing the agent that they are a type for whom a given deviation is not equilibrium dominated. We give the formal definition of RNP, as adapted for our informed principal setting, in Section OA.13.

Strongly Justified Communication Equilibrium: Clark and Fudenberg [2021] developed SJCE as a refinement of Nash equilibrium for signaling games with cheap-talk communication and gave it a learning foundation.[33] The learning foundation assumes the typical sender has much more playing experience than the typical receiver. Identifying senders with principals and receivers with agents, this assumption seems particularly fitting for many principal-agent settings in which the principals are institutions such as firms and agents are individuals or other small units. In Section OA.13 we explain how to adapt SJCE to our informed principal setting.

Loosely, the focus of both RNP and SJCE is on analyzing various sets of types that could gain by identifying themselves at an equilibrium p. In the general-mechanism game, consider a distribution over actions χ along with the mechanism $(\mu_\chi, \{0\})$ com-

[32] Myerson [1983] noted that communication is likely to be important in informed principal problems.

[33] SJCE is a refinement of Nash equilibrium rather than contracting equilibrium because Nash equilibrium is a necessary condition in the learning foundation of Clark and Fudenberg [2021], but sequential equilibrium type solution concepts are not.

mitting to χ. There is a "credible robust neologism" corresponding to $(\mu_\chi, \{0\})$ and the set of types $\widetilde{\Theta}$ if, when the agent responds to the proposal of $(\mu_\chi, \{0\})$ under a belief that the type is in $\widetilde{\Theta}$, proposing $(\mu_\chi, \{0\})$ would give every type in $\widetilde{\Theta}$ a strictly higher payoff than in p and every type outside $\widetilde{\Theta}$ a strictly lower payoff. A contracting equilibrium is RNP precisely when it has no credible robust neologisms. For SJCE, the set of "strongly justified types" for $(\mu_\chi, \{0\})$ is essentially the minimal set of types $\Theta^{SJ}(\chi, p)$ that, when a mixture over agent best responses to $(\mu_\chi, \{0\})$ and beliefs that the type is in $\Theta^{SJ}(\chi, p)$ would make a type weakly prefer proposing $(\mu_\chi, \{0\})$ over their equilibrium outcome, some type in $\Theta^{SJ}(\chi, p)$ would strictly prefer to propose $(\mu_\chi, \{0\})$. SJCE requires that there is some mixture over agent best responses to $(\mu_\chi, \{0\})$ and beliefs concentrating in $\Theta^{SJ}(\chi, p)$ which deters every principal type from proposing $(\mu_\chi, \{0\})$. Analogous constructions hold for the deterministic-mechanism game, except that the mechanisms of interest are those that commit to pure principal actions, rather than possibly non-degenerate distributions over actions.

In OA.14, we show that payoff-plausibility characterizes the predictions of RNP and SJCE in MCS environments. That is, the payoff-plausible outcomes are precisely those that survive RNP and also precisely those that survive SJCE. Here we briefly describe how this proceeds for RNP in the general-mechanism game; the procedures for the deterministic-mechanism game and for SJCE are similar. Intuitively, if an outcome is not payoff-plausible, there is some type θ whose payoff falls below their plausibility threshold. Then the distribution χ that attains this type's plausibility threshold either gives a credible robust neologism corresponding to θ, or there is some higher type θ' who would also prefer proposing χ to their equilibrium payoff when the agent responds under the belief that the type is θ. In the latter case, using the monotonicity and supermodularity conditions of MCS, we can modify χ by increasing the transfer levels so that θ' would strictly prefer to propose the resulting contract (assuming the agent responds under a belief that the type is θ') and all lower types strictly prefer their equilibrium payoffs. This either results in a credible robust neologism corresponding to θ', or there is some yet higher type θ'' that would prefer to propose this modified χ

if the agent responds under the belief the type is θ'. In the latter case, we carry out the same procedure, which must eventually terminate in a credible robust neologism.

Arguing that payoff-plausible outcomes are RNP proceeds as follows. If there is a credible robust neologism corresponding to χ and $\widetilde{\Theta}$, then the lowest type $\underline{\theta}$ in $\widetilde{\Theta}$ must get a strictly higher payoff than in equilibrium by proposing χ when the agent responds under a belief concentrating on $\underline{\theta}$. Moreover, this agent response would lead all types below $\underline{\theta}$ to get a strictly lower payoff by proposing χ. But this implies that the payoff of $\underline{\theta}$ does not meet their plausibility threshold.

9 Conclusion

We developed a general framework for studying informed principals in environments with agent moral hazard, and we established the existence of contracting equilibria. In MCS environments, the principal-optimal safe outcomes are always payoff-plausible contracting equilibrium outcomes and they provide a lower bound for the principal payoffs in any other payoff-plausible outcome. Furthermore, payoff-plausibility often rules out the ex-ante mechanism design benchmarks, which underscores that signaling issues arising from an informed principal's private information at the time of contracting should not be ignored. In contrast to informed principal environments without moral hazard, there can be contracting equilibria that give the principal types lower payoffs than the principal-optimal safe outcomes, and payoff-plausibility can allow more efficient outcomes with pooling.

We conclude with some extensions and possible directions for future research. In some settings, the agent may not be aware of the principal's action until after taking their own action. This can be captured by having the agent observe only the recommendation r before choosing y. Many of the results carry through to this setting; however, the upper frontier of payoffs typically shifts upward due to the principal's greater concealment ability. This is true in particular with the informed firm and employee. For example, the highest equilibrium payoff that the type 4 firm can attain

when the payoff of the type 2 firm is 2 shifts from 4 to 17/4.

Incorporating agent adverse selection alongside agent moral hazard seems desirable for some informed principal settings. Another avenue would be to consider principal information that is verifiable, potentially at some cost. This would lead to issues of informed information design as well as mechanism design, and would relate to a growing literature on information design by an informed designer. (See e.g. Perez-Richet [2014], Hedlund [2017], Chen and Zhang [2020], and Koessler and Skreta [2021].) An interesting possibility would be to consider situations where information can be verified only after a principal-agent relationship has been formed.

References

B. Balkenborg and M. Makris. An undominated mechanism for a class of informed principal problems with common values. *Journal of Economic Theory*, 157:918–958, 2015.

P. Beaudry. Why an informed principal may leave rents to an agent. *International Economic Review*, 35:821–832, 1994.

N. C. Bedard. Contracts in informed-principal problems with moral hazard. *Economic Theory Bulletin*, 5:21–34, 2017.

R. Bénabou and J. Tirole. Intrinsic and extrinsic motivation. *Review of Economic Studies*, 70:489–520, 2003.

M. Cella. Informed principal with correlation. *Games and Economic Behavior*, 64: 433–456, 2008.

H. Chade and R. Silvers. Informed principal, moral hazard, and the value of a more informative technology. *Economics Letters*, 74:291–300, 2002.

Y. Chen and J. Zhang. Signalling by bayesian persuasion and pricing strategy. *The Economic Journal*, 130:976–1007, 2020.

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

D. Clark. Robust neologism proofness. Working Paper, 2021.

D. Clark and D. Fudenberg. Justified communication equilibrium. *American Economic Review*, 111:3004–3034, 2021.

P. M. DeMarzo and D. M. Frankel. Mechanism design with an informed principal: Extensions and generalizations. Working Paper, 2020.

P. M. DeMarzo, D. M. Frankel, and Y. Jin. Portfolio liquidity and security design with private information. Working Paper, 2020.

J. Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5:514–531, 1993.

D. Fudenberg and D. Levine. Limit games and limit equilibria. *Journal of Economic Theory*, 38:261–279, 1986.

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

J. Hedlund. Bayesian persuasion by a privately informed sender. *Journal of Economic Theory*, 167:229–268, 2017.

W. Hildenbrand. *Core and Equilibria of a Large Economy*. Princeton University Press, 1974.

R. Inderst. Incentive schemes as a signaling device. *Journal of Economic Behavior & Organization*, 44:455–465, 2001.

R. Inderst. Matching markets with adverse selection. *Journal of Economic Theory*, 121:145–166, 2005.

F. Koessler and V. Skreta. Informed seller with taste heterogeneity. *Journal of Economic Theory*, 165:456–471, 2016.

F. Koessler and V. Skreta. Information design by an informed principal. Working Paper, 2021.

D. Martimort and W. Sand-Zantman. Signalling and the design of delegated management contracts for public utilities. *Rand Journal of Economics*, 37:763–782, 2006.

E. Maskin and J. Tirole. The principal-agent relationship with an informed principal, i: Private values. *Econometrica*, 58:379–409, 1990.

E. Maskin and J. Tirole. The principal-agent relationship with an informed principal, ii: Common values. *Econometrica*, 60:1–42, 1992.

T. Mekonnen. Informed principal, moral hazard, and limited liability. *Economic Theory Bulletin*, 9:119–142, 2021.

P. R. Milgrom and R. J. Weber. Distributional strategies for games with incomplete information. *Mathematics of Operations Research*, 10:619–632, 1985.

R. B. Myerson. Mechanism design by an informed principal. *Econometrica*, 51:1767–1797, 1983.

T. Mylovanov and T. Tröger. Informed-principal problems in environments with generalized private values. *Theoretical Economics*, 7:465–488, 2012.

T. Mylovanov and T. Tröger. Mechanism design by an informed principal: Private values with transferable utility. *Review of Economic Studies*, 81:1668–1707, 2014.

E. Perez-Richet. Interim bayesian persuasion: First steps. *American Economic Review*, 104:469–474, 2014.

I. Segal and M. Whinston. Robust predictions for bilateral contracting with externalities. *Econometrica*, 71:757–791, 2003.

S. Severinov. An efficient solution to the informed principal problem. *Journal of Economic Theory*, 141:114–133, 2008.

C. Wagner, T. Mylovanov, and T. Tröger. Informed-principal problem with moral hazard, risk-neutrality, and no limited liability. *Journal of Economic Theory*, 159: 280–289, 2015.

A Definition of Contracting Equilibrium

A.1 Sequential Continuation Equilibria

A first step is to define a sequential continuation equilibrium in the subgame in which an arbitrary mechanism is accepted. Given mechanism (μ, M_P), let $\Pi_P \equiv \Delta(M_P)$ be the space of probability distributions over the principal's message. An assessment is a tuple $(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \Lambda, \beta_A)$ of (1) a probability distribution over the principal type $\tilde{\lambda} \in \Delta(\Theta)$, (2) a message strategy profile $(\pi_{\theta_1}, ..., \pi_{\theta_N}) \in (\Pi_P)^{\Theta}$, and (3) a belief updating rule $\Lambda : (\cup_{m_P \in M_P} \text{supp}(\mu(m_P))) \to \Delta(\Theta)$ and an agent action rule $\beta_A : \cup_{m_P \in M_P} \text{supp}(\mu(m_P)) \to \Delta(Y)$, which are measurable mappings taking principal action and recommendation pairs that are possible under (μ, M_P) into $\Delta(\Theta)$ and $\Delta(Y)$, respectively.

We restrict attention to *consistent* assessments. Consistency requires that there be a sequence of full-support beliefs and profiles of full-support message strategies such that (1) the full-support beliefs converge to the belief in the assessment, (2) the profiles

of full-support message strategies converge to the profile of message strategies in the assessment, and (3) the agent's belief update rule in the assessment equals the limit of the agent's beliefs obtained applying Bayes' rule along the sequence of full-support beliefs and profiles of full-support message strategies.

Given mechanism (μ, M_P), let $\Pi_{P,+} \equiv \Delta_+(M_P)$ be the space of full-support probability distributions over the principal message. Given a belief over the principal type $\tilde{\lambda} \in \Delta(\Theta)$ and a profile of full-support principal message strategies $(\pi_{\theta_1}, ..., \pi_{\theta_N}) \in (\Pi_{P,+})^{\Theta}$, let $\widehat{\Lambda}(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N})[(x,r)] \in \Delta(\Theta)$ be the agent's posterior after observing $(x,r) \in \cup_{m_P \in M_P} \text{supp}(\mu(m_P))$.

Definition 13. *An assessment* $(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \Lambda, \beta_A)$ *after mechanism* (μ, M_P) *is accepted is* **consistent** *if there is a sequence* $\{(\tilde{\lambda}_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N})\}_j$ *such that*

$$\lim_{j\to\infty} \tilde{\lambda}_j = \tilde{\lambda}, \lim_{j\to\infty} \pi_{j,\theta} = \pi_\theta \text{ for all } \theta \in \Theta \text{ and}$$
$$\lim_{j\to\infty} \widehat{\Lambda}(\tilde{\lambda}_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N})(x,r) = \Lambda(x,r) \ \forall (x,r) \in \cup_{m_P \in M_P} supp(\mu(m_P)).$$

Definition 14. *A* **sequential continuation equilibrium after mechanism** (μ, M_P) **is accepted** *is a consistent assessment* $(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \Lambda, \beta_A)$ *such that*

1. *For every* $\theta \in \Theta$,

$$\mathbb{E}_{\pi_\theta}[\mathbb{E}_{\mu(m_P)}[\mathbb{E}_{\beta_A(x,r)}[U(\theta,x,y)]]]$$
$$= \max_{m_P \in M_P} \mathbb{E}_{\mu(m_P)}[\mathbb{E}_{\beta_A(x,r)}[U(\theta,x,y)]].$$

2. *For every* $(x,r) \in \cup_{m_P \in M_P} supp(\mu(m_P))$,

$$\beta_A(x,r) \in \Delta(\arg\max_{y \in Y} \mathbb{E}_{\Lambda(x,r)}[V(\theta,x,y)])$$
$$\text{for all } (x,r) \in \cup_{m_P \in M_P} supp(\mu(m_P)).$$

Condition 1 of Definition 14 means that, for every $\theta \in \Theta$, π_θ puts support only on

those messages which are optimal for the type θ principal given the play of the agent β_A. Condition 2 requires that, for every $(x, r) \in \cup_{m_P \in M_P} \text{supp}(\mu(m_P))$, $\beta_A(x, r)$ puts support only on those agent actions which are optimal for the agent given x and belief about the principal's type $\Lambda(x, r)$.

Suppose we fix a sequential continuation equilibrium after a given mechanism has been accepted, and analyze whether the agent should choose to accept or reject the principal's proposal given the prescribed continuation play. Combining this sequential continuation equilibrium with an optimal agent acceptance/rejection choice results in a sequential continuation equilibrium in the earlier subgame where a mechanism has been proposed but not yet accepted or rejected.

Definition 15. *A* **sequential continuation equilibrium after mechanism** (μ, M_P) **is proposed** *consists of a sequential continuation equilibrium after the mechanism is accepted as well as a (possibly randomized) agent acceptance decision that is optimal given the prescribed future play and posterior belief about the principal type.*

A.2 Definition of Contracting Equilibrium in the General-Mechanism Game

We now define contracting equilibrium for the whole principal-agent game. By the Inscrutability Principle (Myerson [1983]), it is without loss of generality to restrict attention to equilibria in which, on the path of play, all principal types propose the same *direct* and *incentive compatible* mechanism. (Of course, there is no restriction on the mechanisms that the principal can propose off the path of play.) A ***direct mechanism*** (μ, Θ) is a special kind of mechanism in which the principal's message space coincides with their type space, i.e. $M_P = \Theta$. Additionally, every possible principal recommendation must contain an explicit action recommendation to the agent. A direct mechanism is ***incentive compatible*** if the induced "truthful and obedient" *outcome* is an incentive compatible outcome in the sense of Definition 9. The truthful and obedient outcome of a direct mechanism is the outcome that results when (1)

the agent accepts the mechanism proposal, (2) each principal type plays the message corresponding to their type, and (3) the agent follows every on-path action recommendation.

Definition 16. *A* **contracting equilibrium** *in the general-mechanism game consists of (1) an incentive compatible direct mechanism, which is proposed by every principal type, and (2) a sequential continuation equilibrium for when any mechanism is proposed that results in a weakly lower expected utility to each principal type than they obtain from truthful and obedient outcome from the mechanism they are supposed to propose.*

A.3 Definition of Contracting Equilibrium in the Deterministic-Mechanism Game

As with the general-mechanism game, we consider sequential continuation equilibria in both the subgames in which a mechanism is proposed and the subgames in which a a mechanism is accepted. The original general-mechanism sequential equilibrium definitions carry over to the restricted setting. When analyzing contracting equilibria in the deterministic-mechanism game, we shall use τ to denote a rule mapping mechanisms into sequential continuation equilibria, so that

$$\tau(\mu, M_P) = (\tilde{\lambda}(\mu, M_P), \pi_{\theta_1}(\mu, M_P), ..., \pi_{\theta_N}(\mu, M_P), \alpha(\mu, M_P), \Lambda(\mu, M_P), \beta_A(\mu, M_P))$$

is the sequential continuation equilibrium following the proposal of (μ, M_P). Additionally, we will let $U(\theta, \tau(\mu, M_P))$ denote the expected payoff to the type θ principal from proposing mechanism (μ, M_P) when subsequent play is governed by $\tau(\mu, M_P)$.

Unlike with the general-mechanism game, we cannot make use of the Inscrutability Principle to justify a restriction to equilibria where all principal types propose the same direct and incentive compatible mechanism. Instead, here we must explicitly consider the possibility that the type θ principal proposes a non-degenerate distribution over

mechanisms $\eta_\theta \in \Delta(\mathcal{M})$. The set of such probability distributions that we consider are the Borel distributions corresponding to the topology induced by the following metric over deterministic mechanisms. To define this metric, denote the metrics over X and R by $d_X : X^2 \to \mathbb{R}_+$ and $d_R : R^2 \to \mathbb{R}_+$, respectively. Let $d_{X\times R} : (X \times R)^2 \to \mathbb{R}_+$ be the metric over $X \times R$ given by $d_{X\times R}((x, r), (x', r')) = \max\{d_X(x, x'), d_R(r, r')\}$, and fix an $A > \max_{x,x',r,r'} d_{X\times R}((x, r), (x', r'))$.

Definition 17. *The* **deterministic mechanism metric** $d_{\mathcal{M}} : \mathcal{M}^2 \to \mathbb{R}_+$ *is given by the following:*

$$d_{\mathcal{M}}((\mu, M_P), (\mu', M'_P)) = \begin{cases} \max_{m_P \in M_P} d_{X\times R}(supp(\mu(m_P)), supp(\mu'(m_P))) & \text{if } M_P = M'_P \\ A & \text{if } M_P \neq M'_P \end{cases}.$$

We will require that the mechanism proposal distributions chosen by the principal types be optimal given the prevailing rule mapping mechanisms into sequential continuation equilibrium play.

Definition 18. *The profile of mechanism proposal distributions* $\{\eta_\theta\}_{\theta\in\Theta}$ *is* **optimal** *given sequential equilibrium rule* τ *if* $\eta_\theta(\arg\max_{(\mu,M_P)} U(\theta, \tau(\mu, M_P))) = 1$ *for all* $\theta \in \Theta$.[34]

Finally, we will require that the agent's intermediate belief about the principal's type at the mechanism proposal stage $\tilde{\lambda} : \mathcal{M} \to \Delta(\Theta)$ comes from a regular conditional distribution derived from their prior λ and the profile of mechanism proposal distributions $\{\eta_\theta\}_{\theta\in\Theta}$ used by the principal types.

Definition 19. *A* **contracting equilibrium** *in the deterministic-mechanism game consists of a profile of mechanism proposal distributions* $\{\eta_\theta\}_{\theta\in\Theta}$ *and rule governing sequential continuation equilibria* τ *such that (1)* $\{\eta_\theta\}_{\theta\in\Theta}$ *is optimal given* τ*, and (2)*

[34]Note that this implicitly assumes that $\arg\max_{(\mu,M_P)} U(\theta, \tau(\mu, M_P))$ is measurable. This will be established in the proof of Theorem 1.

The component of τ giving the agent's intermediate belief about the principal's type at the mechanism proposal stage $\tilde{\lambda} : \mathcal{M} \to \Delta(\Theta)$ comes from a regular conditional distribution derived from their prior λ and the profile of mechanism proposal distributions $\{\eta_\theta\}_{\theta \in \Theta}$ used by the principal types.

B Existence of Contracting Equilibria

B.1 Proof of Theorem 1 for the General-Mechanism Principal-Agent Game

To prove Theorem 1, we construct a sequence of finite approximations of the action and recommendation spaces, and show that the limits of the associated contracting equilibrium outcomes are contracting equilibrium outcomes in the limit environment. We do so by proving a general upper hemicontinuity result concerning the correspondence mapping the primitives of the principal-agent game to its contracting equilibrium outcomes.

Before developing our upper hemicontinuity result, we first define convergence of a sequence of primitives.[35] Throughout, we hold the type space Θ fixed, and we assume that there is some larger metric space that embeds all the principal action spaces and agent action spaces. We use $\mathcal{P}$ to denote a collection of primitives, consisting of a prior λ, principal action space X, agent action space Y, principal payoff function $U : \Theta \times X \times Y \to \mathbb{R}$, and agent payoff function $V : \Theta \times X \times Y \to \mathbb{R}$.

Definition 20. *A sequence of primitives $\{\mathcal{P}_j\}_{j \in \mathbb{N}}$* **converges** *to $\mathcal{P}$ if*

1. $\lim_{j \to} \mu_j = \mu$,
2. $\lim_{j \to \infty} X_j = X$ *and* $\lim_{j \to \infty} Y_j = Y$ *according to the Hausdorff metric, and*
3. *For all $\varepsilon > 0$, there exists J and $\delta > 0$ such that $|U_j(\theta, x', y') - U(\theta, x, y)| < \varepsilon$ and $|V_j(\theta, x', y') - V(\theta, x, y)| < \varepsilon$ if $|x' - x| < \delta$, $|y' - y| < \delta$, and $j > J$.*

[35] Our convergence notion is related to similar notions in e.g. Milgrom and Weber [1985] and Fudenberg and Levine [1986].

Proposition 8. *Suppose that $\{\mathcal{P}_j\}_{j\in\mathbb{N}}$ is a sequence of primitives that converges to $\mathcal{P}$. Suppose further that $p_j \in \Delta(\Theta \times X_j \times Y_j)$ is a contracting equilibrium outcome for $\mathcal{P}_j$ and $\lim_{j\to\infty} p_j = p$ for some $p \in \Delta(\Theta \times X \times Y)$. Then p is a contracting equilibrium outcome for $\mathcal{P}$.*

Aside from its usefulness for the proof of the existence of contracting equilibria, the upper hemicontinuity established Proposition 8 is a desirable property in its own right. In particular, it shows that small perturbations in the underlying primitives do not result in drastically different contracting equilibrium outcomes. This helps justify the study of principal-agent games with continuum action spaces even if the action spaces in reality are finite, as long as the action spaces are very fine.

To prove Proposition 8, we first show that the correspondence mapping primitives into incentive compatible outcomes is upper hemicontinuous.

Lemma 1. *Consider a sequence of primitives $\{\mathcal{P}_j\}_{j\in\mathbb{N}}$ that converges to the primitives $\mathcal{P}$, and suppose that $p_j \in \Delta(\Theta \times X \times Y)$ is an incentive compatible outcome in the game corresponding to $\mathcal{P}_j$ for each $j \in \mathbb{N}$. If $\lim_{j\to\infty} p_j = p$, then p is an incentive compatible outcome in the game corresponding to $\mathcal{P}$.*

Proof. Since each p_j is incentive compatible, $\mathbb{E}_{p_j}[U(\theta, x, y)|\theta] \geq \mathbb{E}_{p_j}[U(\theta, x, y)|\theta']$ for all θ and θ'. As $\lim_{j\to\infty} p_j = p$, it follows that $\lim_{j\to\infty} \mathbb{E}_{p_j}[U(\theta, x, y)|\theta] = \mathbb{E}_p[U(\theta, x, y)|\theta]$ and $\lim_{j\to\infty} \mathbb{E}_{p_j}[U(\theta, x, y)|\theta'] = \mathbb{E}_p[U(\theta, x, y)|\theta']$. Therefore, $\mathbb{E}_p[U(\theta, x, y)|\theta] \geq \mathbb{E}_p[U(\theta, x, y)|\theta']$, so Condition 1 of the definition of incentive compatibility is satisfied. Similar arguments show that Conditions 2 and 3 are satisfied as well.

It remains to show that $\mathbb{P}[y \in \arg\max_{y'\in Y} \mathbb{E}[V(\theta, x, y')|(x, y)]] = 1$. Suppose otherwise that $\mathbb{P}[y \in \arg\max_{y'\in Y} \mathbb{E}[V(\theta, x, y')|(x, y)]] < 1$. Then there are closed sets $\widetilde{X} \subseteq X$ and $\widetilde{Y} \subseteq Y$, as well as an agent action $\hat{y}$, such that $\mathbb{E}_p[\mathbb{1}_{\widetilde{X}\times\widetilde{Y}}(x, y)V(\theta, x, \hat{y})] > \mathbb{E}_p[\mathbb{1}_{\widetilde{X}\times\widetilde{Y}}(x, y)V(\theta, x, y)]$. For every $\varepsilon > 0$, let $\widetilde{X}_{<\varepsilon} = \{x \in X : d(x, \widetilde{X}) < \varepsilon\}$ and $\widetilde{X}_{\leq\varepsilon} = \{x \in X : d(x, \widetilde{X}) \leq \varepsilon\}$, and similarly let $\widetilde{Y}_{<\varepsilon}$, $\widetilde{Y}_{\leq\varepsilon}$ denote the corresponding sets for Y. Additionally, let $\underline{V} = \min_{(\theta,x,y)\in\Theta\times X\times Y} V(\theta, x, y)$. By continuity, there exists

some $\tilde{\varepsilon} > 0$ such that

$$\begin{aligned}&\mathbb{E}_p[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}\times\widetilde{Y}_{<\tilde{\varepsilon}}}(x,y)(V(\theta,x,\hat{y})-\underline{V})]+\mathbb{E}_p[(1-\mathbb{1}_{\widetilde{X}_{\leq\tilde{\varepsilon}}\times\widetilde{Y}_{\leq\tilde{\varepsilon}}}(x,y))(V(\theta,x,y)-\underline{V})]\\ >&\mathbb{E}_p[V(\theta,x,y)]-\underline{V}.\end{aligned}$$

As $\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}\times\widetilde{Y}_{<\tilde{\varepsilon}}}(x,y)(V(\theta,x,\hat{y})-\underline{V})$ is a lower semicontinuous function of (x,y), it follows that $\liminf_{j\to\infty}\mathbb{E}_{p_j}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}\times\widetilde{Y}_{<\tilde{\varepsilon}}}(x,y)(V(\theta,x,\hat{y})-\underline{V})]\geq\mathbb{E}_p[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}\times\widetilde{Y}_{<\tilde{\varepsilon}}}(x,y)(V(\theta,x,\hat{y})-\underline{V})]$. Likewise, $(1-\mathbb{1}_{\widetilde{X}_{\leq\tilde{\varepsilon}}\times\widetilde{Y}_{\leq\tilde{\varepsilon}}}(x,y))(V(\theta,x,y)-\underline{V})$ is a lower semicontinuous function of $(x,y)\in X\times Y$, so $\liminf_{j\to\infty}\mathbb{E}_{p_j}[(1-\mathbb{1}_{\widetilde{X}_{\leq\tilde{\varepsilon}}\times\widetilde{Y}_{\leq\tilde{\varepsilon}}}(x,y))(V(\theta,x,y)-\underline{V})]\geq\mathbb{E}_p[(1-\mathbb{1}_{\widetilde{X}_{\leq\tilde{\varepsilon}}\times\widetilde{Y}_{\leq\tilde{\varepsilon}}}(x,y))(V(\theta,x,y)-\underline{V})]$. Consequently, for sufficiently high $j\in\mathbb{N}$,

$$\begin{aligned}&\mathbb{E}_{p_j}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}\times\widetilde{Y}_{<\tilde{\varepsilon}}}(x,y)(V(\theta,x,\hat{y})-\underline{V})]+\mathbb{E}_p[(1-\mathbb{1}_{\widetilde{X}_{\leq\tilde{\varepsilon}}\times\widetilde{Y}_{\leq\tilde{\varepsilon}}}(x,y))(V(\theta,x,y)-\underline{V})]\\ >&\mathbb{E}_{p_j}[V(\theta,x,y)]-\underline{V}.\end{aligned}$$

This implies that $\mathbb{E}_{p_j}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}\times\widetilde{Y}_{<\tilde{\varepsilon}}}(x,y)V(\theta,x,\hat{y})]+\mathbb{E}_p[(1-\mathbb{1}_{\widetilde{X}_{\leq\tilde{\varepsilon}}\times\widetilde{Y}_{\leq\tilde{\varepsilon}}}(x,y))V(\theta,x,y)]>\mathbb{E}_{p_j}[V(\theta,x,y)]$, which contradicts p_j being incentive compatible. ■

All that remains to be shown is that, for every mechanism, there is a sequential continuation equilibrium that deters every principal type from proposing the mechanism. This is non-trivial because, as illustrated by example in Section OA.6, the correspondence from mechanisms to sequential continuation equilibria is not in general upper hemicontinuous. In Section OA.7, we prove the following lemma, which shows that, for any mechanism in the limit environment, there is a sequence of mechanisms corresponding to the approximating primitives for which any limit of sequential continuation equilibrium outcomes is a sequential continuation equilibrium outcome after the mechanism in the limit environment is proposed. Since each of the mechanisms in the sequence must have a sequential continuation equilibrium outcome that gives every principal type a lower payoff than the corresponding contracting equilibrium outcome, this enables us to show that the final contracting equilibrium outcome condition is satisfied.

Lemma 2. *Consider a sequence of primitives $\{\mathcal{P}_j\}_{j\in\mathbb{N}}$ that converges to the original primitives $\mathcal{P}$. For every mechanism $(\mu, M_P) \in \mathcal{M}$, there is a sequence of mechanisms $(\mu_j, M_P) \in \mathcal{M}_j$ such that any limit of sequential continuation equilibrium outcomes after these mechanisms are proposed is a sequential continuation equilibrium outcome after (μ, M_P) is proposed.*

Proof of Proposition 8. Fix a sequence of primitives $\{\mathcal{P}_j\}_{j\in\mathbb{N}}$ that converges to the original primitives $\mathcal{P}$, and let $p_j \in \Delta(\Theta \times X \times Y)$ be a contracting equilibrium outcome in the game corresponding to $\mathcal{P}_j$. By restricting attention to a convergent subsequence if necessary, there is some $p \in \Delta(\Theta \times X \times Y)$ such that $\lim_{j\to\infty} p_j = p$. By Lemma 1, p is incentive compatible.

We now argue that, for every mechanism $(\mu, M_P) \in \mathcal{M}$, there is a sequential continuation equilibrium after the mechanism is proposed which gives every principal type a weakly lower expected utility than p. Since each p_j is a contracting equilibrium outcome, there is a sequential continuation equilibrium after any mechanism is proposed that gives each principal type a lower payoff than p_j. Moreover, by Lemma 2, there is a sequence of mechanisms $\{(\mu_j, M_P)\}_{j\in\mathbb{N}}$ such that any limit of sequential continuation equilibria after these mechanisms are proposed is a sequential continuation equilibrium after (μ, M_P) is proposed. Therefore, there is some sequential continuation equilibrium after (μ, M_P) is proposed that gives each principal type a lower payoff than p. ■

Now we establish the existence of contracting equilibria when the action and recommendation spaces are finite but we do not restrict the space of mechanisms. Since every action and recommendation space can be approximated to arbitrary accuracy by a sequence of finite action/recommendation spaces, this existence result combined with Proposition 8 implies that contracting equilibria exist in general.

Lemma 3. *If the principal's action space X, recommendation space R, and agent action space are all finite, then a contracting equilibrium exists.*

Proof. For a given $j \in \mathbb{N}$, consider the set of mechanisms $\mathcal{M}_j$ that (1) have no more than j principal messages and (2) are such that the probability of a given principal

action-recommendation pair conditional on any message is some integer multiple of $1/j$:

$$\mathcal{M}_j = \Big\{(\mu, M_P) \in \mathcal{M} : \ (1)\ |M_P| \leq j,$$
$$(2)\ \forall m_P \in M_P, x \in X, r \in R, \exists k \in \{1, ..., j\} \text{ s.t. } \mu(x, r|m_P) = \frac{k}{j}\Big\}.$$

Because both X and R are finite, $\mathcal{M}_j$ is well-defined and non-empty for all sufficiently high j. For the remainder of the proof, we restrict attention to such j.

Consider the modified principal-agent game in which the principal can only propose a mechanism belonging to $\mathcal{M}_j$. Since $\mathcal{M}_j$ is finite, standard arguments show that this game has a contracting equilibrium. Let $p_j \in \Delta(\Theta \times X \times Y)$ be an outcome corresponding to a contracting equilibrium, and suppose (by restricting attention to a convergent subsequence if necessary) that $\lim_{j\to\infty} p_j = p$. As each p_j is incentive compatible, Lemma 1 ensures that p is incentive compatible.

We now show that, for every mechanism $(\mu, M_P) \in \mathcal{M}$, there is a sequential continuation equilibrium after the mechanism is proposed which gives every principal type a weakly lower expected utility than p. By construction, there is some sequence of mechanisms satisfying $(\mu_j, M_{j,P}) \in \mathcal{M}_j$ such that $M_{j,P} = M_P$ for all j, and $\lim_{j\to\infty} \mu_j(x, r|m_P) = \mu(x, r|m_P)$ for all x, r, and m_P. For each j, there is a sequential continuation equilibrium after $(\mu_j, M_{j,P}, M_{j,A})$ is proposed which gives each principal type a weakly lower payoff than p_j. Let $(\tilde{\lambda}_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A}, \Lambda_j, \beta_{j,A})$ be the corresponding sequential continuation equilibrium after $(\mu_j, M_{j,P}, M_{j,A})$ is accepted. By restricting attention to a convergent subsequence if necessary, there is some consistent assessment $(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \Lambda, \beta_A)$ after mechanism (μ, M_P) is accepted that is the limit of the $(\tilde{\lambda}_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A}, \Lambda_j, \beta_{j,A})$. Since X, R, and A are all finite, standard arguments show that $(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \Lambda, \beta_A)$ is a sequential continuation equilibrium after mechanism (μ, M_P) is accepted. Moreover, either all principal types get a lower payoff than p from this sequential continuation equilibrium, or it is optimal for the agent to reject the proposal of (μ, M_P). In either case, there is a sequential continuation equi-

librium such that every principal type gets a lower payoff than from p by proposing (μ, M_P). ■

Proof of Theorem 1. Fix a sequence of primitives $\{\mathcal{P}_j\}_{j\in\mathbb{N}}$ that converges to the original primitives $\mathcal{P}$ and is such that the principal action space X_j, recommendation space R_j, and agent action space Y_j are all finite for every $j \in \mathbb{N}$. For each j, let $p_j \in \Delta(\Theta\times X\times Y)$ be a contracting equilibrium outcome in the game corresponding to $\mathcal{P}_j$, the existence of which is guaranteed by Lemma 3. By restricting attention to a convergent subsequence if necessary, there is some $p \in \Delta(\Theta \times X \times Y)$ such that $\lim_{j\to\infty} p_j = p$. By Proposition 8, p is a contracting equilibrium outcome. ■

B.2 Proof of Theorem 1 for the Deterministic-Mechanism Principal-Agent Game

Our general approach will be to take a sequence of finite principal-agent games that converges in the limit to the true game. We will show that the limits of the equilibrium outcomes of these games correspond to contracting equilibrium outcomes in the true principal-agent game.

Let $\{X_j\}_{j\in\mathbb{N}}$, $\{Y_j\}_{j\in\mathbb{N}}$, $\{R_j\}_{j\in\mathbb{N}}$ be sequences of finite action and recommendation sets such that $\lim_{j\to\infty} X_j = X$, $\lim_{j\to\infty} Y_j = Y$, and $\lim_{j\to\infty} R_j = R$. For a given $j \in \mathbb{N}_{++}$, consider the set of mechanisms

$$\mathcal{M}_j = \Big\{(\mu, M_P) \in \mathcal{M} :\ (1)\ |M_P| \le j,$$
$$(2)\ \forall m_P \in M_P,\ \exists x \in X_j, r \in R_j,\ \text{s.t. } \mu(x, r|m_P) = 1\Big\}$$

that (1) have no more than j principal messages and (2) are such that every principal message results in some principal-action-transfer-recommendation tuple that belongs to $X_j \times R_j$.

We now describe the strategy space of the type θ principal in the (j, k) game. Part of this player's choice is over which mechanisms to propose. We force θ to propose all

mechanisms with probability at least $1/k$, so the distribution over mechanism proposals used by θ must belong to

$$\Delta_{j,k,\theta}(\mathcal{M}_j) = \left\{ \eta \in \Delta(\mathcal{M}_j) : \ \eta[(\mu, M_P)] \geq \frac{1}{k} \ \forall (\mu, M_P) \in \mathcal{M}_j \setminus \mathcal{M}^0_{j,\theta} \right\}.$$

Moreover, when a given mechanism is accepted, we force θ to tremble and play every message in the mechanism with probability at least $1/k$. Formally, the distribution over messages used by θ when mechanism (μ, M_P) is accepted must belong to

$$\Pi_{j,k,P}(\mu, M_P) = \left\{ \pi_P \in \Delta(M_P) : \pi_P[m_P] \geq \frac{1}{k} \ \forall m_P \in M_P \right\}.$$

A valid strategy for θ in the (j,k) game is any pair $(\eta_\theta, \boldsymbol{\pi}_\theta(\cdot))$ consisting of a $\eta_\theta \in \Delta_{j,k,\theta}(\mathcal{M}_j)$ and a rule $\boldsymbol{\pi}_\theta(\cdot)$ for how to play when an arbitrary mechanism is accepted that satisfies $\boldsymbol{\pi}_\theta(\mu, M_P) \in \Pi_{j,k,P}(\mu, M_P)$.

The strategy space of the agent is unaltered from the principal-agent game, aside from the addition of trembles. For every mechanism (μ, M_P), we require the probability α that the agent accepts its proposal to be no less than $1/k$. Additionally, we require the agent to tremble in their choices of actions. In particular, for every mechanism (μ, M_P) and principal action-recommendation pair (x, r), the agent's choice of action must be a distribution belonging to

$$\Delta_k(Y_k) = \left\{ \mathbf{y} \in \Delta(Y_k) : \mathbf{y}[y] \geq \frac{1}{k|Y_k|} \ \forall y \in Y_k \right\}.$$

A valid strategy for the agent in the (j,k) game is any pair $(\boldsymbol{\alpha}(\cdot), \boldsymbol{\beta}(\cdot))$ consisting of (1) a rule governing the probability of mechanism acceptance, $\boldsymbol{\alpha}(\cdot)$, satisfying $\boldsymbol{\alpha}(\mu, M_P) \geq 1/k$ for all $(\mu, M_P) \in \mathcal{M}_j$ and (2) a rule governing the agent's choice of actions $\boldsymbol{\beta}(\cdot)$ satisfying $\boldsymbol{\beta}(\mu, M_P) \in \Delta_k(Y_k)^{X_j \times R_j}$ for all $(\mu, M_P) \in \mathcal{M}_j$.

The payoffs of both the principal and agent are exactly as in the true principal-agent game. Standard arguments show that Nash equilibria exist in the (j,k) game

and that the play after any mechanism is proposed (combined with the corresponding distribution over the principal's type) constitutes a sequential continuation equilibrium in the $k \to \infty$ limit. Throughout the remainder of the argument, we will let $p_{j,k} \in \Delta(\Theta \times \mathcal{M}_j \times [0,1] \times X_j \times Y_k)$ denote an equilibrium outcome of the (j,k) game. Additionally, we assume (by restricting attention to convergent subsequences if necessary) that $\lim_{k\to\infty} p_{j,k} = p_j$ for some $p_j \in \Delta(\Theta \times \mathcal{M}_j \times [0,1] \times X_j \times Y)$, and that $\lim_{j\to\infty} p_j = p$ for some $p \in \Delta(\Theta \times \mathcal{M} \times [0,1] \times X \times Y)$. (Recall that the topology over mechanisms that we use in defining appropriate Borel sets is that induced by the metric in Definition 17.)

Lemma 4. *There is a profile of mechanism proposal distributions $\{\eta_\theta\}_{\theta\in\Theta}$ and a measurable mapping $\tau^* : \mathcal{M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^\Theta$ that takes each mechanism $(\mu, M_P) \in \mathcal{M}$ into a tuple consisting of a distribution over the principal's type and a distribution over $(\alpha, x, y) \in [0,1] \times X \times Y$ for each principal type that corresponds to a single sequential continuation equilibrium after (μ, M_P) is proposed such that*

1. *There is a regular conditional probability distribution obtained from λ and $\{\eta_\theta\}_{\theta\in\Theta}$ that, for every $(\mu, M_P) \in \mathcal{M}$, induces the $\Delta(\Theta)$ component of $\tau^*(\mu, M_P)$ as the belief over the principal's type following the proposal of $(\mu, M_P) \in \mathcal{M}$,*
2. *$U(\theta, \tau^*(\mu, M_P)) \leq U(\theta, p)$ for all $\theta \in \Theta$ and $(\mu, M_P) \in \mathcal{M}$, and*
3. *$\{\eta_{j,\theta}\}_{\theta\in\Theta}$ combined with $\tau^*(\mu, M_P)$ following the proposal of each $(\mu, M_P) \in \mathcal{M}$ induces the same distribution over (θ, α, x, y) as outcome p.*

We handle the proof of Lemma 4, which is given in Section OA.8, in two steps. The first involves constructing valid on-path play consistent with the same distribution over (θ, α, x, y) as in p occurring in a contracting equilibrium outcome. (In terms of the conditions of Lemma 4, this amounts to satisfying Condition 1 and Condition 3 as well as Condition 2 for all on-path mechanisms (μ, M_P).) The second involves showing that there is valid off-path play that deters every principal type when they receive the same payoff as in p. (This corresponds to Condition 2 being satisfied for all off-path mechanisms.)

Whereas in the general-mechanism game the main difficulty in showing the existence of contracting equilibria was identifying sequential continuation equilibria that deterred the principal types from proposing off-path mechanisms, here the main obstacle is handling the first step. Part of the reason for this is the inability to invoke the Inscrutability Principle, which arises from the lessened design power of the principal. Additionally, the space of mechanisms is not compact. These two features complicate the finding of valid distributions over mechanisms that lead to a desired outcome distribution.

We approach the first step by focusing on the class of *binary* and *obedient* mechanisms. These are mechanisms with precisely two messages in which the recommendations tied to the messages encode information that gives a direct recommendation to the agent about which actions to take as well as which beliefs to hold about the principal's type when the mechanism is proposed and the probability distribution over the messages in the mechanism the various principal types use. Such a mechanism is obedient when the content conveyed by the recommendations exactly captures the sequential continuation equilibrium played after the mechanism is proposed.

We focus on these mechanisms for two reasons. First, binary and obedient mechanisms suffice to replicate each of the p_j. Flexible contracts enable outcomes in which there are positive probabilities of principal actions for which the agent's conditional expected utility is negative because these actions can be incorporated in contracts which also incorporate principal actions for which the agent's conditional expected utilities are positive. The second reason for focusing on binary and obedient mechanisms is that the space of such mechanisms is compact, and obedience ensures that the sequential continuation equilibrium following mechanism proposal is a continuous function of the mechanism. These features aid the demonstration of various convergence properties, and facilitate the proofs of various facts about sequential continuation equilibria. Ultimately, we show that there are mechanism proposal distributions concentrating on the class of binary and obedient mechanisms that, when combined with obedient play following the proposal of any mechanism in the support of these distributions, induces

p as the outcome.

Intuitively, two messages are sufficient because principal actions for which the agent's conditional expected utilities are positive can be paired off with principal actions for which the agent's conditional expected utilities. By appropriately choosing the proposal probabilities of these paired-off mechanisms as well as the relative probabilities of each of the two actions being chosen when after mechanism acceptance, we can ensure that the agent is still willing to accept contracts with actions that give them negative conditional expected utility.

To handle the second step, we consider the class of *revealing* mechanisms in which, for each message, the recommendation received by the agent precisely reveals the message chosen by the principal. This class is useful because the correspondence mapping revealing mechanisms into sequential continuation equilibria following their proposal is upper hemicontinuous. (This upper hemicontinuity property fails when considering the full class of mechanisms, as can be seen by example in Section OA.6.) Additionally, there is a canonical way which identifies each mechanism with a unique revealing mechanism. We ultimately use the upper hemicontinuity property discussed above to identify certain measurable mappings from revealing mechanisms into sequential continuation equilibria of interest, and then extend these mappings to the full set of mechanisms using the canonical mapping discussed above.

Proof of Theorem 1 for the Deterministic-Mechanism Game. Let $\hat{\tau}^*$ be a rule governing sequential continuation equilibria that, for every mechanism in $\mathcal{M}$, results in the same sequential continuation equilibrium outcome as that given by τ^*. We will argue that $\{\eta_\theta\}_{\theta\in\Theta}$ and $\hat{\tau}^*$ together constitute a contracting equilibrium. Condition 2 of Definition 19 follows from Condition 1 of Lemma 4. Moreover, the measurability of τ^* guarantees that $\{(\mu, M_P) \in \mathcal{M} : U(\theta, \tau^*(\mu, M_P)) = U(\theta, p)\}$ is a measurable subset of $\mathcal{M}$, and Conditions 2 and 3 of Lemma 4 imply that $\eta_\theta[\{(\mu, M_P) \in \mathcal{M} : U(\theta, \tau^*(\mu, M_P)) = U(\theta, p)\}] = 1$ for all $\theta \in \Theta$. This, along with Condition 2 of Lemma 4, implies that $\arg\max_{(\mu,M_P)} U(\theta, \tau(\mu, M_P)) = \{(\mu, M_P) \in \mathcal{M} : U(\theta, \tau^*(\mu, M_P)) = U(\theta, p)\}$ for all

$\theta \in \Theta$, so we conclude that Condition 1 of Definition 19 is satisfied. ■

C Other Proofs

C.1 Proof of Theorem 2

C.1.1 Proof of Theorem 2 for the General-Mechanism Game

The following lemma shows that there is a sequence of outcomes that has various properties that are useful in the proof of Theorem 2.

Lemma 5. *In MCS environments, there are sequences of full-support distributions over the principal type* $\{\lambda_k\}_{k\in\mathbb{N}}$ *and outcomes* $\{p_k\}_{k\in\mathbb{N}}$ *such that*

1. $marg_\Theta p_k = \lambda_k$ *for all* $k \in \mathbb{N}$,
2. $\liminf_{k\to\infty} \mathbb{E}_{p_k}[v(\theta, x, y) + g(t)|\theta] \geq 0$ *for all* $\theta \in \Theta$,
3. $\mathbb{E}_{p_k}[u(\theta, x, y) - t|\theta] \geq \mathbb{E}_{p_k}[u(\theta, x, y) - t|\theta']$ *for all* $\theta, \theta' \in \Theta$ *and* $k \in \mathbb{N}$,
4. $\mathbb{P}_{p_k}[y = y^*(\theta, x)|\theta, x \neq x_o] = 1$ *for all* $\theta \in \Theta$ *and* $k \in \mathbb{N}$, *and*
5. *For each mechanism* $(\mu, M_P) \in \mathcal{M}$ *and* $k \in \mathbb{N}$, *there is a sequential continuation equilibrium after* (μ, M_P) *is proposed that gives every principal type a payoff no more than* $1/k$ *greater than that from* p_k.

The first condition simply states that each outcome p_k is consistent with the corresponding distribution over principal types λ_k. The second condition says that, in the $k \to \infty$ limit, the agent receives a non-negative expected utility conditional on each principal type. The third condition captures principal incentive compatibility with the outcomes, while the fourth says that, with probability 1, the agent takes the same action they would given the knowledge of the principal's type.

The proof of Lemma 5, which is in Section OA.9.1, constructs a sequence of hypothetical games and establishes that the outcomes and distributions over principal types corresponding to the equilibria of these games have the desired properties. Section 5.2

gave a brief description of some of the modifications we make to the principal-agent games in this sequence. Here we give a fuller description.

As discussed before, the principal-agent game is modified so that there are costs (and benefits) to the principal types from using mechanisms of the form $(\mu_{\chi,\theta}, \{0\})$, the proposal of which is assumed to automatically induce the outcome in which the agent accepts and then plays $y^*(\theta, x)$ after observing any $x \neq x_o$. To avoid possible violations of agent incentive compatibility this modification may cause, we make it prohibitively costly for (or equivalently do not allow) a type θ principal to propose any mechanism of the form $(\mu_{\chi,\theta'}, \{0\})$ where $\theta' \neq \theta$. We also make it costly for the type θ principal to propose a $(\mu_{\chi,\theta}, \{0\})$ mechanism whenever (1) there is some other principal type who would get a higher payoff from proposing $(\mu_{\chi,\theta}, \{0\})$ (if the agent were to accept and respond according to $y^*(\theta, x)$) than they do from the prevailing outcome, or (2) the agent's total expected utility in the prevailing outcome conditional on θ is too low. The costs arising from Case (1) ensure principal incentive compatibility. The costs arising from from Case (2) ensure that the agent's expected utility conditional on any principal type who uses this class of mechanisms is not much lower than their outside option. To encourage pooling when these costs are 0 and a type θ principal would otherwise be indifferent with proposing a $(\mu_{\chi,\theta}, \{0\})$ mechanism, we give the type θ principal a flat benefit of $1/k$ from proposing mechanisms of this form.

Even with these costs, pooling cannot occur in equilibrium. This follows from the modifications described above for the case where the agent's expected utility conditional on the highest principal type involved in pooling is not too low. This is because the highest type involved in pooling would be strictly better off proposing a mechanism that commits to a distribution over action-transfer pairs that matches the distribution they realize in equilibrium but for an increase in the transfer level. By supermodularity, the increase in the transfer level can be chosen so that only the highest type would want to make this proposal. We also rule out the possibility of pooling being sustained by the highest principal type involved in pooling giving the agent a significantly lower conditional expected utility than their outside option. We do so by adding a hypothet-

ical third player to the game who selects the distribution over principal types with the objective of minimizing the agent's total expected utility from play over mechanisms outside the class of $(\mu_{\chi,\theta}, \{0\})$ mechanisms. Since the agent's play regarding mechanisms outside of this class is unrestricted, their total expected utility from such play must (weakly) exceed that of their outside option. So each principal type either (a) only plays mechanisms outside of $(\mu_{\chi,\theta}, \{0\})$ class, in which case they must be giving the agent an expected utility close to the outside option, or (b) plays mechanisms in the $(\mu_{\chi,\theta}, \{0\})$ class, which, as noted before, they are only willing to do when the agent's expected utility conditional on their type is not too low.

Proof of Theorem 2 for the General-Mechanism Game. For each $\theta \in \Theta$ and $k \in \mathbb{N}$, let $q_k(\theta)$ denote the conditional distribution obtained from p_k given θ. By restricting attention to a convergent subsequence if necessary, there is some $q(\theta) \in \Delta(X \times T \times Y)$ such that $\lim_{k\to\infty} q_k(\theta) = q(\theta)$ for all $\theta \in \Theta$. Conditions 2, 3, and 4 of Lemma 5, along with continuity, imply that (1) $\mathbb{P}_{q(\theta)}[y = y^*(\theta, x) | x \neq x_o] = 1$, (2) $\mathbb{E}_{q(\theta)}[v(\theta, x, y) + g(t)] \geq 0$, and (3) $\mathbb{E}_{q(\theta)}[u(\theta, x, y) - t] \leq \mathbb{E}_{q(\theta')}[u(\theta, x, y) - t]$ for all $\theta, \theta' \in \Theta$. By Proposition 2, it follows that each principal type θ obtains a weakly lower payoff from $q(\theta)$ than they do from the principal-optimal safe outcomes. Condition 5 of Lemma 5 implies that, for every mechanism, there is a sequential continuation equilibrium after the mechanism is proposed that gives every principal type a weakly lower payoff than $q(\theta)$, and thus from the principal-optimal safe outcomes. Combining this with the incentive compatibility of the principal-optimal safe outcome shows that it is a contracting equilibrium outcome. ■

C.1.2 Proof of Theorem 2 for the Deterministic-Mechanism Game

The following lemma plays an analogous role in the proof of Theorem 2 game to that of Lemma 5 in the proof for the general-mechanism game.

Lemma 6. *In MCS environments, there are sequences of full-support distributions over the principal type* $\{\lambda_k\}_{k\in\mathbb{N}}$ *and outcomes* $\{p_k\}_{k\in\mathbb{N}}$ *such that*

1. $marg_{\Theta} p_k = \lambda_k$ *for all* $k \in \mathbb{N}$,
2. $\liminf_{k\to\infty} \mathbb{E}_{p_k}[\alpha(v(\theta, x, y) + g(t))|\theta] \geq 0$ *for all* $\theta \in \Theta$,
3. $\mathbb{P}_{p_k}[U(\theta, p_k) \geq \alpha(u(\theta, x, y^*(\theta', x)) - t)|\theta', x, t, \alpha] = 1$ *for all* $\theta, \theta' \in \Theta$ *and* $k \in \mathbb{N}$,
4. $\mathbb{P}_{p_k}[y = y^*(\theta, x)|\theta, x \neq x_o] = 1$ *for all* $\theta \in \Theta$ *and* $k \in \mathbb{N}$, *and*
5. *For each mechanism* $(\mu, M_P) \in \mathcal{M}$ *and* $k \in \mathbb{N}$, *there is a sequential continuation equilibrium after* (μ, M_P) *is proposed that gives every principal type a payoff no more than* $1/k$ *greater than that from* p_k.

Conditions 1, 2, 4, and 5 are exactly as in Lemma 5. Condition 3 is similar to the corresponding principal incentive compatibility condition in Lemma 5, though it is strengthened to fit the principal's lower commitment power in the deterministic-mechanism game.

The proof of Lemma 6, which is in Section OA.9.2, proceeds in a similar fashion to the proof of Lemma 5: It too constructs a sequence of hypothetical games and uses essentially the same arguments, adapted to the deterministic-mechanism game, to establish that the outcomes and distributions over principal types corresponding to the equilibria of these games have the desired properties.

Proof of Theorem 2 for the Deterministic-Mechanism Game. For each $\theta \in \Theta$ and $k \in \mathbb{N}$, let $q_k(\theta)$ denote the conditional distribution obtained from p_k given θ. By restricting attention to a convergent subsequence if necessary, there is some $q(\theta) \in \Delta([0, 1] \times X \times T \times Y)$ such that $\lim_{k\to\infty} q_k(\theta) = q(\theta)$ for all $\theta \in \Theta$. Conditions 2, 3, and 4 of Lemma 6, along with continuity, imply that (1) $\mathbb{P}_{q(\theta)}[y = y^*(\theta, x)|x \neq x_o] = 1$, (2) $\mathbb{E}_{q(\theta)}[\alpha(v(\theta, x, y) + g(t))] \geq 0$, and (3) $\mathbb{P}_{q(\theta)}[U(\theta', q(\theta')) \geq \alpha(u(\theta', x, y) - t)|\theta, x, t, \alpha] = 1$ for all $\theta, \theta' \in \Theta$. Thus, for each $\theta \in \Theta$, there is some $(x, t) \in X \times T$ and $\alpha \in [0, 1]$ such that (1) $U(\theta, q(\theta)) = \alpha(u(\theta, x, y^*(\theta, x)) - t)$, (2) $U(\theta', q(\theta')) \geq \alpha(u(\theta', x, y^*(\theta, x)) - t)$ for all $\theta' \neq \theta$, and (3) $v(\theta, x, y^*(\theta, x)) + g(t) \geq 0$. Observe that, for $t' = \alpha t + (1 - \alpha)u(\theta, x, y^*(\theta, x))$, (1) $U(\theta, q(\theta)) = u(\theta, x, y^*(\theta, x)) - t'$, (2) $U(\theta', q(\theta')) \leq u(\theta', x, y^*(\theta, x)) = t'$, and (3) $v(\theta, x, y^*(\theta, x)) + g(t) \geq 0$. By Proposition OA 6, it follows that each principal type θ obtains a weakly lower payoff from $q(\theta)$

than they do from the principal-optimal safe outcomes. Condition 5 of Lemma 6 implies that, for every mechanism, there is a sequential continuation equilibrium after the mechanism is proposed that gives every principal type a weakly lower payoff than $q(\theta)$, and thus from the principal-optimal safe outcomes. We conclude the principal-optimal safe outcomes are contracting equilibrium outcomes. ■

C.2 Proof of Proposition 3

Proof. Consider a principal-optimal safe outcome $p \in \Delta(\Theta \times X \times Y)$, and suppose that $p' \in \Delta(\Theta \times X \times Y)$ is an incentive compatible outcome that payoff dominates p. Because p is a contracting equilibrium outcome, by definition there is a sequential continuation equilibrium after any mechanism is proposed that gives each principal type a weakly lower payoff than they obtain from p. Combining this with the fact that p' payoff dominates p, we conclude that (μ', Θ) corresponds to a contracting equilibrium with outcome p'. ■

C.3 Proof of Theorem 3.1

Proof of Theorem 3.1 for the General-Mechanism Game. Let p be a payoff-plausible outcome. Suppose towards a contradiction that there is some $n \in \{1, ..., N\}$ for which θ_n obtains a lower expected utility than their principal-optimal safe payoff, and let n be the lowest such value. Let $q^*(\theta_n)$ be the conditional outcome given θ_n in a principal-optimal safe outcome. By definition, $\text{marg}_{X \times T} q^*(\theta_n)$ satisfies the agent individual rationality constraints in the type-θ_n optimization problem in (1). Moreover, since each $\theta_{n'}$ for $n < n'$ obtains a weakly higher expected utility than their principal-optimal safe payoff, it follows that $\text{marg}_{X \times T} q^*(\theta_n)$ satisfies the third constraint in the optimization problem. Thus, we have $U(\theta_n, p) \geq \mathbb{E}_{q^*(\theta_n)}[u(\theta_n, x, y) - t]$, which contradicts θ_n obtaining a lower expected utility than their principal-optimal safe payoff. ■

Proof of Theorem 3.1 for the Deterministic-Mechanism Game. Let p be a payoff-plausible outcome. Suppose towards a contradiction that there is some $n \in \{1, ..., N\}$ for which

θ_n obtains a lower expected utility than their principal-optimal safe payoff, and let n be the lowest such value. Let $(x, t, y^*(\theta_n, x))$ be such that it gives type θ_n their principal-optimal safe payoff, it gives every lower type $\theta_{n'}$ for $n' < n$ a weakly lower payoff than their principal-optimal safe payoff, and it gives the agent a weakly positive utility conditional on type θ_n. Then (x, t) satisfies the agent incentive compatibility and individual rationality constraints in the type-θ_n optimization problem in (2). Moreover, since each $\theta_{n'}$ for $n < n'$ obtains a weakly higher expected utility than their principal-optimal safe payoff, it follows that (x, t) satisfies the third constraint in the optimization problem. Thus, we have $U(\theta_n, p) \geq u(\theta_n, x, y) - t$, which contradicts θ_n obtaining a lower expected utility than their principal-optimal safe payoff. ■

Online Appendix for "The Informed Principal with Agent Moral Hazard"

Daniel Clark

November 8, 2021

OA.1 Firm-Worker Equilibrium Payoffs Computations

OA.1.1 Upper Envelope with Flexible Contracts

Proposition OA 1. *The upper envelope of the firm payoffs sustainable in contracting equilibrium with flexible contracts is*

$$\big\{(U(2),U(4)) \in \mathbb{R}^2 : 2 \leq U(2) \leq 3 \text{ and } U(4) = U(2)+2, \text{ or } 3 \leq U(2) \leq 5 \text{ and } U(4) = -\frac{1}{2}U(2)^2 + 4U(2) - \frac{5}{2}\big\}.$$

Lemma OA 1. *In any Pareto-optimal contracting equilibrium,* $\mathbb{P}[(s,e) = (1,2)|\theta = 2] = 1$.

Proof. Consider an outcome p for which $\mathbb{P}_p[(s,e) = (1,2)|\theta = 2] < 1$. Let p' be the outcome obtained by modifying p as follows: Conditional on 2, every (s,t,e) is changed to $(1,-U(2,p),2)$, and, conditional on 4, every (s,t,e) is shifted to $(s,t,4s)$.

By construction, the expected utility of the 2 firm is the same under p and p'. Moreover, since $e \leq 4s$ holds with probability 1 under p, it follows that the expected utility of the 4 firm is weakly higher under p'. Finally, observe that setting $s = 1$ and having the agent take effort $e = 2$ uniquely maximizes the total surplus given type 2. Combining this with the fact that the type 2 firm gets the same utility under p and p' as well as the fact that the worker's best response to any s is $4s$ given type 4, we conclude that the worker's expected utility is strictly higher under p' than p. Thus, p is not Pareto-optimal. ■

Lemma OA 2. *In any Pareto-optimal contracting equilibrium, there is some* $s^* \in [0, 1]$ *such that* $\mathbb{P}[(s, e) = (s^*, 4s^*)|\theta = 4] = 1$.

Proof. Consider a Pareto-optimal contracting equilibrium outcome p. By Lemma OA 1, $\mathbb{P}_p[(s, e) = (1, 2)|\theta = 2] = 1$, which implies that there is no pooling between the two firm types. Since $e = 4s$ is the worker's best response to any s under the belief that the firm's type is 4, we have that $\mathbb{P}_p[e = 4s|\theta = 4] = 1$. Therefore, the expected utility of each firm type θ from the conditional distribution of p given 4 is $\mathbb{E}_p[U(\theta, s, t, e)|\theta = 4] = \mathbb{E}_p[s(1 - s)|\theta = 4]\theta - \mathbb{E}_p[t|\theta = 4]$, while the corresponding expected utility of the worker is $\mathbb{E}_p[V(4, s, t, e)|\theta = 4] = 8\mathbb{E}_p[s^2|\theta = 4] + \mathbb{E}_p[t|\theta = 4]$. Let $s^* = \sqrt{\mathbb{E}_p[s^2|\theta = 4]}$. Since, for $s \geq 0$, $s(1-s)$ is a strictly concave function of s^2, Jensen's inequality implies that $\mathbb{E}_p[s(1 - s)|\theta = 4] \leq s^*(1 - s^*)$, with the inequality strict if $\mathbb{P}_p[s = s^*|\theta = 4] < 1$. Consider $t' = \mathbb{E}_p[t|\theta = 4] + 4(s^*(1 - s^*) - \mathbb{E}_p[s(1 - s)|\theta = 4])$. By construction, the outcome p' obtained from modifying p so that, conditional on 4, every (s, t, e) is changed to $(s^*, t', 4s^*)$ is incentive compatible and gives both firm types the same payoff as p. Moreover, p' would give the employee a strictly higher payoff than p if $\mathbb{P}_p[s = s^*|\theta = 4] < 1$. We thus conclude that $\mathbb{P}_p[s = s^*|\theta = 4] = 1$ since p is the outcome of a Pareto-optimal contracting equilibrium. ■

Proof of Proposition OA 1. First, observe that the type 2 firm can never get a lower payoff than 2 in a contracting equilibrium. The reason is that, for any $\varepsilon > 0$, the employee will accept the offer $(s, t) = (1, -2 + \varepsilon)$, which results in a payoff of $2 - \varepsilon$ to

the type 2 firm. Moreover, the type 2 firm can never achieve get a higher payoff than the maximum total expected surplus of 5 in a contracting equilibrium. This is because the payoff of the type 2 firm must always be weakly lower than the payoff of the type 4 firm, and the expected value of the firm's payoff can be no more than the maximum total expected surplus.

By Lemmas OA 1 and OA 2, the maximum payoff that the type 4 firm can obtain across the contracting equilibria in which the type 2 firm obtains a payoff of $U(2)$ is given by

$$\begin{aligned} &\max_{(s,t)\in[0,1]\times\mathbb{R}} 16s(1-s)-t \\ \text{s.t. } &\text{AIR: } \frac{1}{2}\left(8s^2+t\right)+\frac{1}{2}\left(2-U(2)\right)\geq 0, \\ &\text{PIC: } 8s(1-s)-t\leq U(2). \end{aligned} \tag{OA 1}$$

To understand the AIR constraint, observe that $8s^2+t$ is the agent's expected utility given (s,t) and type 4 when they respond with $4s$, and the agent's expected utility must be $2-U(2)$ when the type 2 firm plays $s=1$ with probability 1 and receives a payoff of $U(2)$.

We first solve this problem under the assumption that only the AIR constraint binds. When this is the case, $t=-8s^2+U(2)-2$ must hold at the optimum, so the optimization problem in (OA 1) reduces to

$$\max_{s\in[0,1]} 16s(1-s)+8s^2+U(2)-2.$$

The objective function is strictly increasing in s, and so has a unique maximizer of $s^*=1$ (which gives a corresponding value of $t^*=-10+U(2)$), from which we obtain a type 4 firm payoff of $10-U(2)$. This solution satisfies the PIC constraint if and only if $10-U(2)\leq U(2)$, which is equivalent to $U(2)\geq 5$. As observed above, the type 2 firm can never achieve a payoff strictly above this threshold, so we conclude that $U(2)=5$ is the unique contracting equilibrium payoff of type 2 at which only the AIR constraint

binds, and the corresponding maximum payoff that the type 4 firm can obtain is also $U(4) = 5$.

Now we solve (OA 1) under the assumption that the AIR constraint does not bind. When this is the case, $t = 8s(1-s) - U(2)$ must hold at the optimum, so the optimization problem in (OA 1) reduces to

$$\max_{s \in [0,1]} U(2) + 8s(1-s).$$

The objective function is single-peaked with a unique maximizer of $s^* = 1/2$ (which gives a corresponding value of $t^* = 2 - U(2)$), from which we obtain a type 4 firm payoff of $U(2) + 2$.

We determine the values of $U(2)$ for which this solution actually constitutes the optimum. Given $s = 1/2$ and $t = 2 - U(2)$, the agent's expected utility is $3 - U(2)$. Thus, the AIR constraint is satisfied when $U(2) \leq 3$.

We thus have that type 2 payoffs of $U(2) \in [2, 3]$ are possible in contracting equilibrium, and the corresponding maximum payoff of the type 4 firm is $U(2) + 2$.

Now we solve (OA 1) for $U(2) \in [3, 5]$. We have established that here both the AIR and PIC constraints must bind at the optimum. Setting the AIR and PIC inequalities to be equalities and then solving for t and $U(2)$ gives

$$\begin{aligned} t &= -4s^2 + 4s(1-s) - 1, \\ U(2) &= 4s^2 + 8s(1-s) + 1. \end{aligned}$$

Consequently, the payoffs of the two firm types, as parametrized by $s \in (1/2, 1)$ are

$$\begin{aligned} U(2) &= -4s^2 + 8s + 1, \\ U(4) &= -8s^2 + 12s + 1. \end{aligned}$$

Solving for $U(4)$ in terms of $U(2)$ then gives $U(4) = -U(2)^2/2 + 4U(2) - 5/2$. ■

OA.1.2 Upper Envelope with Explicit Contracts

Proposition OA 2. *The upper envelope of the firm payoffs sustainable in contracting equilibrium with explicit contracts is*

$$\Big\{(U(2),U(4)) \in \mathbb{R}^2 : (U(2),U(4)) = (2,4),\ 2 < U(2) \leq \frac{21}{8} \text{ and } U(4) = U(2) + \frac{3}{2},\ \text{or}$$
$$\frac{21}{8} \leq U(2) \leq \frac{9}{2} \text{ and } U(4) = 5U(2) - 6\sqrt{36 - 6U(2)} - 36\Big\}.$$

Lemma OA 3. *In an equilibrium with payoff $U(2) > 2$ to the type 2 firm, the payoff of the type 4 firm is bounded from above by*

$$\begin{aligned} &\max_{(s,t)\in[0,1]\times\mathbb{R}} 12s(1-s) - t \\ s.t.\ &\textit{AIR: } t + \frac{9}{2}s^2 \geq 0, \\ &\textit{PIC: } 6s(1-s) - t = U(2). \end{aligned} \tag{OA 2}$$

Proof. For the type 2 firm to obtain a payoff strictly higher than 2 with explicit contracts, the type 2 firm must pool with the type 4 firm with probability 1. There would then need to be (s,t) and $\tilde{\lambda}(2) \in [1/2, 1]$ such that both types of the firm obtain their equilibrium payoff from (s,t) when the employee responds with some play consistent with a posterior belief putting probability $\tilde{\lambda}(2)$ on $\theta = 2$. Since the type 2 firm obtains a strictly higher payoff, the employee must accept the proposal with strictly positive probability α, so $U(2) = \alpha(4(2-\tilde{\lambda}(2))s(1-s) - t)$, $U(4) = \alpha(8(2-\tilde{\lambda}(2))s(1-s) - t)$, and $t + (2-\tilde{\lambda}(2))^2 s^2/2 \geq 0$. Thus, the payoff of the type 4 firm is bounded from above

by

$$\max_{(s,t,\tilde{\lambda}(2),\alpha)\in[0,1]\times\mathbb{R}\times\left[\frac{1}{2},1\right]\times(0,1]} \alpha(8(2-\tilde{\lambda}(2))s(1-s)-t)$$

$$\text{s.t. } \tilde{\lambda}(2)\in\left[\frac{1}{2},1\right],$$

$$\text{AIR: } t+\frac{1}{2}(2-\tilde{\lambda}(2))^2s^2\geq 0,$$

$$\text{PIC: } \alpha(4(2-\tilde{\lambda}(2))s(1-s)-t)=U(2).$$

Observe that, for any $\tilde{\lambda}>1/2$, decreasing $\tilde{\lambda}$ to $1/2$ and increasing t by $4(\tilde{\lambda}(2)-1/2)s(1-s)$ preserves the AIR constraint, keeps the PIC constraint satisfied, and weakly increases the payoff of the type 4 firm. A similar shift can be done for any $\alpha<1$. Thus, the optimum must be attained with $\tilde{\lambda}(2)=1/2$ and $\alpha=1$. Substituting these values into the constrained optimization problem and deleting the belief constraint results in (OA 2). ■

Lemma OA 4. *With explicit contracts, the payoff of the type* 2 *firm in a contracting equilibrium can never be more than* $9/2$.

Proof. As established in the proof of Lemma OA 3, there must be some (s,t), $\tilde{\lambda}(2)\in[1/2,1]$, and $\alpha\in[0,1]$ such that $U(2)=\alpha(4(2-\tilde{\lambda}(2))s(1-s)-t)$ and $t+(2-\tilde{\lambda}(2))^2s^2/2\geq 0$. Thus, we have $U(2)\leq\alpha(4(2-\tilde{\lambda}(2))s(1-s)+(2-\tilde{\lambda}(2))^2s^2/2)$. Standard computations show that $\max_{(s,t,\tilde{\lambda}(2),\alpha)}\alpha(4(2-\tilde{\lambda}(2))s(1-s)+(2-\tilde{\lambda}(2))^2s^2/2)=9/2$. ■

Proof of Proposition OA 2. The same argument as in the proof of Proposition OA 1 shows that the type 2 firm can never get a lower payoff than 2 in a contracting equilibrium. Moreover, Proposition OA 1 established that 4 is the maximum payoff that the type 4 firm can get in equilibrium with flexible contracts when the type 2 firm receives a payoff of 2. This is also true when only explicit contracts. The reason is that because it is the maximum payoff of the type 4 firm with flexible contracts, 4 provides an upper bound for the payoff of the type 4 firm with explicit contracts, and this upper bound is attained at the least-cost separating outcome.

We now turn our attention to when the type 2 firm receives a higher equilibrium payoff than 2. We first solve (OA 2) under the assumption that the AIR constraint does not bind. When this is the case, $t = 6s(1-s) - U(2)$ must hold at the optimum, so the optimization problem in (OA 1) reduces to

$$\max_{s\in[0,1]} U(2) + 6s(1-s).$$

The objective function is single-peaked with a unique maximizer of $s^* = 1/2$ (which gives a corresponding value of $t^* = 3/2 - U(2)$), from which we obtain a type 4 firm payoff of $U(2) + 3/2$.

We determine the values of $U(2)$ for which this solution actually constitutes the optimum. Given $s = 1/2$ and $t = 3/2 - U(2)$, the agent's expected utility is $21/8 - U(2)$. Thus, the AIR constraint is satisfied when $U(2) \leq 21/8$.

We thus have that type 2 payoffs of $U(2) \in (2, 21/8]$ are possible in contracting equilibrium, and the corresponding maximum payoff of the type 4 firm is $U(2) + 3/2$.

Now we solve (OA 1) for $U(2) \in [21/8, 9/2]$. We have established that here the AIR constraint must bind in must bind at the optimum. Setting the AIR to be an equality, combining this with the PIC equality, and then solving for t and $U(2)$ gives

$$\begin{aligned} t &= -\frac{9}{2}s^2, \\ U(2) &= \frac{9}{2}s^2 + 6s(1-s). \end{aligned}$$

Consequently, the payoffs of the two firm types, as parametrized by $s \in (1/2, 1)$ are

$$\begin{aligned} U(2) &= -\frac{3}{2}s^2 + 6s, \\ U(4) &= -\frac{15}{2}s^2 + 12s. \end{aligned}$$

Solving for $U(4)$ in terms of $U(2)$ then gives $U(4) = 5U(2) - 6\sqrt{36 - 6U(2)} - 36$. ■

OA.1.3 Plausible Payoffs with Flexible Contracts

Proposition OA 3. *The set of firm payoffs sustainable in payoff-plausible contracting equilibria with flexible contracts is*

$$\big\{(U(2),U(4)) \in \mathbb{R}^2 : 2 \leq U(2) \leq 3, U(4) = U(2)+2\big\}.$$

Proof. We first establish that no equilibrium where the payoffs are below the upper envelope characterized in Proposition OA 1 is payoff-plausible. This is because, in any contracting equilibrium, the worker's expected utility conditional on 2 is weakly negative, so the AIR constraint in (OA 1) implies the AIR constraint for the type 4 payoff-benchmark problem in (1). Thus, given a payoff-plausible equilibrium where the type 2 firm obtains a payoff of $U(2)$, the payoff of the type 4 firm must exceed that given in (OA 1).

We now argue that none of the payoff profiles in the upper envelope with $U(2) > 3$ are plausible. Fix such a payoff profile, and note that it corresponds to an outcome in which, conditional on 4, the profit share is $s > 1/2$ and the worker's expected utility is strictly positive. If this outcome were payoff-plausible, then the payoff of the type 4 firm would exceed that obtained from

$$\begin{aligned} &\max_{(s,t)\in[0,1]\times\mathbb{R}} 16s(1-s) - t \\ \text{s.t. } &\text{AIR: } 8s^2 + t \geq 0, \\ &\text{PIC: } 8s(1-s) - t \leq U(2). \end{aligned}$$

This alters the optimization problem in (OA 1) so that the AIR constraint only requires $8s^2 + t \geq 0$, i.e. that the worker's expected utility conditional on 4 be weakly positive. Since the worker's expected utility conditional on 4 is strictly positive in the outcome being considered, this relaxed AIR constraint is slack. However, this relaxed AIR constraint cannot be slack at an optimal solution of $s > 1/2$, for essentially the same

reason that the true AIR constraint in (OA 1) cannot be slack at an optimal solution of $s > 1/2$. Thus, the payoff of the type 4 firm does not exceed the required benchmark for plausibility.

We conclude by showing that every payoff profile in the upper envelope with $U(2) \leq 3$ is plausible. Each of these payoffs can be attained by taking the principal-optimal safe outcome and shifting the expected transfers given by both firm types down by the same amount. The resulting outcome necessarily satisfies the payoff-plausibility bounds for each firm type. Otherwise, the firm type whose payoff-plausibility bound exceeded their payoff from this outcome could obtain a safe payoff equal to the payoff-plausibility bound minus the transfer shift, which would exceed their optimal safe payoff and thus result in a contradiction. ■

OA.2 Proof of Proposition 1

Proposition 1. *In both the general-mechanism and deterministic-mechanism game, principal-optimal safe outcomes exist.*

Here we give the proof for the general-mechanism game. The proof for the deterministic-mechanism game is similar.

Proof of Proposition 1 for the General-Mechanism Game. Let $\mathcal{M}_{safe}$ denote the set of safe mechanisms. Throughout the proof, we identify every direct mechanism with the corresponding collection of allocations $\{q(\theta)\}_{\theta\in\Theta}$ induced for each type. Moreover, for each principal type $\theta \in \Theta$ we let $U(\theta, q) \equiv \mathbb{E}_q[U(\theta, x, y)]$ and $V(\theta, q) \equiv \mathbb{E}_q[V(\theta, x, y)]$ denote the expected utility of the principal and the agent, respectively, from allocation q when the principal's type is θ. We first note that $\mathcal{M}_{safe}$ is non-empty since every direct mechanism in which each principal type commits to x_o, i.e. $q(\theta)[x_o] = 1$ for all $\theta \in \Theta$, is safe.

We argue that $\mathcal{M}_{safe}$ is a sequentially compact space. Let $\{\{q_j(\theta)\}_{\theta\in\Theta}\}_{j\in\mathbb{N}}$ be an arbitrary sequence of safe mechanisms. Since $\Delta(X \times Y)$ is itself sequentially com-

pact, it follows that $\{\{q_j(\theta)\}_{\theta\in\Theta}\}_{j\in\mathbb{N}}$ has a limit point. Let $\{q^*(\theta)\}_{\theta\in\Theta}$ denote such a limit point, and suppose without loss of generality (by restricting attention to a convergent subsequence if necessary) that $\lim_{j\to\infty} q_j(\theta) = q^*(\theta)$ for all $\theta \in \Theta$. Since $U(\theta, q_j(\theta)) \geq \max\{\max_{\theta'\in\Theta} U(\theta, q_j(\theta')), 0\}$ for all $j \in \mathbb{N}$ and $\theta \in \Theta$, continuity implies that $U(\theta, q^*(\theta)) \geq \max\{\max_{\theta'\in\Theta} U(\theta, q^*(\theta')), 0\}$ for all $\theta \in \Theta$. For identical reasons, $V(\theta, q(\theta)) \geq 0$ also holds for all $\theta \in \Theta$. To conclude that $\{q^*(\theta)\}_{\theta\in\Theta}$ is an safe mechanism, all that remains is to show that $\mathbb{P}_{q^*(\theta)}[y \in \arg\max_{y'\in Y} V(\theta, x, y')] = 1$ for all $\theta \in \Theta$. Suppose otherwise that $\mathbb{P}_{q^*(\theta)}[y \in \arg\max_{y'\in Y} V(\theta, x, y')] < 1$ for some θ. Then there is some closed set $\widetilde{X} \subseteq X$ and agent action $\tilde{y} \in Y$ such that $\mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}}(x)V(\theta, x, \tilde{y})] > \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}}(x)V(\theta, x, y)]$. For every $\varepsilon > 0$, let $\widetilde{X}_{<\varepsilon} = \{x \in X : d(x, \widetilde{X}) < \varepsilon\}$, $\widetilde{X}_{=\varepsilon} = \{x \in X : d(x, \widetilde{X}) = \varepsilon\}$, and $\widetilde{X}_{>\varepsilon} = \{x \in X : d(x, \widetilde{X}) > \varepsilon\}$. Additionally, let $\underline{V}(\theta) = \min_{(x,y)\in X\times Y} V(\theta, x, y)$. By continuity, there exists some $\tilde{\varepsilon} > 0$ such that

$$\begin{aligned}&\mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}}(x)(V(\theta, x, \tilde{y}) - \underline{V}(\theta))] + \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}_{>\tilde{\varepsilon}}}(x)(V(\theta, x, y) - \underline{V}(\theta))]\\ >&\mathbb{E}_{q^*(\theta)}[V(\theta, x, y)] - \underline{V}(\theta).\end{aligned}$$

As $\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}}(x)(V(\theta, x, \tilde{y}) - \underline{V}(\theta))$ is a lower semicontinuous function of $x \in X$, it follows that $\liminf_{j\to\infty} \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}}(x)(V(\theta, x, \tilde{y}) - \underline{V}(\theta))] \geq \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}}(x)(V(\theta, x, \tilde{y}) - \underline{V}(\theta))]$. Likewise, $\mathbb{1}_{\widetilde{X}_{>\tilde{\varepsilon}}}(x)(V(\theta, x, y) - \underline{V}(\theta))$ is a lower semicontinuous function of $(x, y) \in X \times Y$, so $\liminf_{j\to\infty} \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}_{>\tilde{\varepsilon}}}(x)(V(\theta, x, y) - \underline{V}(\theta))] \geq \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\widetilde{X}_{>\tilde{\varepsilon}}}(x)(V(\theta, x, y) - \underline{V}(\theta))]$. Consequently, for sufficiently high $j \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{E}_{q_j(\theta)}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}}(x)(V(\theta, x, \tilde{y}) - \underline{V}(\theta))] + \mathbb{E}_{q_j(\theta)}[\mathbb{1}_{\widetilde{X}_{>\tilde{\varepsilon}}}(x)(V(\theta, x, y) - \underline{V}(\theta))]\\ >&\mathbb{E}_{q_j(\theta)}[V(\theta, x, y)] - \underline{V}(\theta).\end{aligned}$$

This implies that $\mathbb{E}_{q_j(\theta)}[\mathbb{1}_{\widetilde{X}_{<\tilde{\varepsilon}}}(x)V(\theta, x, \tilde{y})] + \mathbb{E}_{q_j(\theta)}[\mathbb{1}_{\widetilde{X}_{=\tilde{\varepsilon}}\cup\widetilde{X}_{>\tilde{\varepsilon}}}(x)V(\theta, x, y)] > \mathbb{E}_{q_j(\theta)}[V(\theta, x, y)]$, which contradicts $q_j(\theta)$ being safe.

Let $\overline{U}_{safe}(\theta) \equiv \sup_{\{q(\theta')\}_{\theta'\in\Theta}\in\mathcal{M}_{safe}} U(\theta, q(\theta))$ denote the supremum of the type θ principal's payoff over all safe mechanisms. Let $\{\{q_{j,\theta}(\theta')\}_{\theta'\in\Theta}\}_{j\in\mathbb{N}}$ be a sequence

of safe mechanisms that converges to the safe mechanism $\{q^*_\theta(\theta')\}_{\theta' \in \Theta}$ and attains $\overline{U}_{safe}(\theta)$ for the type θ principal: that is, $\lim_{j\to\infty} U(\theta, q_j(\theta)) = \overline{U}_{safe}(\theta)$. By continuity, $U(\theta, q^*_\theta(\theta)) = \overline{U}_{safe}(\theta)$. Consider the direct mechanism given by $\{q^*_\theta(\theta)\}_{\theta\in\Theta}$. By construction, this mechanism satisfies the agent's incentive compatibility requirements. Moreover, $U(\theta, q^*_\theta(\theta)) = \overline{U}_{safe}(\theta) \geq U(\theta, q^*_{\theta'}(\theta'))$ for all $\theta, \theta' \in \Theta$. Thus, this mechanism is safe and attains each principal type's highest possible payoff over all safe mechanisms. ■

OA.3 Proof of Proposition 2

Proposition 2. *In MCS environments, the conditional distributions of the principal-optimal safe outcomes $\{q^*(\theta)\}_{\theta\in\Theta}$ in the general-mechanism game are characterized inductively by*

$$
\begin{aligned}
q^*(\theta_n) \in \underset{q\in\Delta(X\times T\times Y)}{\arg\max}\ & \mathbb{E}_q[u(\theta_n, x, y) - t] \\
s.t.\ \ & \textit{AIC: } \mathbb{P}_q[y = y^*(\theta_n, x)|x \neq x_o] = 1, \\
& \textit{AIR: } \mathbb{E}_q[v(\theta_n, x, y) + g(t)] \geq 0, \\
& \textit{PIC: } \mathbb{E}_q[u(\theta_{n'}, x, y) - t] \leq \mathbb{E}_{q^*(\theta_{n'})}[u(\theta_{n'}, x, y) - t]\ \forall n' < n,
\end{aligned}
$$

for all $n \in \{1, ..., N\}$. Moreover, the same inductive characterization holds for the deterministic-mechanism game when the PIC constraint is strengthened to $\mathbb{P}_q[u(\theta_{n'}, x, y) - t \leq U(\theta'_n, q^*(\theta_{n'}))] = 1$ for all $n' < n$.

Lemma OA 5. *In MCS environments, the conditional distributions of the principal-optimal safe outcomes $\{q^*(\theta)\}_{\theta\in\Theta}$ in the general-mechanism game are characterized*

inductively by

$$
\begin{aligned}
q^*(\theta_n) \in \underset{q \in \Delta(X \times T \times Y)}{\arg\max} \; & \mathbb{E}_q[u(\theta_n, x, y) - t] \\
s.t. \;\; & \textit{AIC: } \mathbb{P}_q[y = y^*(\theta_n, x) | x \neq x_o] = 1, \\
& \textit{AIR: } \mathbb{E}_q[v(\theta_n, x, y) + g(t)] \geq 0, \\
& \textit{PIC: } \mathbb{E}_q[u(\theta_n, x, y) - t] \leq \mathbb{E}_{q^*(\theta_{n'})}[u(\theta_{n'}, x, y) - t] \; \forall n' < n,
\end{aligned}
\tag{OA 3}
$$

for all $n \in \{1, ..., N\}$.

Proof. The conditional distributions of a principal-optimal safe outcome solve the constraints given in (OA 3). Thus, any conditional distributions which solve the problem necessarily result in a weakly higher payoff to the corresponding principal type than their principal-optimal safe payoff.

To complete the proof, we show that every outcome whose conditional distribution for every type is a solution to the problem in (OA 3) is safe. Fix such an outcome, and, for each $\theta \in \Theta$, let $q^*(\theta)$ the corresponding conditional distribution. The agent incentive compatibility and individual rationality constraints are satisfied by definition. So all that remains is to check that principal incentive compatibility holds. Consider a principal type θ_n. By construction, every type $\theta_{n'}$ with $n' < n$ (weakly) prefers their conditional distribution $q^*(\theta_{n'})$ to $q^*(\theta_n)$. Therefore, we need only consider whether some type $\theta_{n'}$ with $n' > n$ would prefer the conditional distribution $q^*(\theta_n)$ than $q^*(\theta_{n'})$. Suppose that there is such a type and that $\theta_{n'}$ is the smallest type for which this is true. Consider now the distribution $\tilde{q}(\theta_{n'}) \in \Delta(X \times T \times Y)$ that is obtained from $q^*(\theta_n)$ by setting $y = y^*(\theta_{n'}, x)$ whenever $x \neq x_o$ and shifting every t to $t + \mathbb{E}_{q^*(\theta_n)}[u(\theta_{n'}, x, y^*(\theta_{n'}, x)) - u(\theta_{n'}, x, y^*(\theta_n, x))]$. This conditional distribution gives $\theta_{n'}$ the same expected utility as $q^*(\theta_n)$, and, by supermodularity and the fact that $y^*(\theta_{n'}, x) > y^*(\theta_n, x)$ for all $x \neq x_o$, satisfies the corresponding constraints in (OA 3). This means that $\theta_{n'}$ must obtain a payoff from $q^*(\theta_{n'})$ that is weakly higher than the payoff they obtain from $q^*(\theta_n)$, which is a contradiction. ■

Lemma OA 6. *In MCS environments, the conditional distributions of the principal-optimal safe outcomes* $\{q^*(\theta)\}$ *in the deterministic-mechanism game are characterized inductively by*

$$\begin{aligned} q^*(\theta_n) = & \underset{q\in\Delta(X\times T\times Y)}{\arg\max} \ \mathbb{E}_q[u(\theta_n, x, y^*(\theta_n, x)) - t] \\ & s.t. \ \ AIC\text{:}\ \mathbb{P}_q[y = y^*(\theta_n, x)|x \neq x_o] = 1, \\ & \quad AIR\text{:}\ \mathbb{E}_q[v(\theta_n, x, y) + g(t)] \geq 0, \\ & \quad PIC\text{:}\ \mathbb{P}_q[u(\theta_{n'}, x, y) - t \leq U(\theta'_n, q^*(\theta_{n'}))] = 1 \ \forall n' < n, \end{aligned} \tag{OA 4}$$

for all $n \in \{1, ..., N\}$.

Proof. A similar argument to those in the proof of Proposition 2 shows that any of these outcomes are safe. Since the conditional distributions of any principal-optimal safe outcome solve the constraints given in (OA 4), we conclude that the conditional distributions identified by (OA 4) do in fact characterize the principal-optimal safe outcomes. ■

OA.4 Proof of Proposition 4

Proposition 4. *For each* $\theta \in \Theta$*, let* $x_\theta^{CI} \in X$ *be the principal action in the complete-information benchmark when the principal's type is known to be* θ*. Suppose the environment is MCS and that the ex-ante mechanism design benchmarks have the same actions as the complete-information benchmark but different expected transfers for at least one principal type. If, for each* $\theta \in \Theta$*, there is a sequence* $\{x_i\}$ *converging to* x_θ^{CI} *such that* $u(\theta, x_i, y^*(\theta, x_i)) - u(\theta, x_\theta^{CI}, y^*(\theta, x_\theta^{CI})) > u(\theta', x_i, y^*(\theta, x_i)) - u(\theta', x_\theta^{CI}, y^*(\theta, x_\theta^{CI}))$ *for all* $\theta' < \theta$ *and* i*, then the ex-ante mechanism design benchmarks are not payoff-plausible.*

Proof. Fix an ex-ante mechanism design benchmark. The complete-information benchmark for each type θ gives the agent a utility of exactly 0. Combining this with the fact that the agent's expected utility in any individually rational outcome must be non-

negative, it follows that, in the ex-ante mechanism design benchmark, there must be at least one type, say $\overline{\theta}$, that gives the agent a strictly positive expected utility. Let $t_{\overline{\theta}}$ be the expected transfer played by $\overline{\theta}$ in the ex-ante mechanism design benchmark. For each $i \in \mathbb{N}$, consider the transfer given by $t_i = t_{\overline{\theta}} + u(\overline{\theta}, x_i, y^*(\overline{\theta}, x_i)) - u(\overline{\theta}, x_{\overline{\theta}}^{CI}, y^*(\overline{\theta}, x_{\overline{\theta}}^{CI}))$. By construction, the type $\overline{\theta}$ would obtain the same payoff from (x_i, t_i) and the agent responding with $y^*(\overline{\theta}, x_i)$ as in the ex-ante mechanism design benchmark, while all lower types would obtain a strictly lower payoff. Thus, for all sufficiently large i, there is a small but strictly positive ε such that $(x_i, t_i - \varepsilon)$ and the agent responding with $y^*(\overline{\theta}, x_i)$ gives the type $\overline{\theta}$ a strictly higher payoff than the ex-ante mechanism design benchmark, all types lower than $\overline{\theta}$ a strictly lower payoff than the ex-ante mechanism design benchmark, and the agent a strictly positive expected utility when the type is $\overline{\theta}$, which means that the payoff of the type $\overline{\theta}$ does not meet their plausibility threshold. ■

OA.5 Proof of Proposition 6

The following is a generalization of Proposition 6 that implies that payoff-plausibility selects the principal-optimal safe outcomes in the deterministic-mechanism game of the doubly supermodular firm and employee example.

Proposition 4′. *Suppose the environment is MCS with definite gains and that, for every $\tilde{\lambda} \in \Delta(\Theta)$ and $x \neq x_o$, either quasi-strictness holds at x, or there exists a sequence $\{x_i\}$ converging to x such that $y^*(\tilde{\lambda}, x_i)$ converges to $y^*(\tilde{\lambda}, x)$, quasi-strictness holds at each x_i, and either one of the following conditions hold:*

1. *(a) $u(\theta, x, y^*(\tilde{\lambda}, x))$ is constant in θ.*
   (b) $u(\theta, x_i, y^*(\tilde{\lambda}, x_i)) > u(\theta, x, y^*(\tilde{\lambda}, x))$ for all i.
   (c) $v(\theta, x_i, y^*(\tilde{\lambda}, x_i)) > v(\theta, x, y^*(\tilde{\lambda}, x))$ for all i.
2. *(a) $u(\theta, x, y^*(\tilde{\lambda}, x))$ is constant in θ.*
   (b) $v(\theta, x, y^*(\tilde{\lambda}, x))$ is strictly increasing in θ.

Then payoff-plausibility selects the principal-optimal safe outcomes in the deterministic-

mechanism game.

Proof. We first show that every payoff-plausible contracting equilibrium outcome must be always-accepting. Suppose that p is a contracting equilibrium that is not always-accepting. Then there is some $\overline{\theta} \in \Theta$, $x \in X$, $t \in T$, $\tilde{\lambda} \in \Delta(\Theta)$, and $\alpha \in (0,1)$ such that (1) $\alpha(u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) - t) = U(\overline{\theta}, p)$, (2) $\alpha(u(\theta, x, y) - t) \leq U(\theta, p)$ for all $\theta \neq \overline{\theta}$, (3) $\tilde{\lambda}$ is weakly below $\delta_{\overline{\theta}}$ under FOSD, and (4) $v(\overline{\theta}, x, y^*(\overline{\theta}, x)) + g(t) \geq 0$. Consider (x', t') such that $t' = \alpha t + u(\overline{\theta}, x', y^*(\overline{\theta}, x')) - \alpha u(\overline{\theta}, x, y^*(\tilde{\lambda}, x))$. By construction, this (x', t') is such that, when the agent responds with $y^*(\overline{\theta}, x')$, the type $\overline{\theta}$ principal obtains the same payoff as in p. Moreover, we can take $x' > x$ to be close enough to x so that all lower type principals would achieve a strictly lower payoff from $(x', t', y^*(\overline{\theta}, x'))$ than p and the agent gets a strictly higher utility from $\overline{\theta}$ playing (x', t') than their outside option. Thus, for sufficiently small $\varepsilon > 0$, $(x', t' - \varepsilon)$ would satisfy the constraints of the type $\overline{\theta}$ optimization problem in (2) and give $\overline{\theta}$ a strictly higher payoff than in p, so p cannot be payoff-plausible.

We now show that $\mathbb{P}[v(\theta, x, y) + g(t) \leq 0] = 1$ in any payoff-plausible outcome. Suppose towards a contradiction that there is some θ such that $\mathbb{P}[v(\theta, x, y) + g(t) > 0 | \theta] > 0$, and suppose that $\overline{\theta}$ is the highest type for which this is true. Then there is some $x \in X$, $t \in T$, and $\tilde{\lambda} \in \Delta(\Theta)$ such that (1) $u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) - t = U(\overline{\theta}, p)$, (2) $u(\theta, x, y) - t \leq U(\theta, p)$ for all $\theta \neq \overline{\theta}$, (3) $\tilde{\lambda}$ is weakly below $\delta_{\overline{\theta}}$ under FOSD, and (4) $v(\overline{\theta}, x, y^*(\overline{\theta}, x)) + g(t) > 0$. Consider (x', t') such that $t' = t + u(\overline{\theta}, x', y^*(\overline{\theta}, x')) - u(\overline{\theta}, x, y^*(\tilde{\lambda}, x))$. By construction, this (x', t') is such that, when the agent responds with $y^*(\overline{\theta}, x')$, the type $\overline{\theta}$ principal obtains the same payoff as in p. Moreover, we can take $x' > x$ to be close enough to x so that all lower type principals would achieve a strictly lower payoff from $(x', t', y^*(\overline{\theta}, x'))$ than p and the agent gets a strictly higher utility from $\overline{\theta}$ playing (x', t') than their outside option. Thus, for sufficiently small $\varepsilon > 0$, $(x', t' - \varepsilon)$ would satisfy the constraints of the type $\overline{\theta}$ optimization problem in (2) and give $\overline{\theta}$ a strictly higher payoff than in p, which contradicts payoff-plausibility.

Since the agent's total expected utility must be weakly positive, it thus follows

that $\mathbb{P}[v(\theta, x, y) + t = 0] = 1$ must hold in any payoff-plausible contracting equilibrium outcome p. Since the agent's utility is strictly increasing in the principal's type, this means that there can be no pooling between different principal types, so any outcome that is payoff-plausible must be safe. As every payoff-plausible outcome must principal-payoff-dominate the principal-optimal safe outcome, it thus follows that only the principal-optimal safe outcomes can be payoff-plausible. ■

OA.6 Omitted Example

We show by example that the correspondence mapping mechanisms into sequential continuation equilibria is not necessarily upper hemicontinuous.

Example OA 1. Suppose that $\Theta = \{-1, 1\}$, $X = [-1, 1]^2$, $Y = R = [-1, 1]$, $U(\theta, x_1, x_2, y) = \theta y - x_2$, and $V(\theta, x_1, x_2, y) = x_1 y - \alpha|x_1| + x_2$ for some $\alpha \in (1, 3/2)$. Consider the sequence of mechanisms (μ_j, M_P) indexed by $j \in \mathbb{N}$, where $M_P = \{m_{P,1}, m_{P,2}, m_{P,3}, m_{P,4}\}$ and

$$\mu_j(m_P) = \begin{cases} \delta_{((\frac{1}{j+1},0),0)} & \text{if } m_P = m_{P,1} \\ \delta_{((-\frac{1}{j+1},0),0)} & \text{if } m_P = m_{P,2} \\ \delta_{((1,\frac{1}{2}),1)} & \text{if } m_P = m_{P,3} \\ \delta_{((-1,\frac{1}{2}),-1)} & \text{if } m_P = m_{P,4} \end{cases}.$$

As $j \to \infty$, this sequence of mechanisms converges to the mechanism (μ, M_P) given by

$$\mu(m_P) = \begin{cases} \delta_{((0,0),0)} & \text{if } m_P \in \{m_{P,1}, m_{P,2}\} \\ \delta_{((1,\frac{1}{2}),1)} & \text{if } m_P = m_{P,3} \\ \delta_{((-1,\frac{1}{2}),-1)} & \text{if } m_P = m_{P,4} \end{cases}.$$

The unique sequential continuation equilibrium after mechanism (μ_j, M_P) is accepted has the type 1 principal playing $m_{P,1}$, the type -1 principal playing $m_{P,2}$, and the agent responding with $y = 1$ to any positive x_1 and with $y = -1$ to any negative

x_1. Consequently, the agent's expected utility from accepting this mechanism is strictly negative, so mechanism (μ_j, M_P) must be rejected in any contracting equilibrium.

In every sequential continuation equilibrium where mechanism (μ, M_P) is accepted, either the type 1 principal plays $m_{P,3}$ or the type -1 principal plays $m_{P,4}$. Thus the agent's expected utility from accepting this mechanism is weakly positive conditional on either principal type and strictly positive conditional on at least one type, so it must be accepted in any contracting equilibrium. □

OA.7 Proof of Lemma 2

Lemma 2. *Consider a sequence of primitives $\{\mathcal{P}_j\}_{j\in\mathbb{N}}$ that converges to the original primitives $\mathcal{P}$. For every mechanism $(\mu, M_P) \in \mathcal{M}$, there is a sequence of mechanisms $(\mu_j, M_P) \in \mathcal{M}_j$ such that any limit of sequential continuation equilibrium outcomes after these mechanisms are proposed is a sequential continuation equilibrium outcome after (μ, M_P) is proposed.*

Construction of Mechanism. Let $\nu = \sum_{m_P} \mu(m_P)/|M_P|$ be the distribution over principal action-recommendation pairs that is obtained by drawing (x, r) from $\mu(m_P)$ with probability $1/|M_P|$ uniform over each (m_P). Let $f_{m_P} : \cup_{m'_P}\text{supp}(\mu(m'_P)) \to [0, |M_P|]$ be the Radon-Nikodym derivative of the $\mu(m_P)$ distribution with respect to ν. Note that $\sum_{m_P} f_{(m_P)}(x, r)/|M_P| = 1$ for all $(x, r) \in \cup_{m'_P}\text{supp}(\mu(m'_P))$.

Let $P_+(M_P) = P(M_P) \setminus \{\emptyset\}$ be the set of non-empty subsets of M_P. For a given $(x, r) \in \cup_{m'_P}\text{supp}(\mu(m'_P))$, let $\mathbf{M}(x, r) = \{(m_P) \in M_P : f_{(m_P)}(x, r) > 0\}$ be the set of principal messages for which the corresponding distribution over principal action-recommendation pairs has a strictly positive Radon-Nikodym derivative at (x, r).

We enlarge the principal recommendation space so that, in addition to some $r \in R$, each recommendation includes a non-empty subset of M_P. Formally, the enlarged recommendation space corresponds to $\widetilde{R} = R \times P_+(M_P)$. The modified mechanism is $(\tilde{\mu}, M_P)$, where the principal message space is the same as in (μ, M_P), and the

$\tilde{\mu}$ is induced from μ by replacing each principal-action recommendation pair $(x, r) \in \cup_{m'_P} \text{supp}(\mu(m'_P))$ with $(x, (r, \mathbf{M}(x, r)))$. Let $\tilde{\nu} = \sum_{m_P} \tilde{\mu}(m_P)/|M_P|$ and $\tilde{f}_{(m_P)} : \cup_{m'_P} \text{supp}(\tilde{\mu}(m'_P)) \to [0, |M_P|]$ be the Radon-Nikodym derivative of the $\tilde{\mu}(m_P)$ distribution with respect to $\tilde{\nu}$. Then, by construction, $\tilde{f}_{(m_P)}(x, (r, \mathbf{M}(x, r))) = f_{(m_P)}(x, r)$ for all $m_P \in M_P$ and $(x, r) \in X \times R$, while $\tilde{f}_{m_P}(x, (r, \mathrm{M})) = 0$ for all $m_P \in M_P$, $(x, r) \in X \times R$, and $\mathrm{M} \neq \mathbf{M}(x, r)$. Moreover, both (μ, M_P) and $(\tilde{\mu}, M_P)$ have the same sequential continuation equilibrium outcomes. ■

Construction of Sequences of Mechanisms. Suppose, by restricting attention to a subsequence if necessary, that there is a finite $\widetilde{X}_j \subseteq X_j$ such that, for all $j \in \mathbb{N}$ and $x \in X$, there is an $x' \in \widetilde{X}_j$ satisfying $|x - x'| \leq 1/j$. Additionally, suppose that there is some finite $\widetilde{R}_j \subset R$ such that $|\widetilde{R}_j| \leq |R_j|/2^{|M_P|}$ and, for all $r \in R$, there is an $r' \in \widetilde{R}_j$ satisfying $|r - r'| \leq 1/j$. The requirement on the relative sizes of $\widetilde{R}_j$ and R_j means that, for each $r' \in \widetilde{R}_j$ and $\mathrm{M}' \in P_+(M_P)$, we can identify (r', M') with some element of R_j. Consider the mechanism $(\tilde{\mu}_j, M_P)$, where $\tilde{\mu}_j$ is determined as follows. For each $x \in \widetilde{X}_j$, $r \in \widetilde{R}_j$, and $\mathrm{M} \in P_+(M_P)$, let

$$\tilde{\mu}_j(m_P)[x', (r', \mathrm{M}')] = \mathbb{E}_{(x,r)\sim\mu(m_P)} \left[\frac{\mathbb{1}(\mathbf{M}(x, r) = \mathrm{M}') \mathbb{1}\left(|x - x'| \leq \frac{1}{j}\right) \mathbb{1}\left(|r - r'| \leq \frac{1}{j}\right)}{\sum_{(x'', r'') \in \widetilde{X}_j \times \widetilde{R}_j} \mathbb{1}\left(|x - x''| \leq \frac{1}{j}\right) \mathbb{1}\left(|r - r''| \leq \frac{1}{j}\right)} \right]$$

and $\tilde{\mu}_j(m_P)[x', (r', \mathrm{M}'')] = 0$ for all $\mathrm{M}'' \neq \mathrm{M}'$. By construction, $\tilde{\mu}_j(m_P)[x', (r', \mathrm{M}')] \geq 0$ for all $x \in \widetilde{X}_j$, $r \in \widetilde{R}_j$, and $\mathrm{M} \in P_+(M_P)$, and $\sum_{x', r', \mathrm{M}'} \tilde{\mu}_j(m_P)[x', (r', \mathrm{M}')] = 1$. Therefore, $\tilde{\mu}_j(m_P) \in \Delta(\widetilde{X}_j \times \widetilde{R}_j \times P_+(M_P))$. Moreover, $\tilde{\mu}_j(m_P)[\mathrm{M}'] = \mathbb{E}_{(x,r)\sim\mu(m_P)}[\mathbb{1}(\mathbf{M}(x, r) = \mathrm{M}')]$ is the probability of realizing some (x, r) for which $\mathbf{M}(x, r) = \mathrm{M}'$ under $\mu(m_P)$. ■

Lemma OA 7. *For all $m_P \in M_P$, $\lim_{j\to\infty} \tilde{\mu}_j(m_P) = \tilde{\mu}(m_P)$.*

Proof. Let $\mathcal{O}$ be an arbitrary open subset of $X \times R$ and M be an arbitrary element of $P_+(M_P)$. We need to show that $\liminf_{j\to\infty} \tilde{\mu}_j(m_P)[\mathcal{O} \times \{\mathrm{M}\}] \geq \tilde{\mu}(m_P)[\mathcal{O} \times \{\mathrm{M}\}]$. For any $\varepsilon > 0$, let $((X \times \widetilde{R}) \setminus \mathcal{O})_{\geq\varepsilon} = \{(x, r) \in \mathcal{O} : \forall\ (x', r') \notin \mathcal{O},\ |x - x'| \geq \varepsilon \text{ or } |r - r'| \geq \varepsilon\}$ be the subset of points in $\mathcal{O}$ that are of distance at least ε from $(X \times R) \setminus \mathcal{O}$. By

construction, for every $j > 1/\varepsilon$, we have $\tilde{\mu}_j(m_P)[\mathcal{O} \times \{\mathrm{M}\}] \geq \tilde{\mu}(m_P)[((X \times \widetilde{R}) \setminus \mathcal{O})_{\geq \varepsilon} \times \{\mathrm{M}\}]$. Since $\lim_{\varepsilon \to 0} \tilde{\mu}(m_P)[((X \times \widetilde{R}) \setminus \mathcal{O})_{\geq \varepsilon} \times \{\mathrm{M}\}] = \tilde{\mu}(m_P)[\mathcal{O} \times \{\mathrm{M}\}]$, the claim follows. ■

Lemma OA 8. *Fix an* $\mathrm{M} \in P_+(M_P)$ *and* $m_P, m'_P \in M_P$ *such that* $m_P, m'_P \in \mathrm{M}$. *For any* $j \in \mathbb{N}$, *let* $q_j \in \Delta(\widetilde{X}_j \times \widetilde{R}_j \times Y_j)$ *and* $q'_j \in \Delta(\widetilde{X}_j \times \widetilde{R}_j \times Y_j)$ *be the distributions induced by the conditional distributions given* M *of* $\tilde{\mu}_j(m_P)$ *and* $\tilde{\mu}_j(m'_P)$, *respectively, when the agent responds to any* (x, r) *according to a fixed action rule* $\beta_{j,A}(x, r) \in \Delta(Y_j)$. *Suppose that* $\lim_{j \to \infty} q_j = q$ *and* $\lim_{j \to \infty} q'_j = q'$. *Then, with probability* 1 *under both* q *and* q', *the conditional distribution of* y *given any* (x, r) *is the same under* q *and* q'.

Proof. Suppose otherwise that the conditional distributions are different under q and q'. Then there is some closed $\mathcal{C} \subseteq X \times \widetilde{R}$, closed $\widehat{Y} \subseteq Y$, and $\kappa > 0$ such that either (1) $\mathbb{P}_q[\mathcal{C}] > 0$ and $\mathbb{P}_q[\widehat{Y}|x, r] > (1 + \kappa)\mathbb{P}_{q'}[\widehat{Y}|x, r]$ for all $(x, r) \in \mathcal{C}$, or (2) $\mathbb{P}_{q'}[\mathcal{C}] > 0$ and $\mathbb{P}_{q'}[\widehat{Y}|x, r] > (1 + \kappa)\mathbb{P}_q[\widehat{Y}|x, r]$ for all $(x, r) \in \mathcal{C}$. Assume without loss of generality that the former holds. Since $h(x, r) \equiv \tilde{f}_{m_P}(x, r)/\tilde{f}_{m'_P}(x, r)$ is measurable, Lusin's theorem implies that there is some closed $\widetilde{\mathcal{C}} \subseteq \mathcal{C}$ satisfying $\mathbb{P}_q[\widetilde{\mathcal{C}}] > 0$ and on which h is continuous.

Fix $\eta > 0$. Since h is continuous and strictly positive on $\widetilde{\mathcal{C}}$, for any $(x, r) \in \widetilde{\mathcal{C}}$, there exists some $\delta_{(x,r)} > 0$ such that $(1 - \eta)h(x, r) < h(x', r') < (1 + \eta)h(x, r)$ whenever $|x' - x|, |r' - r| \leq \delta_{(x,r)}$. Consider the open cover of $\widetilde{\mathcal{C}}$ given by $\{B_{\delta_{(x,r)}}(x, r)\}_{(x,r) \in \widetilde{\mathcal{C}}}$, where, for any $(x, r) \in \widetilde{\mathcal{C}}$ and $\delta > 0$, $B_\delta(x, r) = \{(x', r') \in \widetilde{\mathcal{C}} : |x' - x|, |r' - r| < \delta\}$ is the set of points in $\widetilde{\mathcal{C}}$ of distance less than δ to (x, r). As $\widetilde{\mathcal{C}}$ is compact, this open cover has a finite sub-cover $\{B_{\delta_{(x_k,r_k)}}(x_k, r_k)\}_{1 \leq k \leq K}$. Thus, for at least one $k \in \{1, ..., K\}$, $\mathbb{P}_q[B_{\delta_{(x_k,r_k)}}(x_k, r_k)] > 0$. Throughout the remainder of the proof, we let $\widehat{\mathcal{C}} = B_{\delta_{(x_k,r_k)}}(x_k, r_k)$. Note that, by construction, $(1 - \eta)\rho < \tilde{f}_{m_P}(x, r)/\tilde{f}_{m'_P}(x, r) < (1 + \eta)\rho$

for all $(x,r) \in \widehat{\mathcal{C}}$ where $\rho = \tilde{f}_{m_P}(x_k, r_k)/\tilde{f}_{m'_P}(x_k, r_k) > 0$. From this, it follows that

$$\begin{aligned}
\mathbb{P}_{q'}[\widehat{\mathcal{C}} \times \widehat{Y}] &= \mathbb{E}_{q'}[\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)\mathbb{P}_{q'}[\widehat{Y}|x,r]] \\
&= \mathbb{E}_{q}\left[\frac{\tilde{f}_{m'_P}(x,r)}{\tilde{f}_{m_P}(x,r)}\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)\mathbb{P}_{q'}[\widehat{Y}|x,r]\right] \\
&\leq \left(\frac{1+\eta}{1+\kappa}\right)\rho\mathbb{E}_{q}[\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)\mathbb{P}_{q}[\widehat{Y}|x,r]] \\
&= \left(\frac{1+\eta}{1+\kappa}\right)\rho\mathbb{P}_{q}[\widehat{\mathcal{C}} \times \widehat{Y}],
\end{aligned}$$

where the second equality follows from the Radon-Nikodym theorem. Similarly,

$$\begin{aligned}
\mathbb{P}_{q'}[\widehat{\mathcal{C}}] &= \mathbb{E}_{q'}[\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)] \\
&= \mathbb{E}_{q}\left[\frac{\tilde{f}_{m'_P}(x,r)}{\tilde{f}_{m_P}(x,r)}\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)\right] \\
&\geq (1-\eta)\,\rho\mathbb{E}_{q}[\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)] \\
&= (1-\eta)\,\rho\mathbb{P}_{q}[\widehat{\mathcal{C}}].
\end{aligned}$$

Combining these two inequalities gives

$$\mathbb{P}_{q'}[\widehat{Y}|\widehat{\mathcal{C}}] = \frac{\mathbb{P}_{q'}[\widehat{\mathcal{C}} \times \widehat{Y}]}{\mathbb{P}_{q'}[\widehat{\mathcal{C}}]} \leq \left(\frac{1}{1+\kappa}\right)\left(\frac{1+\eta}{1-\eta}\right)\mathbb{P}_{q}[\widehat{Y}|\widehat{\mathcal{C}}]. \tag{OA 5}$$

For any $\varepsilon > 0$, let $\widehat{\mathcal{C}}_{\leq\varepsilon} = \{(x,r) \in X \times \widetilde{R} :\ \exists(x',r')\ \in \widehat{\mathcal{C}}$ s.t. $|x-x'|, |r-r'| \leq \varepsilon\}$ be the set of points in $X \times R$ that are of distance no more than ε from $\widehat{\mathcal{C}}$. Likewise, let $\widehat{Y}_{\leq\varepsilon} = \{y \in Y :\ \exists\ \hat{y} \in \widehat{Y}$ s.t. $|y-\hat{y}| \leq \varepsilon\}$ be the set of points in Y that are of distance no more than ε from $\widehat{Y}$. Note that $\mathbb{P}_{q'}[\widehat{\mathcal{C}} \times \widehat{Y}] = \lim_{\varepsilon\to 0} \liminf_{j\to\infty} \mathbb{P}_{q'_j}[\widehat{\mathcal{C}}_{\leq\varepsilon} \times \widehat{Y}_{\leq\varepsilon}]$. For

any $j > 1/\varepsilon$,

$$
\begin{aligned}
&\mathbb{P}_{q_j}[\widehat{\mathcal{C}}_{\leq\varepsilon} \times \widehat{Y}_{\leq\varepsilon}] \\
=&\sum_{(x,r)\in(\widetilde{X}_j\times\widetilde{R}_j)\cap\widehat{\mathcal{C}}_{\leq\varepsilon}} \frac{\tilde{\mu}_j(m_P)[x,(r,\mathrm{M})]}{\sum_{(x',r')\in\widetilde{X}_j\times\widetilde{R}_j}\tilde{\mu}_j(m_P)[x',(r',\mathrm{M})]}\beta_{j,A}(x,r)[\widehat{Y}_{\leq\varepsilon}] \\
\leq&\mathbb{E}_q\left[\mathbb{1}((x,r)\in\widehat{\mathcal{C}})\frac{\sum_{(x',r')\in(\widetilde{X}_j\times\widetilde{R}_j)\cap\widehat{\mathcal{C}}_{\leq\varepsilon}}\mathbb{1}\left(|x-x'|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r'|\leq\frac{1}{j}\right)}{\sum_{(x'',r'')\in\widetilde{X}_j\times\widetilde{R}_j}\mathbb{1}\left(|x-x''|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r''|\leq\frac{1}{j}\right)}\beta_{j,A}(x',r')[\widehat{Y}_{\varepsilon}]\right] \\
&+\mathbb{P}_q[\widehat{\mathcal{C}}_{\leq 2\varepsilon}\setminus\widehat{\mathcal{C}}],
\end{aligned}
$$

where the first equality follows by definition and the inequality follows from the construction of $\tilde{\mu}_j(m_P)$ and the fact that no (x,r) of distance more than 2ε from $\widehat{\mathcal{C}}$ contributes positive probability to any $(x',r') \in (\widetilde{X}_j \times \widetilde{R}_j) \cap \widehat{\mathcal{C}}_\varepsilon$. Similarly,

$$
\begin{aligned}
&\mathbb{P}_{q'_j}[\widehat{\mathcal{C}}_{\leq\varepsilon} \times \widehat{Y}_{\leq\varepsilon}] \\
=&\sum_{(x,r)\in(\widetilde{X}_j\times\widetilde{R}_j)\cap\widehat{\mathcal{C}}_{\leq\varepsilon}} \frac{\tilde{\mu}_j(m'_P)[x,(r,\mathrm{M})]}{\sum_{(x',r')\in\widetilde{X}_j\times\widetilde{R}_j}\tilde{\mu}_j(m'_P)[x',(r',\mathrm{M})]}\beta_{j,A}(x,r)[\widehat{Y}_{\leq\varepsilon}] \\
\geq&\mathbb{E}_{q'}\left[\mathbb{1}((x,r)\in\widehat{\mathcal{C}})\frac{\sum_{(x',r')\in(\widetilde{X}_j\times\widetilde{R}_j)\cap\widehat{\mathcal{C}}_{\leq\varepsilon}}\mathbb{1}\left(|x-x'|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r'|\leq\frac{1}{j}\right)}{\sum_{(x'',r'')\in\widetilde{X}_j\times\widetilde{R}_j}\mathbb{1}\left(|x-x''|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r''|\leq\frac{1}{j}\right)}\beta_{j,A}(x',r')[\widehat{Y}_{\varepsilon}]\right] \\
=&\mathbb{E}_q\left[\frac{\tilde{f}_{m'_P}(x,r)}{\tilde{f}_{m_P}(x,r)}\mathbb{1}((x,r)\in\widehat{\mathcal{C}})\frac{\sum_{(x',r')\in(\widetilde{X}_j\times\widetilde{R}_j)\cap\widehat{\mathcal{C}}_{\leq\varepsilon}}\mathbb{1}\left(|x-x'|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r'|\leq\frac{1}{j}\right)}{\sum_{(x'',r'')\in\widetilde{X}_j\times\widetilde{R}_j}\mathbb{1}\left(|x-x''|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r''|\leq\frac{1}{j}\right)}\beta_{j,A}(x',r')[\widehat{Y}_{\varepsilon}]\right] \\
\geq&(1-\eta)\rho\mathbb{E}_q\left[\mathbb{1}((x,r)\in\widehat{\mathcal{C}})\frac{\sum_{(x',r')\in(\widetilde{X}_j\times\widetilde{R}_j)\cap\widehat{\mathcal{C}}_{\leq\varepsilon}}\mathbb{1}\left(|x-x'|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r'|\leq\frac{1}{j}\right)}{\sum_{(x'',r'')\in\widetilde{X}_j\times\widetilde{R}_j}\mathbb{1}\left(|x-x''|\leq\frac{1}{j}\right)\mathbb{1}\left(|r-r''|\leq\frac{1}{j}\right)}\beta_{j,A}(x',r')[\widehat{Y}_{\varepsilon}]\right] \\
\geq&(1-\eta)\rho\mathbb{P}_{q_j}[\widehat{\mathcal{C}}_{\leq\varepsilon} \times \widehat{Y}_{\leq\varepsilon}] - (1-\eta)\rho\mathbb{P}_q[\widehat{\mathcal{C}}_{\leq 2\varepsilon}\setminus\widehat{\mathcal{C}}],
\end{aligned}
$$

where the last inequality comes from the previously established inequality for $\mathbb{P}_{q_j}[\widehat{\mathcal{C}}_{\leq\varepsilon} \times \widehat{Y}_{\leq\varepsilon}]$. Since $\lim_{\varepsilon\to 0}\mathbb{P}_q[\widehat{\mathcal{C}}_{\leq 2\varepsilon}\setminus\widehat{\mathcal{C}}] = 0$, we thus have $\mathbb{P}_{q'_j}[\widehat{\mathcal{C}}\times\widehat{Y}] = \lim_{\varepsilon\geq 0}\liminf_{j\to\infty}\mathbb{P}_{q'_j}[\widehat{\mathcal{C}}_{\leq\varepsilon}\times$

$\widehat{Y}_{\leq\varepsilon}] \geq (1-\eta)\rho \lim_{\varepsilon\to 0} \liminf_{j\to\infty} \mathbb{P}_{q_j}[\widehat{\mathcal{C}}_{\leq\varepsilon} \times \widehat{Y}_{\leq\varepsilon}] = (1-\eta)\rho\mathbb{P}_{q_j}[\widehat{\mathcal{C}} \times \widehat{Y}]$. Moreover, since

$$\begin{aligned}
\mathbb{P}_{q'}[\widehat{\mathcal{C}}] &= \mathbb{E}_{q'}[\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)] \\
&= \mathbb{E}_{q}\left[\frac{\tilde{f}_{m'_P}(x,r)}{\tilde{f}_{m_P}(x,r)}\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)\right] \\
&\leq (1+\eta)\rho\mathbb{E}_q[\mathbb{1}_{\widehat{\mathcal{C}}}(x,r)] \\
&= (1+\eta)\rho\mathbb{P}_q[\widehat{\mathcal{C}}],
\end{aligned}$$

we obtain

$$\mathbb{P}_{q'}[\widehat{Y}|\widehat{\mathcal{C}}] = \frac{\mathbb{P}_{q'}[\widehat{\mathcal{C}} \times \widehat{Y}]}{\mathbb{P}_{q'}[\widehat{\mathcal{C}}]} \geq \left(\frac{1-\eta}{1+\eta}\right)\mathbb{P}_q[\widehat{Y}|\widehat{\mathcal{C}}]. \tag{OA 6}$$

For any $\kappa > 0$, (OA 5) and (OA 6) contradict each other for η sufficiently close to 0. Hence, the conditional distributions under q and q' must be the same. ■

Proof of Lemma 2. The mechanism $(\tilde{\mu}, M_P)$ has the same sequential continuation equilibrium outcomes as (μ, M_P). Thus, to prove Lemma 2, we will show that any limit of sequential continuation equilibrium outcomes after the mechanisms $(\tilde{\mu}_j, M_P)$ in the constructed sequence are proposed is a sequential continuation equilibrium outcome after $(\tilde{\mu}, M_P)$ is proposed. To do so, we need only show that any limit of sequential continuation equilibrium outcomes after the mechanisms $(\tilde{\mu}_j, M_P)$ in the constructed sequence are accepted is a sequential continuation equilibrium outcome after $(\tilde{\mu}, M_P)$ is accepted. The reason is that this, along with the convergence of the agent's utility function, implies that a limit of the agent's acceptance probabilities in the j-th game must be an optimal acceptance probability in the true game given the sequential continuation equilibrium after $(\tilde{\mu}, M_P)$ is accepted.

Let $(\tilde{\lambda}_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \Lambda_j, \beta_{j,A})$ be a consistent assessment in a sequential continuation equilibrium after $(\tilde{\mu}_j, M_P)$ is accepted in the j-th game. We will use this sequence to construct an assessment $(\tilde{\lambda}^*, \pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \Lambda^*, \beta^*_A)$ in the limit game. (In doing so, we assume that all relevant objects have a $j \to \infty$ limit, which is without loss since we can always restrict attention to subsequences of j.) Let $\tilde{\lambda}^* = \lim_{j\to\infty} \tilde{\lambda}_j$ and

$\pi^*_\theta = \lim_{j\to\infty} \pi_{j,\theta}$ for all $\theta \in \Theta$. For any $\mathrm{M} \in P_+(M_P)$ such that$\mathbf{M}(m_P) = \mathrm{M}$ for some $m_P \in M_P$, let $p_j(\mathrm{M}) \in \Delta(\Theta \times X \times \widetilde{R} \times Y)$ be the conditional distribution given M that is induced by this assessment, and let $p^*(\mathrm{M}) = \lim_{j\to\infty} p_j(\mathrm{M})$. The agent's belief updating rule Λ^* is such that $\Lambda^*(x,(r,\mathbf{M}(x,r)))$ equals the conditional distribution of θ given (x,r) under $p^*(\mathbf{M}(x,r))$ for all $(x,r) \in \cup_{m_P\in M_P}\mathrm{supp}(\mu(m_P))$. Likewise, the agent's action rule β^*_A is such that $\beta^*_A(x,(r,\mathbf{M}(x,r)))$ equals the conditional distribution of y given (x,r) under $p^*(\mathbf{M}(x,r))$ for all $(x,r) \in \cup_{m_P\in M_P}\mathrm{supp}(\mu(m_P))$. The construction of $(\tilde{\lambda}^*, \pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \Lambda^*, \beta^*_A)$ guarantees that it is consistent.

We now argue that $(\tilde{\lambda}^*, \pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \Lambda^*, \beta^*_A)$ constitutes a contracting equilibrium after $(\tilde{\mu}, M_P)$ is accepted. For any $m_P \in M_P$, let $q_j(m_P) \in \Delta(X \times Y)$ be the distribution that results from $\tilde{\mu}_j(m_P)$ and the agent responding according to $\beta_{j,A}$. Likewise, let $q^*(m_P) \in \Delta(X\times Y)$ be the distribution that results from $\tilde{\mu}(m_P)$ and the agent responding according to β^*_A. Lemmas OA 7 and OA 8 imply that $\lim_{j\to\infty} q_j(m_P) = q^*(m_P)$ for all $m_P \in M_P$. Then, since the message choices of the principal prescribed in $(\pi_{j,\theta_1}, ..., \pi_{j,\theta_N})$ are optimal given the other's play, it follows that the message choices prescribed in $(\pi^*_{\theta_1}, ..., \pi^*_{\theta_N})$ are also optimal given the other's play. Moreover, a similar argument to that used in the proof of Lemma 1 establishes that the agent's action rule β^*_A assigns probability 1 to best responses to their posterior beliefs about the principal's type. ■

OA.8 Proof of Lemma 4

Lemma 4. *There is a profile of mechanism proposal distributions $\{\eta_\theta\}_{\theta\in\Theta}$ and a measurable mapping $\tau^* : \mathcal{M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^\Theta$ that takes each mechanism $(\mu, M_P) \in \mathcal{M}$ into a tuple consisting of a distribution over the principal's type and a conditional distribution over (α, x, y) for each principal type that corresponds to a single sequential continuation equilibrium after (μ, M_P) is proposed such that*

1. *There is a regular conditional probability distribution obtained from λ and $\{\eta_\theta\}_{\theta\in\Theta}$ that, for every $(\mu, M_P) \in \mathcal{M}$, induces the $\Delta(\Theta)$ component of $\tau^*(\mu, M_P)$ as the*

belief over the principal's type following the proposal of $(\mu, M_P) \in \mathcal{M}$,

2. $U(\theta, \tau^*(\mu, M_P)) \leq U(\theta, p)$ *for all* $\theta \in \Theta$ *and* $(\mu, M_P) \in \mathcal{M}$, *and*
3. $\{\eta_{j,\theta}\}_{\theta\in\Theta}$ *combined with* $\tau^*(\mu, M_P)$ *following the proposal of each* $(\mu, M_P) \in \mathcal{M}$ *induces the same distribution over* (θ, α, x, y) *as outcome* p.

We first develop the class of binary and obedient mechanisms, described in Appendix B.2, that we use to show that there is valid on-path play consistent with the same distribution over (θ, α, x, y) as in p occurring in a contracting equilibrium outcome. For any $(x_1, x_2) \in X \times X$, let $\mu_{(x_1,x_2)} : \{1, 2\} \to \Delta(X \times \mathbb{N})$ be the mapping given by $\mu_{(x_1,x_2)}(m) = \delta_{(x_m,m)}$ for both $m \in \{1, 2\}$, and let $\sigma(x_1, x_2)$ be the set of sequential continuation equilibria after $(\mu_{(x_1,x_2)}, \{1, 2\})$ is proposed. Let $\Sigma = \{(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta) \in X^2 \times \Delta(\Theta) \times \Delta(\{1, 2\})^\Theta \times [0, 1] \times \Delta(Y)^2 : (\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta) \in \sigma(x_1, x_2)\}$. Observe that Σ is a compact subset of $X^2 \times \Delta(\Theta) \times \Delta(\{1, 2\})^\Theta \times [0, 1] \times \Delta(Y)^2$.

Let

$$\mathcal{M}^{bin*} = \{(\mu, \{1, 2\}) \in \mathcal{M} : \exists (x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta) \in \Sigma \text{ s.t.}$$
$$\text{supp}(\mu(1)) = (x_1, (\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta(1))) \text{ and } \text{supp}(\mu(2)) = (x_2, (\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta(2)))\}.$$

We will say that there is *obedient play* following the proposal of the mechanism in $\mathcal{M}^{bin*}$ corresponding to $(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta) \in \Sigma$ if each principal type θ plays according to π_θ and the agent plays according to (α, β). For every θ and $(\mu, \{1, 2\}) \in \mathcal{M}^{bin*}$, we let $\tau^{obed}(\mu, \{1, 2\}) \in \Delta(\Theta) \times \Delta([0, 1] \times X \times Y)^\Theta$ denote the tuple consisting of the distribution over the principal's type $\tilde{\lambda}$ and the distribution over (α, x, y) for each principal type that results from the proposal of $(\mu, \{1, 2\})$ if it is followed by obedient play.

Lemma OA 9. *There is a profile of mechanism proposal distributions* $\{\eta_{j,\theta}\}_{\theta\in\Theta} \subset \Delta(\mathcal{M}^{bin*})$ *such that*

1. *There is a regular conditional probability distribution obtained from* λ *and* $\{\eta_{j,\theta}\}_{\theta\in\Theta}$ *that, for every* $(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta) \in \Sigma$, *induces* $\tilde{\lambda}$ *as the belief over the*

principal's type following the proposal of the mechanism in $\mathcal{M}^{bin*}$ *corresponding to* $(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta)$.

2. $\{\eta_{j,\theta}\}_{\theta\in\Theta}$ *combined with the principal and agent playing obediently for each mechanism in* $\mathcal{M}^{bin*}$ *induces the same distribution over* (θ, α, x, y) *as outcome* p_j.
3. $U(\theta, \tau^{obed}(\mu, \{1,2\})) \leq U(\theta, p_j)$ *for all* $\theta \in \Theta$ *and* $(\mu, \{1,2\}) \in \cup_{\theta'\in\Theta} supp(\eta_{j,\theta'})$.

We prove Lemma OA 9 in the following way. For any mechanism (μ, M_P) that is proposed with positive probability under p_j, we construct a joint distribution over principal types and mechanisms in $\mathcal{M}^{bin*}$ that, when coupled with obedient play, leads to the same distribution over (θ, α, x, y) as the conditional distribution of p_j given the proposal of (μ, M_P). The key will be to pair off the various actions that occur with positive probability in p_j after the acceptance of (μ, M_P) into separate binary mechanisms in such a way that the agent is willing to accept these mechanisms with precisely the same probability with which they accept (μ, M_P) in p_j. This requires appropriately choosing the various mechanism proposal probabilities and probabilities of each of the two actions being chosen after any given mechanism is accepted. We then aggregate over the distributions of principal types and mechanisms in $\mathcal{M}^{bin*}$ identified separately for each on-path mechanism in $\mathcal{M}_j$ to obtain a profile of proposal distributions over mechanisms in $\mathcal{M}^{bin*}$ that results in an outcome of p_j.

Proof. Consider the equilibrium $((\eta^*_{j,k,\theta_1}, \boldsymbol{\pi}^*_{j,k,\theta_1}), ..., (\eta^*_{j,k,\theta_N}, \boldsymbol{\pi}^*_{j,k,\theta_N}), (\boldsymbol{\alpha}^*_{j,k}(\cdot), \boldsymbol{\beta}^*_{j,k}(\cdot)))$ of the (j,k) game, and, restricting attention to a convergent subsequence if necessary, let

$$\begin{aligned}&((\eta^*_{j,\theta_1}, \boldsymbol{\pi}^*_{j,\theta_1}), ..., (\eta^*_{j,\theta_N}, \boldsymbol{\pi}^*_{j,\theta_N}), (\boldsymbol{\alpha}^*_{j}(\cdot), \boldsymbol{\beta}^*_{j}(\cdot)))\\ =&\lim_{k\to\infty}((\eta^*_{j,k,\theta_1}, \boldsymbol{\pi}^*_{j,k,\theta_1}), ..., (\eta^*_{j,k,\theta_N}, \boldsymbol{\pi}^*_{j,k,\theta_N}), (\boldsymbol{\alpha}^*_{j,k}(\cdot), \boldsymbol{\beta}^*_{j,k}(\cdot))).\end{aligned}$$

Fix an arbitrary $(\mu, M_P) \in \mathcal{M}_j$ that is proposed with positive probability under $(\eta^*_{j,\theta_1}, ..., \eta^*_{j,\theta_N})$. Let $\tilde{\lambda}^*_j(\mu, M_P) \in \Delta(\Theta)$ be the posterior distribution over the principal's type conditional on the proposal of (μ, M_P). Further, let $(x_1, \mathbf{y}_1), ..., (x_M, \mathbf{y}_M) \in$

$X_j \times \Delta(Y)$ be the pairs of principal actions and agent action distributions that occur with positive probability under $(\boldsymbol{\pi}^*_{j,\theta_1}(\mu, M_P), ..., \boldsymbol{\pi}^*_{j,\theta_N}(\mu, M_P), \boldsymbol{\beta}^*_j(\cdot))$ when (μ, M_P) is accepted. For every $\theta \in \Theta$ and $m \in \{1, ..., M\}$, we use $q_{j,(x_m,\mathbf{y}_m)}(\theta)$ to denote the probability of $(x_m, \mathbf{y}_m)$ conditional on type θ and (μ, M_P) being accepted under $(\boldsymbol{\pi}^*_{j,\theta_1}(\mu, M_P), ..., \boldsymbol{\pi}^*_{j,\theta_N}(\mu, M_P), \boldsymbol{\beta}^*_j(\cdot))$.

The $k \to \infty$ limit of the expected utility of the agent from accepting (μ, M_P) in the (j, k) equilibrium is thus $V_j(\mu, M_P) = \sum_{m\in\{1,...,M\}} \sum_{\theta\in\Theta} \tilde{\lambda}^*_j(\theta|\mu, M_P) q_{j,(x_m,\mathbf{y}_m)}(\theta) \mathbb{E}_{\mathbf{y}_m}[V(\theta, x_m, y)]$. Observe that there are collections of pairs $\{(m_{l,1}, m_{l,2})\}_{l\in\{1,...,L\}}$ and $\{(s_{l,1}, s_{l,2})\}_{l\in\{1,...,L\}}$ for some $L \in \mathbb{N}$ such that

(a) $s_{l,1} > 0$ and $s_{l,2} \geq 0$ for all $l \in \{1, ..., L\}$,

(b) $\sum_{l\in\{1,...,L\}} (\mathbb{1}_{m_{l,1}=m}(l) s_{l,1} + \mathbb{1}_{m_{l,2}=m}(l) s_{l,2}) = 1$ for all $m \in \{1, ..., M\}$, and

(c) For all $l \in \{1, ..., L\}$,

$$\begin{aligned}
&\text{sign}\Big(s_{l,1} \sum_{\theta\in\Theta} \tilde{\lambda}^*_j(\theta|\mu, M_P) q_{j,l,1}(\theta) \mathbb{E}_{\mathbf{y}_{l,1}}[V(\theta, x_{l,1}, y)] \\
&\qquad + s_{l,2} \sum_{\theta\in\Theta} \tilde{\lambda}^*_j(\theta|\mu, M_P) q_{j,l,2}(\theta) \mathbb{E}_{\mathbf{y}_{l,2}}[V(\theta, x_{l,2}, y)]\Big) \\
=&\text{sign}\left(V_j(\mu, M_P)\right),
\end{aligned}$$

where $x_{l,1} = x_{m_{l,1}}$, $x_{l,2} = x_{m_{l,2}}$, $q_{j,l,1}(\theta) = q_{j,(x_{m_{l,1}},\mathbf{y}_{m_{l,1}})}(\theta)$, and $q_{j,l,2}(\theta) = q_{j,(x_{m_{l,2}},\mathbf{y}_{m_{l,2}})}(\theta)$ for all $\theta \in \Theta$ and $l \in \{1, ..., L\}$.

For each $l \in \{1, ..., L\}$, we create a mechanism $(\mu_l, \{1, 2\}) \in \mathcal{M}^{bin*}$ in which $m_P = 1$ maps to $x_{l,1}$ and a recommended action distribution of $\mathbf{y}_{l,1}$, and $m_P = 2$ maps to $x_{l,2}$ and a recommended action distribution of $\mathbf{y}_{l,2}$. We will have each type θ to propose the $(\mu_l, \{1, 2\})$ mechanism with probability $s_{l,1} q_{j,l,1}(\theta) + s_{l,2} q_{j,l,2}(\theta)$. Conditions (a) and (b) ensure that this constitutes a valid mechanism proposal distribution. Moreover, after the acceptance of a $(\mu_l, \{1, 2\})$ that they propose with positive probability, we will have the type θ principal play $m_P = 1$ with probability $s_{l,1} q_{j,l,1}(\theta) / (s_{l,1} q_{j,l,1}(\theta) + s_{l,2} q_{j,l,2}(\theta))$ and $m_P = 2$ with complementary probability $s_{l,2} q_{j,l,2}(\theta) / (s_{l,1} q_{j,l,1}(\theta) + s_{l,2} q_{j,l,2}(\theta))$. (For any mechanism that they propose with 0 probability, we will have the type θ play

$m_P = 1$ whenever they weakly prefer $(x_{l,1}, \mathbf{y}_{l,1})$ to $(x_{l,2}, \mathbf{y}_{l,2})$ and otherwise play $m_P = 2$, but the precise message selection rules in these cases are irrelevant.) We will also have the agent follow every action recommendation. Condition (c) then implies that the agent's expected utility conditional on the acceptance of any of the $(\mu_l, \{1,2\})$ mechanisms has the same sign as their expected utility from accepting (μ, M_P) in the $k \to \infty$ limit of the (j,k) equilibria. Thus, the agent will be willing to accept each of the $(\mu_l, \{1,2\})$ mechanisms with the same probability $\boldsymbol{\alpha}^*_j(\mu, M_P)$, and indeed we will have them do so. This means that the conditional distribution of (α, x, y) given type θ is exactly as when (μ, M_P) is proposed in the $k \to \infty$ limit of the (j,k) equilibria.

The specific mechanisms are as follows. For each $(m_{l,1}, m_{l,2})$, consider the $(\mu_l, \{1,2\}) \in \mathcal{M}^{bin*}$ given by

$$\begin{aligned}
\text{supp}(\mu(1)) &= (x_{l,1}, (\tilde{\lambda}^*_{j,l}, \pi^*_{j,l,\theta_1}, ..., \pi^*_{j,l,\theta_N}, \boldsymbol{\alpha}^*_j(\mu, M_P), \mathbf{y}_{l,1})), \\
\text{supp}(\mu(2)) &= (x_{l,2}, (\tilde{\lambda}^*_{j,l}, \pi^*_{j,l,\theta_1}, ..., \pi^*_{j,l,\theta_N}, \boldsymbol{\alpha}^*_j(\mu, M_P), \mathbf{y}_{l,2})),
\end{aligned}$$

where, for each $\theta \in \Theta$,

$$\begin{aligned}
\tilde{\lambda}^*_{j,l}(\theta) &= \frac{\tilde{\lambda}^*_j(\theta|\mu, M_P)(s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta))}{\sum_{\theta' \in \Theta} \tilde{\lambda}^*_j(\theta'|\mu, M_P)(s_{l,1}q_{j,l,1}(\theta') + s_{l,2}q_{j,l,2}(\theta'))}, \\
\pi^*_{j,l,\theta}(1) &= \begin{cases} \frac{s_{l,1}q_{j,l,1}}{s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta)} & \text{if } s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta) > 0 \\ 0 & \text{if } \mathbb{E}_{\mathbf{y}_{l,1}}[U(\theta, x_{l,1}, y)] < \mathbb{E}_{\mathbf{y}_{l,2}}[U(\theta, x_{l,2}, y)] \\ 1 & \text{otherwise} \end{cases}, \\
\pi^*_{j,l,\theta}(2) &= 1 - \pi^*_{j,l,\theta}(1).
\end{aligned}$$

Note that $\tilde{\lambda}^*_{j,l}$ is the posterior distribution induced by having each type θ propose the $(\mu_l, \{1,2\})$ mechanism with probability $s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta)$ given prior distribution $\tilde{\lambda}^*_j(\mu, M_P)$, while $\pi^*_{j,l,\theta}$ gives the mechanism selection probabilities for each type θ as discussed above.

We have already established that if each principal type θ proposes the binary mech-

anism corresponding to $(m_{l,1}, m_{l,2})$ with probability $s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta)$ and the subsequent play is obedient, then the resulting outcome distribution is the same as that which follows the proposal of (μ, M_P). We now argue that obedient play is consistent with these proposal probabilities and sequential continuation equilibrium. To see this, observe that, the only $(x, \mathbf{y})$ pairs that occur with positive probability conditional on type θ are those which occurred with positive probability conditional on θ under the $k \to \infty$ limit of the (j, k) equilibrium outcome following the acceptance of (μ, M_P). Because the principal's trembles vanish in the $k \to \infty$ limit, these must be the optimal $(x, \mathbf{y})$ pairs for the principal type. Additionally, in the mechanism corresponding to $(m_{l,1}, m_{l,2})$, the posterior over the principal's type after $(x_{l,1}, \mathbf{y}_{l,2})$ is the same as the posterior after $(x_{m_{l,1}}, \mathbf{y}_{l,2})$ in the $k \to \infty$ of the (j, k) equilibrium. Likewise, for $(x_{l,1}, \mathbf{y}_{l,2})$. So the prescribed agent play (after acceptance) is always optimal since the agent also has vanishing trembles (and the agent's action space converges to their true action space) as $k \to \infty$. Finally, as noted above, Condition (c) ensures that the prescribed acceptance probability of $\boldsymbol{\alpha}_j^*(\mu, M_P)$ is optimal for the agent.

We have thus obtained a profile of mechanism proposal distributions $\{\eta_{j,\mu,M_P,\theta}\}_{\theta\in\Theta}$ corresponding to an arbitrary on-path (in the $k \to \infty$ limit of the (j, k) equilibrium) mechanism $(\mu, M_P) \in \mathcal{M}_j$ that satisfies the three conditions of Lemma OA 9 with the following qualification: In Condition 2, the outcome p_j needs to be replaced with the outcome conditional on (μ, M_P) being proposed. This can be done for every mechanism in $\mathcal{M}_j^{on} = \{(\mu, M_P) \in \mathcal{M} : \exists\theta \in \Theta \text{ s.t. } \eta_{j,\theta}^*(\mu, M_P) > 0\}$, the set of on-path mechanisms (according to the $k \to \infty$ limit of the (j, k) equilibrium). Averaging over the mechanism proposal distributions for type θ weighted by the probability of proposing each (μ, M_P) then gives $\eta_{j,\theta} = \sum_{(\mu,M_P)\in\mathcal{M}_j} \eta_{j,\theta}^*(\mu, M_P)\eta_{j,\mu,M_P,\theta}$. The profile of these mechanism proposal distributions satisfies all the conditions of Lemma OA 9. ■

Lemma OA 10. *There is a profile of mechanism proposal distributions* $\{\eta_\theta\}_{\theta\in\Theta} \subset \Delta(\mathcal{M}^{bin*})$ *such that*

1. *There is a regular conditional probability distribution obtained from* λ *and* $\{\eta_\theta\}_{\theta\in\Theta}$

that, for every $(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta) \in \Sigma$*, induces* $\tilde{\lambda}$ *as the belief over the principal's type following the proposal of the mechanism in* $\mathcal{M}^{bin*}$ *corresponding to* $(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \alpha, \beta)$*.*

2. $\{\eta_\theta\}_{\theta\in\Theta}$ *combined with the principal and agent playing obediently for each mechanism in* $\mathcal{M}^{bin*}$ *induces the same distribution over* (θ, α, x, y) *as outcome* p*.*
3. $U(\theta, \tau^{obed}(\mu, \{1,2\})) \leq U(\theta, p)$ *for all* $\theta \in \Theta$ *and* $(\mu, \{1,2\}) \in \cup_{\theta'\in\Theta} supp(\eta_{\theta'})$*.*

Proof. For each $\theta \in \Theta$, let $\eta_\theta \in \Delta(\mathcal{M})$ be a limit point of the sequence $\{\eta_{j,\theta}\}_{j\in\mathbb{N}}$. Without loss, suppose that $\lim_{j\to\infty} \eta_{j,\theta} = \eta_\theta$ for all $\theta \in \Theta$. Since $\Delta(\mathcal{M}^{bin*})$ is closed, it follows that each $\eta_\theta \in \Delta(\mathcal{M}^{bin*})$.

We first demonstrate Condition 1. For every $j \in \mathbb{N}$, let $\eta_j = \sum_{\theta\in\Theta} \lambda(\theta)\eta_{j,\theta} \in \Delta(\mathcal{M}^{bin*})$, and likewise let $\eta = \sum_{\theta\in\Theta} \lambda(\theta)\eta_\theta \in \Delta(\mathcal{M}^{bin*})$. Let $p_{j,\Theta\times\mathcal{M}^{bin*}} \in \Delta(\Theta \times \mathcal{M}^{bin*})$ be the probability distribution over pairs of principal types and binary and obedient mechanisms induced by η_j as the distribution over mechanisms and $\tilde{\lambda}(\cdot|(M, \mu_P))$ the conditional distribution over the principal's type given mechanism $(M, \mu_P) \in \mathcal{M}^{bin*}$. Similarly, let $p_{\Theta\times\mathcal{M}^{bin*}} \in \Delta(\Theta \times \mathcal{M}^{bin*})$ be the probability distribution induced by η and $\tilde{\lambda}(\cdot|(M, \mu_P))$. Fix $\theta \in \Theta$ and let $\overline{\mathcal{M}}$ be a closed subset of $\mathcal{M}^{bin*}$. Since $\tilde{\lambda}(\theta|(M, \mu_P))\mathbb{1}_{\overline{\mathcal{M}}}(M, \mu_P)$ is an upper semicontinuous function of $(M, \mu_P) \in \mathcal{M}^{bin*}$, it follows that

$$\begin{aligned}
\mathbb{P}_{p_{\Theta\times\mathcal{M}^{bin*}}}[\{\theta\} \times \overline{\mathcal{M}}] &= \mathbb{E}_\eta[\tilde{\lambda}(\theta|(M, \mu_P))\mathbb{1}_{\overline{\mathcal{M}}}(M, \mu_P)] \\
&\geq \limsup_{j\to\infty} \mathbb{E}_{\eta_j}[\tilde{\lambda}(\theta|(M, \mu_P))\mathbb{1}_{\overline{\mathcal{M}}}(M, \mu_P)] \\
&= \limsup_{j\to\infty} \mathbb{P}_{p_{j,\Theta\times\mathcal{M}^{bin*}}}[\{\theta\} \times \overline{\mathcal{M}}].
\end{aligned}$$

Because θ and $\overline{\mathcal{M}}$ are arbitrary, we conclude that $\lim_{j\to\infty} p_{j,\Theta\times\mathcal{M}^{bin*}} = p_{\Theta\times\mathcal{M}^{bin*}}$, which means Condition 1 is satisfied.

Now we show that Condition 2 holds. Fix $\theta \in \Theta$ and let $\overline{X}$, $\overline{A}$, and $\overline{Y}$ be arbitrary closed subsets of X, $[0, 1]$, and Y, respectively. Since $\tilde{\lambda}(\theta|(\mu, \{1,2\}))\tau^{obed}(\overline{X} \times \overline{A} \times \overline{Y}|\theta, \mu, \{1,2\})$ is a continuous function of $(\mu, \{1,2\}) \in \mathcal{M}^{bin*}$, it follows that

$\mathbb{E}_{\eta}[\tilde{\lambda}(\theta|(\mu,\{1,2\}))\tau^{obed}(\overline{X}\times\overline{A}\times\overline{Y}|\theta,\mu,\{1,2\})] = \lim_{j\to\infty}\mathbb{E}_{\eta_j}[\tilde{\lambda}(\theta|(\mu,\{1,2\}))\tau^{obed}(\overline{X}\times\overline{A}\times\overline{Y}|\theta,\mu,\{1,2\})] = \lim_{j\to\infty}\mathbb{P}_{p_j}[\{\theta\}\times\overline{X}\times\overline{A}\times\overline{Y}] = \mathbb{P}_p[\{\theta\}\times\overline{X}\times\overline{A}\times\overline{Y}]$. Because θ, $\overline{X}$, $\overline{A}$, and $\overline{Y}$ are arbitrary, we conclude that $\{\eta_\theta\}_{\theta\in\Theta}$, together with obedient play, induces p.

Finally, since $U(\theta,\tau^{obed}(\mu,\{1,2\})) \leq U(\theta,p_j)$ for all $\theta\in\Theta$ and $(\mu,\{1,2\})\in\cup_{\theta'\in\Theta}\text{supp}(\eta_{j,\theta'})$, standard continuity arguments show that $U(\theta,\tau^{obed}(\mu,\{1,2\})) \leq U(\theta,p)$ for all $\theta\in\Theta$ and $(\mu,\{1,2\})\in\cup_{\theta'\in\Theta}\text{supp}(\eta_{\theta'})$. ■

We now develop the class of revealing mechanisms, also described in Appendix B.2, to show that there is valid off-path play consistent with contracting equilibria in which the principal types receive the same payoffs as they get in p.

For every $M\in\mathbb{N}$, let

$$\mathcal{M}^{rev,M} = \{(\mu,\{1,...,M\})\in\mathcal{M} : \exists(x_1,...,x_M)\in X^M \text{ s.t. } \text{supp}(\mu(m)) = (x_m,m)\ \forall m\in\{1,...,M\}\}.$$

be the set of deterministic mechanisms with M messages in which the message chosen by the principal constitutes the recommendation received by the agent.

Lemma OA 11. *For every $M\in\mathbb{N}$ and $(\mu,\{1,...,M\})\in\mathcal{M}^{rev,M}$, there is a sequential continuation equilibrium after $(\mu,\{1,...,M\})$ is proposed that gives every principal type a weakly lower payoff than $U(\theta,p)$.*

Proof. Note that any $(\mu,\{1,...,M\})\in\mathcal{M}^{rev,M}$ can be approximated to arbitrary accuracy by some sequence of mechanisms $\{(\mu_j,\{1,...,M\})\}_{j\in\mathbb{N}}$, where $(\mu_j,\{1,...,M\})\in\mathcal{M}_j$ for all large enough j. Because of the vanishing trembles of the principal, it follows that, in the $k\to\infty$ limit, for the proposal of any mechanism in $\mathcal{M}_j$, there is a sequential continuation equilibrium which gives the principal types a lower payoff than what they receive from p_j. For all sufficiently large j, let $(\tilde{\lambda}_j,\pi_{j,\theta_1},...,\pi_{j,\theta_N},\alpha_j,\beta_j)$ denote such a sequential continuation equilibrium for the proposal of $(\mu_j,\{1,...,M\})$. Standard arguments show that any limit point of these sequential continuation equilibria is itself a sequential continuation equilibrium following the proposal of $(\mu_j,\{1,...,M\})$.

Continuity ensures that this sequential continuation equilibrium gives each principal type a lower payoff than what they receive from p. ■

Lemma OA 12. *For every $M \in \mathbb{N}$, there is a measurable mapping $\tau^{M*} : \mathcal{M}^{rev,M} \to \Delta(\Theta) \times \Delta(\{1, ..., M\})^{\Theta} \times [0,1] \times \Delta(Y)^M$ such that, for every $(\mu, \{1, ..., M\}) \in \mathcal{M}^{rev,M}$, $\tau^{M*}(\mu, \{1, ..., M\})$ is a sequential continuation equilibrium after $(\mu, \{1, ..., M\})$ is proposed that gives every principal type a weakly lower payoff than $U(\theta, p)$.*

Proof. For each $(\mu, \{1, ..., M\}) \in \mathcal{M}^{rev,M}$, let $\sigma^{M*}(\mu, \{1, ..., M\})$ be the set of sequential continuation equilibria after (μ, M_P) is proposed which give the principal types weakly lower payoffs than they obtain from p.

By Lemma OA 12, $\sigma^{M*}(\mu, \{1, ..., M\})$ is non-empty for all $(\mu, \{1, ..., M\}) \in \mathcal{M}^{rev,M}$. Additionally, standard arguments show that $\sigma^{M*} : \mathcal{M}^{rev,M} \rightrightarrows \Delta(\Theta) \times \Delta(\{1, ..., M\})^{\Theta} \times [0,1] \times \Delta(Y)^M$ is an upper hemicontinuous correspondence. Since $\mathcal{M}^{rev,M}$ is compact, Lemma 1 in Section D of Hildenbrand [1974] guarantees that there is a measurable selection of σ^{M*} and hence the desired τ^{M*}. ■

Lemma OA 13. *There is a measurable mapping $\tau^{\dagger} : \mathcal{M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^{\Theta}$ that takes each mechanism $(\mu, M_P) \in \mathcal{M}$ into a tuple consisting of a distribution over the principal's type and a distribution over (α, x, y) for each principal type that corresponds to a single sequential continuation equilibrium outcome after (μ, M_P) is proposed that gives every principal type a weakly lower payoff than $U(\theta, p)$.*

Proof. Consider arbitrary $M \in \mathbb{N}$ and the mapping $\tau^{M*} : \mathcal{M}^{rev,M} \to \Delta(\Theta) \times \Delta(\{1, ..., M\})^{\Theta} \times [0,1] \times \Delta(Y)^M$ identified in Lemma OA 12. Let $\tau^{M\dagger} : \mathcal{M}^{rev,M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^{\Theta}$ be the mapping that specifies the probability distribution over types and the distributions over (α, x, y) corresponding to $\tau^{M*}(\mu, \{1, ..., M\})$ for each $(\mu, \{1, ..., M\})$. Note that $\tau^{M\dagger}$ is measurable.

Fix some $M \in \mathbb{N}$. Consider $\mathcal{M}^{eff,M} = \{(\mu, M_P) \in \mathcal{M} : |\cup_{m \in M_P} \text{supp}(\mu(m))| = M\}$, the set of mechanisms which can effectively induce exactly M distinct principal action-recommendation pairs. For each $(\mu, M_P) \in \mathcal{M}^{eff,M}$, let $\mathcal{X}^M(\mu, M_P) = (x_1, ..., x_M)$ be

the M-tuple of principal actions where, for every $m \in \{1, ..., M\}$, x_m is the action induced by the m-th distinct principal action-transfer pair, as determined by the natural order from the messages $M_P = \{1, ..., M'\}$. Further, let $f : \mathcal{M}^{eff,M} \to \mathcal{M}^{rev,M}$ be the mapping such that $f(\mu, M_P) = \mathcal{X}^M(\mu, M_P)$ for every $(\mu, M_P) \in \mathcal{M}^{eff,M}$. By construction, f is measurable. Moreover, the sets of sequential continuation equilibrium outcomes following the proposal of (μ, M_P) or $f(\mu, M_P)$ are precisely the same for all $(\mu, M_P) \in \mathcal{M}^{eff,M}$. Thus, the composition $\tau^{M*} \circ f : \mathcal{M}^{eff,M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^\Theta$ is a measurable mapping that, for any $(\mu, M_P) \in \mathcal{M}^{eff,M}$, gives a distribution over the principal's type and a distribution over (α, x, y) for each principal type that corresponds to a single sequential continuation equilibrium outcome after (μ, M_P) is proposed in which every principal type receives a weakly lower payoff than $U(\theta, p)$.

Since the $\mathcal{M}^{eff,M}$ are disjoint, measurable subsets of $\mathcal{M}$ satisfying $\cup_{M \in \mathbb{N}} \mathcal{M}^{eff,M} = \mathcal{M}$, the mapping $\tau^\dagger : \mathcal{M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^\Theta$ such that $\tau^\dagger(\mu, M_P) = \tau^{M*}(f(\mu, M_P))$ for any $(\mu, M_P) \in \mathcal{M}^{eff,M}$ has all the desired properties. ■

Proof of Lemma 4. Let the profile of mechanism proposal distributions be the same $\{\eta_\theta\}_{\theta \in \Theta}$ as identified in Corollary OA 10. Additionally, consider the mapping $\tau^* : \mathcal{M} \to \Delta(\Theta) \times \Delta([0,1] \times X \times Y)^\Theta$ defined by

$$\tau^*(\mu, M_P) = \begin{cases} \tau^{obed}(\mu, \{1,2\}) & \text{if } (\mu, M_P) = (\mu, \{1,2\}) \in \cup_{\theta \in \Theta} \text{supp}(\eta_\theta) \\ \tau^\dagger(\mu, M_P) & \text{if } (\mu, M_P) \notin \cup_{\theta \in \Theta} \text{supp}(\eta_\theta). \end{cases}$$

By construction, τ^* is measurable. Additionally, Lemma OA 10 guarantees that Conditions 1 and 3 of Lemma 4 hold with this $\{\eta_\theta\}_{\theta \in \Theta}$ and τ^*, while Lemmas OA 10 and OA 13 together ensure that Condition 2 of Lemma 4 is satisfied. ■

OA.9 Proofs of Lemmas 5 and 6

OA.9.1 Proof of Lemma 5

Lemma 5. *In MCS environments, there are sequences of full-support distributions over the principal type* $\{\lambda_k\}_{k\in\mathbb{N}}$ *and outcomes* $\{p_k\}_{k\in\mathbb{N}}$ *such that*

1. $marg_\Theta p_k = \lambda_k$ *for all* $k \in \mathbb{N}$,
2. $\liminf_{k\to\infty} \mathbb{E}_{p_k}[v(\theta, x, y) + g(t)|\theta] \geq 0$ *for all* $\theta \in \Theta$,
3. $\mathbb{E}_{p_k}[u(\theta, x, y) - t|\theta] \geq \mathbb{E}_{p_k}[u(\theta, x, y) - t|\theta']$ *for all* $\theta, \theta' \in \Theta$ *and* $k \in \mathbb{N}$,
4. $\mathbb{P}_{p_k}[y = y^*(\theta, x)|\theta, x \neq x_o] = 1$ *for all* $\theta \in \Theta$ *and* $k \in \mathbb{N}$, *and*
5. *For each mechanism* $(\mu, M_P) \in \mathcal{M}$ *and* $k \in \mathbb{N}$, *there is a sequential continuation equilibrium after* (μ, M_P) *is proposed that gives every principal type a weakly lower payoff than* p_k.

Construction of Hypothetical Games. Let $\overline{V} = \max_{(\theta,x,t,y)} v(\theta, x, y)+g(t)$. For all $k \in \mathbb{N}$ satisfying $k > |\Theta|$, let

$$V_k^\dagger = -(|\Theta|\overline{V} + 1)/(k - |\Theta|). \tag{OA 7}$$

Note that $V_k^\dagger$ is such that $(1 - (|\Theta| - 1)/k)V_k^\dagger + (|\Theta| - 1)\overline{V}/k < -1/k$. This means that, if the agent's conditional expected utility given some principal type is weakly lower than $V_k^\dagger$ and the probability of this type is at least $1 - (|\Theta| - 1)/k$, the agent's total expected utility is less than $-1/k$.

Let $\{X_j\}_{j\in\mathbb{N}}$, $\{T_j\}_{j\in\mathbb{N}}$, $\{Y_j\}_{j\in\mathbb{N}}$, $\{R_j\}_{j\in\mathbb{N}}$ be sequences of finite action, transfer, and recommendation sets such that $\lim_{j\to\infty} X_j = X$, $\lim_{j\to\infty} T_j = T$, $\lim_{j\to\infty} Y_j = Y$, and $\lim_{j\to\infty} R_j = R$. For a given $j \in \mathbb{N}_{++}$, consider the set of mechanisms

$$\mathcal{M}_j = \Big\{(\mu, M_P) \in \mathcal{M} : \ (1)\ |M_P| \leq j,$$
$$(2)\ \forall m_P \in M_P, x \in X_j, t \in T_j, r \in R_j, \exists k \in \mathbb{N} \text{ s.t. } \mu((x, t), r|m_P) = \frac{k}{j|X_j||T_j||R_j|}\Big\}$$

that (1) have no more than j principal messages and (2) are such that the probability of a given principal action-transfer-recommendation tuple conditional on any message

is some integer multiple of $1/(j|X_j||T_j|)$. Similarly, let

$$\Delta_j(X_j \times T_j) = \left\{ \chi \in \Delta(X_j \times T_j) : \ \forall x \in X_j, t \in T_j, \ \exists k \in \mathbb{N} \text{ s.t. } \chi[(x,t)] = \frac{k}{j|X_j||T_j|} \right\}$$

be the set of distributions over $X_j \times T_j$ such that the probability of a given principal action-transfer pair is some integer multiple of $1/(j|X_j||T_j|)$. We suppose that, for all $j \in \mathbb{N}$, the recommendation space is strictly larger than the set of principal types, i.e. $|R_j| > |\Theta|$. For notational convenience, we will assume that the power set of principal types is in fact a strict subset of the recommendation spaces.

We now describe the strategy space of the type θ principal in the j-th game. Part of this player's choice is over which mechanisms to propose. We force θ to propose almost all mechanisms with positive probability. The exceptions are mechanisms which commit to some χ as the distribution over principal actions and some $\theta' \neq \theta$ as the recommendation received by the agent; θ is required to propose these mechanisms with 0 probability. Formally, let $\mu_{\chi,\theta'} \in \Delta(X \times T \times R)$ be the distribution satisfying $\text{marg}_X \mu_{\chi,\theta'} = \chi$ and $\mu_{\chi,\theta'}[\theta'] = 1$, and let $\mathcal{M}^c_{j,\theta'} = \{(\mu_{\chi,\theta'}, \{0\}) : \chi \in \Delta_j(X_j \times T_j), \ \theta' \neq \theta\}$ be the set of mechanisms in the j-th game that commit to some χ as the distribution over principal actions and θ' as the recommendation received by the agent. Additionally, let $\mathcal{M}^0_{j,\theta} = \cup_{\theta' \neq \theta} \mathcal{M}^c_{j,\theta'}$. The distribution over mechanism proposals used by θ must belong to

$$\Delta_{j,\theta}(\mathcal{M}_j) = \Big\{ \eta \in \Delta(\mathcal{M}_j) : \ (1) \ \eta[(\mu, M_P)] \geq \frac{1}{j|\mathcal{M}_j|} \ \forall (\mu, M_P) \in \mathcal{M}_j \setminus \mathcal{M}^0_{j,\theta}, \\ (2) \ \eta[(\mu, M_P)] = 0 \ \forall (\mu, M_P) \in \mathcal{M}^0_{j,\theta} \Big\}.$$

Moreover, when a given mechanism is accepted, we force θ to tremble and play every message in the mechanism with positive probability. Formally, the distribution over messages used by θ when mechanism (μ, M_P) is accepted must belong to

$$\Pi_{j,P}(\mu, M_P) = \Big\{ \pi_P \in \Delta(M_P) : \pi_P[m_P] \geq \frac{1}{j|M_P|} \ \forall m_P \in M_P \Big\}.$$

A valid strategy for θ in the j-th game is any pair $(\eta_\theta, \boldsymbol{\pi}_\theta(\cdot))$ consisting of a $\eta_\theta \in \Delta_{j,\theta}(\mathcal{M}_j)$ and a rule $\boldsymbol{\pi}_\theta(\cdot)$ for how to play when an arbitrary mechanism is accepted that satisfies $\boldsymbol{\pi}_\theta(\mu, M_P) \in \Pi_{j,P}(\mu, M_P)$.

The strategy space of the agent is unaltered from the principal-agent game, aside from the addition of trembles. For every mechanism (μ, M_P), we require the probability α that the agent accepts its proposal to be no less than $1/j$. Additionally, we require the agent to tremble in their choices of actions. In particular, for every mechanism (μ, M_P) and principal action-recommendation pair (x, r), the agent's choice of action must be a distribution belonging to

$$\Delta_j(Y_j) = \left\{ \mathbf{y} \in \Delta(Y_j) : \mathbf{y}[y] \geq \frac{1}{j|Y_j|} \ \forall y \in Y_j \right\}.$$

A valid strategy for the agent in the j-th game is any pair $(\boldsymbol{\alpha}(\cdot), \boldsymbol{\beta}(\cdot))$ consisting of (1) a rule governing the probability of mechanism acceptance, $\boldsymbol{\alpha}(\cdot)$, satisfying $\boldsymbol{\alpha}(\mu, M_P) \geq 1/j$ for all $(\mu, M_P) \in \mathcal{M}_j$ and (2) a rule governing the agent's choice of actions $\boldsymbol{\beta}(\cdot)$ satisfying $\boldsymbol{\beta}(\mu, M_P) \in \Delta_j(Y_j)^{X_j \times T_j \times R_j}$ for all $(\mu, M_P) \in \mathcal{M}_j$.

In addition to the principal types and agent, we introduce a hypothetical player who determines the distribution over principal types. This player can choose any distribution that puts probability at least $1/k$ on every type. Formally, the strategy space of this player is $\{\lambda' \in \Delta(\Theta) : \lambda'(\theta) \geq 1/k \ \forall \theta \in \Theta\}$.

We now develop the payoffs of the various players for an arbitrary strategy profile ζ. For any $\theta \in \Theta$, let $\widetilde{U}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_P, \boldsymbol{\beta}_A)$ and $\widetilde{V}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_P, \boldsymbol{\beta}_A)$ be the unmodified expected payoffs to the principal and agent, respectively, when the principal's type is θ, the mechanism $(\mu, M_P) \in \mathcal{M}_j$ is proposed, the agent uses the acceptance probability rule $\boldsymbol{\alpha} \in [0,1]^{\mathcal{M}_j}$, and subsequent play is governed by the rules $\boldsymbol{\pi}_P$ and $\boldsymbol{\beta}_A$.

The agent's payoff is

$$V_j(\zeta) = \sum_{\theta \in \Theta} \lambda'(\theta) \left[\sum_{(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}^c_{j,\theta'})} \eta_\theta[(\mu, M_P)] \widetilde{V}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_\theta, \boldsymbol{\beta}_A) \right].$$

This is precisely the agent's total expected utility from play over mechanisms in $\mathcal{M} \setminus (\cup_{\theta \in \Theta} \mathcal{M}^c_{j,\theta})$. The payoff of the player who controls the distribution of principal types is $W_j(\zeta) = -V_j(\zeta)$, i.e. the negative of the agent's payoff. Thus, this player desires to minimize the agent's total expected utility from play over mechanisms in $\mathcal{M} \setminus (\cup_{\theta \in \Theta} \mathcal{M}^c_{j,\theta})$.

We require more notation to specify the payoffs of the principal types.

$$\begin{aligned}
\widehat{U}_j(\theta, \zeta) = & \sum_{(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}^c_{j,\theta'})} \eta_\theta[(\mu, M_P)] \widetilde{U}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_\theta, \boldsymbol{\beta}_A) \\
& + \sum_{(\mu_{\chi,\theta}, \{0\}) \in \mathcal{M}^c_{j,\theta}} \eta_\theta[(\mu_{\chi,\theta}, \{0\})] \sum_{x,t} \chi[(x,t)](u(\theta, x, y^*(\theta, x)) - t) \text{ and} \\
\widehat{V}_j(\theta, \zeta) = & \sum_{(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}^c_{j,\theta'})} \eta_\theta[(\mu, M_P)] \widetilde{V}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_\theta, \boldsymbol{\beta}_A) \\
& + \sum_{(\mu_{\chi,\theta}, \{0\}) \in \mathcal{M}^c_{j,\theta}} \eta_\theta[(\mu_{\chi,\theta}, \{0\})] \sum_{x,t} \chi[(x,t)](v(\theta, x, y^*(\theta, x)) + g(t))
\end{aligned}$$

would be the total expected utilities of the principal and agent, respectively, when the principal's type is θ, the principal follows the mechanism proposal rule η_θ, and the play that follows a mechanism proposal of $(\mu, M_P) \in \mathcal{M}_j$ proceeds as follows: For $(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}^c_{j,\theta'})$, play proceeds according to the rules α, $\boldsymbol{\pi}_P$, and $\boldsymbol{\beta}_A$; for $(\mu_{\chi,\theta'}, \{0\}) \in \mathcal{M}^c_{j,\theta'}$, the agent accepts with probability 1 and then takes action $y^*(\theta', x)$ after observing any $x \in X_j$. We will impose modifications to the payoffs of the principal types so that it is costly for θ to propose any $(\mu_{\chi,\theta}, \{0\}) \in \mathcal{M}^c_{j,\theta}$ whenever either some principal type $\theta' \neq \theta$ would prefer to propose $(\mu_{\chi,\theta}, \{0\})$ (and have the agent respond according to $y^*(\theta, x)$) to their outcome or the agent gets a low expected utility conditional on θ. Let $A > 2 \max_{(\theta,x,t,y)} |u(\theta, x, y) - t|$, and let $f_j : \mathbb{R} \to \mathbb{R}_+$ be the family of continuous functions given by $f_j(z) = \max\{0, A \min\{jz, 1\}\}$. Note that $f_j(z) = 0$ for all $z \leq 0$ and j, and $\lim_{j \to \infty} f_j(z) = A$ for all $z > 0$. Let $c_{j,\theta,\zeta} : \mathcal{M}^c_{j,\theta} \to \mathbb{R}_+$

be the "cost" function given by

$$c_{j,\theta,\zeta}(\mu_{\chi,\theta},\{0\}) = \sum_{\theta'\neq\theta} f_j\left(\sum_{x,t}\chi[x,t](u(\theta',x,y^*(\theta,x))-t) - \widehat{U}_j(\theta',\zeta)\right) + f_j\left(V_k^\dagger - \widehat{V}_j(\theta,\zeta)\right).$$

Note that $c_{j,\widetilde{\Theta},\zeta}(\mu_{\chi,\theta},\{0\}) \geq A$ if some principal type $\theta' \notin \widetilde{\Theta}$ would get a payoff from proposing $(\mu_{\chi,\theta},\{0\})$ that exceeds their payoff from ζ by $1/j$, while $c_{j,\widetilde{\Theta},\zeta}(\mu_{\chi,\theta},\{0\}) = 0$ if every principal type $\theta' \notin \widetilde{\Theta}$ gets a weakly higher payoff from ζ than they would by proposing $(\mu_{\chi,\theta},\{0\})$. We set the payoff of θ from the strategy profile ζ in the j-th game to be

$$U_j(\theta,\zeta) = \widehat{U}_j(\theta,\zeta) - \sum_{(\mu_{\chi,\theta},\{0\})\in\mathcal{M}^c_{j,\theta}} \eta_\theta[(\mu_{\chi,\theta},\{0\})]\left(c_{j,\theta,\zeta}(\mu_{\chi,\theta},\{0\}) - \frac{1}{k}\right).$$

The important feature of the cost terms is that θ would never want to propose a $(\mu_{\chi,\theta},\{0\}) \in \mathcal{M}^c_{j,\theta}$ if either $\sum_{x,t}\chi[x,t](u(\theta',x,y^*(\theta,x))-t) \geq \widehat{U}_j(\theta',\zeta)+1/j$ for some $\theta'\neq\theta$ or $\widehat{V}_j(\theta,\zeta) \leq V_k^\dagger - 1/j$. On the other hand, if $\sum_{x,t}\chi[x,t](u(\theta',x,y^*(\theta,x))-t) \leq \widehat{U}_j(\theta',\zeta)$ holds for all $\theta'\neq\theta$ and $\widehat{V}_j(\theta,\zeta) \geq V_k^\dagger$, then the artificial cost from proposing $(\mu_{\chi,\theta},\{0\})$ is 0 for θ. In this case, θ would want to propose such a mechanism (if the agent responded according to $y^*(\theta,x)$) whenever they would get a higher payoff from it than from the outcome under ζ. ■

Construction of Limit Outcomes and Distributions over Principal Types. Fixing $k \in \mathbb{N}$, standard arguments show that the j-th game has a Nash equilibrium. Let $\lambda_{j,k}$ be the distribution over the principal's type induced by a Nash equilibrium of the j-th game. For the same Nash equilibrium, let $p_{j,k} \in \Delta(\Theta\times X\times T\times Y)$ be the outcome induced by the corresponding mechanism proposal strategies used by the principal types and the following continuation play for each mechanism: For any $(\mu,M_P) \in \mathcal{M}_j \setminus (\cup_{\theta\in\Theta}\mathcal{M}^c_{j,\theta})$, the principal types and agent play as they do in the Nash equilibrium, i.e. θ plays according to $\boldsymbol{\pi}_\theta(\mu,M_P)$ while the agent accepts the mechanism with probability $\boldsymbol{\alpha}(\mu,M_P)$ and then plays according to $\boldsymbol{\beta}_A(\mu,M_P)$; for any $(\mu_{\chi,\theta},\{0\}) \in \mathcal{M}^c_{j,\theta}$, the agent accepts

with probability 1 and then plays $y^*(\theta, x)$ when they observe x. Suppose (by restricting attention to a convergent subsequence if necessary) that $\lim_{j\to\infty} p_{j,k} = p_k$ and $\lim_{j\to\infty} \lambda_{j,k} = \lambda_k$. Since $\text{marg}_\Theta p_{j,k} = \lambda_{j,k}$ and $\lambda_{j,k}(\theta) \geq 1/k$ hold for each $\theta \in \Theta$ and $j \in \mathbb{N}$, we have that $\text{marg}_\Theta p_k = \lambda_k$ and $\lambda_k(\theta) \geq 1/k$ for all $\theta \in \Theta$. ■

Proof of Lemma 5. Condition 1 of Lemma 5 holds by construction. The remainder of this proof shows that the other four conditions are satisfied.

To establish Condition 2, it suffices to show that $\mathbb{E}_{p_k}[v(\theta, x, y) + g(t)|\theta] \geq V_k^\dagger$ holds for all θ, for $V_k^\dagger$ defined in (OA 7), since $\lim_{k\to\infty} V_k^\dagger = 0$. To see that $\mathbb{E}_{p_k}[v(\theta, x, y) + g(t)|\theta] \geq V_k^\dagger$, consider the following two exhaustive possibilities: Case (1) in which the Nash equilibria inducing the $p_{j,k}$ outcomes put probabilities on θ proposing mechanisms in $\mathcal{M}_{j,\theta}^c$ that are uniformly bounded away from 0 for infinitely many j, and Case (2) in which the Nash equilibria inducing the $p_{j,k}$ outcomes put probabilities on θ proposing mechanisms in $\mathcal{M}_{j,\theta}^c$ that converge to 0 for some subsequence of j. Recall that the construction of the hypothetical games ensures that θ is willing to propose mechanisms in $\mathcal{M}_{j,\theta}^c$ only if the agent's conditional expected utility in the prevailing outcome given θ is no less than $V_k^\dagger$. Thus, in Case (1), it must be that $\mathbb{E}_{p_{j,k}}[v(\theta, x, y) + g(t)|\theta] \geq V_k^\dagger$ for infinitely many j, which guarantees $\mathbb{E}_{p_k}[v(\theta, x, y) + g(t)|\theta] \geq V_k^\dagger$ by continuity. For Case (2), suppose towards a contradiction that there is some $\varepsilon > 0$ such that $\mathbb{E}_{p_{j,k}}[v(\theta, x, y)+g(t)|\theta] < V_k^\dagger - \varepsilon$ holds along a subsequence of j for which the probability that θ proposes mechanisms in $\mathcal{M}_{j,\theta}^c$ converges to 0. By construction, this means that the distribution over the principal types is such that the agent's conditional expected utility from the play over mechanisms belonging to $\mathcal{M}_j \setminus (\cup_{\theta'\in\Theta}\mathcal{M}_{j,\theta'}^c)$ is less than $-1/k$ for sufficiently high j in the subsequence. However, the agent's conditional expected utility from the play over mechanisms belonging to $\mathcal{M}_j \setminus (\cup_{\theta'\in\Theta}\mathcal{M}_{j,\theta'}^c)$ cannot be uniformly bounded below 0 as $j \to \infty$. Thus, it must be that $\mathbb{E}_{p_k}[v(\theta, x, y)+g(t)|\theta] \geq V_k^\dagger$.

To see that Condition 3 holds, observe that whenever $\theta' \neq \theta$ is willing to play a mechanism in $\mathcal{M}_{j,\theta'}^c$, θ must (weakly) prefer to not play said mechanism given the

prevailing outcome. Moreover, since θ can always mimic the play of θ' in mechanisms in $\mathcal{M}_j \setminus \mathcal{M}^c_{j,\theta'}$, it follows that, θ must weakly prefer their conditional outcome under $p_{j,k}$ to that of θ' in the $j \to \infty$ limit, which gives Condition 3.

We establish Condition 4 by induction over θ_n, beginning with θ_N as the base case. Suppose towards a contradiction that $\mathbb{P}_{p_k}[x \neq x_o|\theta_N] > 0$ and $\mathbb{P}_{p_k}[y = y^*(\theta_N, x)|\theta_N, x \neq x_o] < 1$. Then, since it is never optimal for an agent to play any action strictly greater than $y^*(\theta_N, x)$ given an $x \neq x_o$, it must be that $\mathbb{P}_{p_k}[y > y^*(\theta_N, x)|\theta_N, x \neq x_o] = 0$ and $\mathbb{P}_{p_k}[y < y^*(\theta_N, x)|\theta_N, x \neq x_o] > 0$. Consider the distribution $\chi \in \Delta(X \times T)$ that is obtained from taking the conditional distribution of p_k given θ_N and shifting every t to $t + \mathbb{E}_{p_k}[u(\theta_N, x, y^*(\theta_N, x)) - u(\theta_N, x, y)|\theta_N]$. When the agent accepts a mechanism committing to χ and plays $y^*(\theta_N, x)$ in response to any x, θ_N obtains the same expected utility as they do under p_k while every other type obtains a weakly lower expected utility than under p_k. Moreover, as previously established, the agent's expected utility from p_k conditional on θ_N is no less than $V_k^\dagger$. Thus, the agent would obtain an expected utility that is weakly greater than $V_k^\dagger$ from accepting a proposal of χ by θ_N. So for sufficiently high j, the type θ_N principal can achieve a payoff in the j-th game that is uniformly bounded above their payoff from p_k by proposing some mechanism $(\mu_{\chi'_j,\theta}, \{0\})$ where $\chi'_j \in \Delta_j(X_j \times T_j)$ sufficiently closely approximates χ', but this contradicts the fact that their payoff should be no more than that under p_k in the $j \to \infty$ limit.

Since $\mathbb{P}_{p_k}[y = y^*(\theta_N, x)|\theta_N, x \neq x_o] = 1$, it follows from the fact that λ_k puts strictly positive probability on θ_{N-1} that $\mathbb{P}_{p_k}[y > y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] = 0$ (assuming $\mathbb{P}_{p_k}[x \neq x_o|\theta_{N-1}] > 0$ so that this conditional probability is even relevant). Therefore, if $\mathbb{P}_{p_k}[y = y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] < 1$, it must be that $\mathbb{P}_{p_k}[y < y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] > 0$. The same argument as for the θ_N case shows that this is not possible. Proceeding with this argument inductively by moving down the θ_n establishes that Condition 4 holds for all θ.

To see why Condition 5 holds, note that, in the $j \to \infty$ limit, every θ must get a weakly higher payoff from $p_{j,k}$ than $1/k$ less the payoff they would get from proposing any $(\mu, M_P) \in \mathcal{M}_j \setminus M^0_{j,\theta}$. Fix some $r \in R_j \setminus \Theta$. For any $\chi \in \Delta_j(X_j \times T_j)$,

every mechanism of the form $(\mu_{\chi,\theta'}, \{0\})$ for some $\theta' \in \Theta$ can be identified with $(\mu_{\chi,r}, \{0\}) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}^0_{j,\theta'})$. This means that, for every mechanism in $\mathcal{M}_j$, there is a corresponding outcome that occurs after either this mechanism is proposed in the equilibrium of the j-th game or, if the mechanism belongs to some $M^0_{j,\theta'}$, after the proposal of the mechanism in which the action recommendation θ' is replaced by r. Thus, in the $j \to \infty$ limit, every principal type must get a weakly higher payoff from $p_{j,k}$ than $1/k$ less the payoff they would get from proposing some mechanism in $\mathcal{M}_j$ if the subsequent play results in this outcome. Similar arguments to those in the proof of Lemma 2 then show that there is a sequential continuation equilibrium outcome after an arbitrary mechanism $(\mu, M_P) \in \mathcal{M}$ is proposed which gives every principal type a weakly lower payoff than they obtain from p_k. ■

OA.9.2 Proof of Lemma 6

Lemma 6. *In MCS environments, there are sequences of full-support distributions over the principal type $\{\lambda_k\}_{k\in\mathbb{N}}$ and outcomes $\{p_k\}_{k\in\mathbb{N}}$ such that*

1. *$marg_\Theta p_k = \lambda_k$ for all $k \in \mathbb{N}$,*
2. *$\liminf_{k\to\infty} \mathbb{E}_{p_k}[\alpha(v(\theta, x, y) + g(t))|\theta] \geq 0$ for all $\theta \in \Theta$,*
3. *$\mathbb{P}_{p_k}[U(\theta, p_k) \geq \alpha(u(\theta, x, y^*(\theta', x)) - t)|\theta', x, t, \alpha] = 1$ for all $\theta, \theta' \in \Theta$ and $k \in \mathbb{N}$,*
4. *$\mathbb{P}_{p_k}[y = y^*(\theta, x)|\theta, x \neq x_o] = 1$ for all $\theta \in \Theta$ and $k \in \mathbb{N}$, and*
5. *For each mechanism $(\mu, M_P) \in \mathcal{M}$ and $k \in \mathbb{N}$, there is a sequential continuation equilibrium after (μ, M_P) is proposed that gives every principal type a weakly lower payoff than p_k.*

Construction of Hypothetical Games. Let $\{X_j\}_{j\in\mathbb{N}}$, $\{T_j\}_{j\in\mathbb{N}}$, $\{Y_j\}_{j\in\mathbb{N}}$, $\{R_j\}_{j\in\mathbb{N}}$ be sequences of finite action, transfer, and recommendation sets such that $\lim_{j\to\infty} X_j = X$, $\lim_{j\to\infty} T_j = T$, $\lim_{j\to\infty} Y_j = Y$, and $\lim_{j\to\infty} R_j = R$. For a given $j \in \mathbb{N}_{++}$, consider

the set of mechanisms

$$\mathcal{M}_j = \Big\{(\mu, M_P) \in \mathcal{M} : \ (1)\ |M_P| \leq |X_j||T_j||R_j|,$$
$$(2)\ \forall m_P \in M_P,\ \exists x \in X_j, t \in T_j, r \in R_j \text{ s.t. } \mu((x,t), r|m_P) = 1\Big\}$$

that (1) have no more than $|X_j||T_j||R_j|$ principal messages and (2) are such that every principal message results in some principal-action-transfer-recommendation tuple that belongs to $X_j \times T_j \times R_j$. We suppose that, for all $j \in \mathbb{N}$, the recommendation space is strictly larger than the set of principal types, i.e. $|R_j| > |\Theta|$. For notational convenience, we will assume that the power set of principal types is in fact a strict subset of the recommendation spaces.

We now describe the strategy space of the type θ principal in the j-th game. Part of this player's choice is over which mechanisms to propose. We force θ to propose almost all mechanisms with positive probability. The exceptions are mechanisms which commit to some $(x,t) \in X_j \times T_j$ as the principal action and some $\theta' \neq \theta$ as the recommendation received by the agent; θ is required to propose these mechanisms with 0 probability. Formally, let $\mathcal{M}^c_{j,\theta'} = \{(\delta_{((x,t),\theta')}, \{0\}) : (x,t) \in X_j \times T_j,\ \theta' \neq \theta\}$ be the set of mechanisms in the j-th game that commit to some $(x,t) \in X_j \times T_j$ as the distribution over principal actions and θ' as the recommendation received by the agent. Additionally, let $\mathcal{M}^0_{j,\theta} = \cup_{\theta' \neq \theta} \mathcal{M}^c_{j,\theta'}$. The distribution over mechanism proposals used by θ must belong to

$$\Delta_{j,\theta}(\mathcal{M}_j) = \Big\{\wp \in \Delta(\mathcal{M}_j) : \ (1)\ \wp[(\mu, M_P)] \geq \frac{1}{j|\mathcal{M}_j|}\ \forall (\mu, M_P) \in \mathcal{M}_j \setminus \mathcal{M}^0_{j,\theta},$$
$$(2)\ \wp[(\mu, M_P)] = 0\ \forall (\mu, M_P) \in \mathcal{M}^0_{j,\theta}\Big\}.$$

Moreover, when a given mechanism is accepted, we force θ to tremble and play every message in the mechanism with positive probability. Formally, the distribution over

messages used by θ when mechanism (μ, M_P) is accepted must belong to

$$\Pi_{j,P}(\mu, M_P) = \Big\{\pi_P \in \Delta(M_P) : \pi_P[m_P] \geq \frac{1}{j|M_P|} \ \forall m_P \in M_P\Big\}.$$

A valid strategy for θ in the j-th game is any pair $(\eta_\theta, \boldsymbol{\pi}_\theta(\cdot))$ consisting of a $\eta_\theta \in \Delta_{j,\theta}(\mathcal{M}_j)$ and a rule $\boldsymbol{\pi}_\theta(\cdot)$ for how to play when an arbitrary mechanism is accepted that satisfies $\boldsymbol{\pi}_\theta(\mu, M_P) \in \Pi_{j,P}(\mu, M_P)$.

The strategy space of the agent is unaltered from the principal-agent game, aside from the addition of trembles. For every mechanism (μ, M_P), we require the probability α that the agent accepts its proposal to be no less than $1/j$. Additionally, we require the agent to tremble in their choices of actions. In particular, for every mechanism (μ, M_P) and principal action-recommendation pair (x, r), the agent's choice of action must be a distribution belonging to

$$\Delta_j(Y_j) = \left\{\mathbf{y} \in \Delta(Y_j) : \mathbf{y}[y] \geq \frac{1}{j|Y_j|} \ \forall y \in Y_j\right\}.$$

A valid strategy for the agent in the j-th game is any pair $(\boldsymbol{\alpha}(\cdot), \boldsymbol{\beta}(\cdot))$ consisting of (1) a rule governing the probability of mechanism acceptance, $\boldsymbol{\alpha}(\cdot)$, satisfying $\boldsymbol{\alpha}(\mu, M_P) \geq 1/j$ for all $(\mu, M_P) \in \mathcal{M}_j$ and (2) a rule governing the agent's choice of actions $\boldsymbol{\beta}(\cdot)$ satisfying $\boldsymbol{\beta}(\mu, M_P) \in \Delta_j(Y_j)^{X_j \times T_j \times R_j}$ for all $(\mu, M_P) \in \mathcal{M}_j$.

In addition to the principal types and agent, we introduce a hypothetical player who determines the distribution over principal types. This player can choose any distribution that puts probability at least $1/k$ on every type. Formally, the strategy space of this player is $\{\lambda' \in \Delta(\Theta) : \lambda'(\theta) \geq 1/k \ \forall \theta \in \Theta\}$.

We now develop the payoffs of the various players for an arbitrary strategy profile ζ. For any $\theta \in \Theta$, let $\widetilde{U}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_P, \boldsymbol{\beta}_A)$ and $\widetilde{V}_j(\theta, \mu, M_P, \boldsymbol{\alpha}, \boldsymbol{\pi}_P, \boldsymbol{\beta}_A)$ be the unmodified expected payoffs to the principal and agent, respectively, when the principal's type is θ, the mechanism $(\mu, M_P) \in \mathcal{M}_j$ is proposed, the agent uses the acceptance probability rule $\boldsymbol{\alpha} \in [0,1]^{\mathcal{M}_j}$, and subsequent play is governed by the rules $\boldsymbol{\pi}_P$ and $\boldsymbol{\beta}_A$.

The agent's payoff is

$$V_j(\zeta) = \sum_{\theta\in\Theta} \lambda'(\theta) \left[\sum_{(\mu,M_P)\in\mathcal{M}_j\setminus(\cup_{\theta'\in\Theta}\mathcal{M}^c_{j,\theta'})} \eta_\theta[(\mu,M_P)]\widetilde{V}_j(\theta,\mu,M_P,\boldsymbol{\alpha},\boldsymbol{\pi}_\theta,\boldsymbol{\beta}_A) \right].$$

This is precisely the agent's total expected utility from play over mechanisms in $\mathcal{M} \setminus (\cup_{\theta\in\Theta}\mathcal{M}^c_{j,\theta})$. The payoff of the player who controls the distribution of principal types is $W_j(\zeta) = -V_j(\zeta)$, i.e. the negative of the agent's payoff. Thus, this player desires to minimize the agent's total expected utility from play over mechanisms in $\mathcal{M} \setminus (\cup_{\theta\in\Theta}\mathcal{M}^c_{j,\theta})$.

We require more notation to specify the payoffs of the principal types.

$$\begin{aligned}
\widehat{U}_j(\theta,\zeta) = &\sum_{(\mu,M_P)\in\mathcal{M}_j\setminus(\cup_{\theta'\in\Theta}\mathcal{M}^c_{j,\theta'})} \eta_\theta[(\mu,M_P)]\widetilde{U}_j(\theta,\mu,M_P,\boldsymbol{\alpha},\boldsymbol{\pi}_\theta,\boldsymbol{\beta}_A) \\
&+ \sum_{(\delta_{((x,t),\theta)},\{0\})\in\mathcal{M}^c_{j,\theta}} \eta_\theta[(\delta_{((x,t),\theta)},\{0\})](u(\theta,x,y^*(\theta,x)) - t) \text{ and} \\
\widehat{V}_j(\theta,\zeta) = &\sum_{(\mu,M_P)\in\mathcal{M}_j\setminus(\cup_{\theta'\in\Theta}\mathcal{M}^c_{j,\theta'})} \eta_\theta[(\mu,M_P)]\widetilde{V}_j(\theta,\mu,M_P,\boldsymbol{\alpha},\boldsymbol{\pi}_\theta,\boldsymbol{\beta}_A) \\
&+ \sum_{(\delta_{((x,t),\theta)},\{0\})\in\mathcal{M}^c_{j,\theta}} \eta_\theta[(\delta_{((x,t),\theta)},\{0\})](v(\theta,x,y^*(\theta,x)) + g(t))
\end{aligned}$$

would be the total expected utilities of the principal and agent, respectively, when the principal's type is θ, the principal follows the mechanism proposal rule η_θ, and the play that follows a mechanism proposal of $(\mu,M_P) \in \mathcal{M}_j$ proceeds as follows: For $(\mu,M_P) \in \mathcal{M}_j \setminus (\cup_{\theta'\in\Theta}\mathcal{M}^c_{j,\theta'})$, play proceeds according to the rules α, $\boldsymbol{\pi}_P$, and $\boldsymbol{\beta}_A$; for $(\delta_{((x,t),\theta')},\{0\}) \in \mathcal{M}^c_{j,\theta'}$, the agent accepts with probability 1 and then takes action $y^*(\theta',x)$. We will impose modifications to the payoffs of the principal types so that it is costly for θ to propose any $(\delta_{((x,t),\theta)},\{0\}) \in \mathcal{M}^c_{j,\theta}$ whenever either some principal type $\theta' \neq \theta$ would prefer to propose $(\delta_{((x,t),\theta)},\{0\})$ (and have the agent respond according to $y^*(\theta,x)$) to their outcome or the agent gets a low expected utility conditional on θ. Let $A > 2\max_{(\theta,x,t,y)}|u(\theta,x,y) - t|$, and let $f_j : \mathbb{R} \to \mathbb{R}_+$ be the family of continuous

functions given by $f_j(z) = \max\{0, A\min\{jz, 1\}\}$. Note that $f_j(z) = 0$ for all $z \leq 0$ and j, and $\lim_{j\to\infty} f_j(z) = A$ for all $z > 0$. Let $c_{j,\theta,\zeta} : \mathcal{M}^c_{j,\theta} \to \mathbb{R}_+$ be the "cost" function given by

$$c_{j,\theta,\zeta}(\mu_{\chi,\theta}, \{0\}) = \sum_{\theta' \neq \theta} f_j\left(u(\theta', x, y^*(\theta, x)) - t - \widehat{U}_j(\theta', \zeta)\right) + f_j\left(V_k^\dagger - \widehat{V}_j(\theta, \zeta)\right).$$

Note that $c_{j,\widetilde{\Theta},\zeta}(\delta_{((x,t),\theta)}, \{0\}) \geq A$ if some principal type $\theta' \notin \widetilde{\Theta}$ would get a payoff from proposing $(\delta_{((x,t),\theta)}, \{0\})$ that exceeds their payoff from ζ by $1/j$, while $c_{j,\widetilde{\Theta},\zeta}(\delta_{((x,t),\theta)}, \{0\}) = 0$ if every principal type $\theta' \notin \widetilde{\Theta}$ gets a weakly higher payoff from ζ than they would by proposing $(\delta_{((x,t),\theta)}, \{0\})$. We set the payoff of θ from the strategy profile ζ in the j-th game to be

$$U_j(\theta, \zeta) = \widehat{U}_j(\theta, \zeta) - \sum_{(\delta_{((x,t),\theta)}, \{0\}) \in \mathcal{M}^c_{j,\theta}} \eta_\theta[(\delta_{((x,t),\theta)}, \{0\})] \left(c_{j,\theta,\zeta}(\delta_{((x,t),\theta)}, \{0\}) - \frac{1}{k}\right).$$

The important feature of the cost terms is that θ would never want to propose a $(\delta_{((x,t),\theta)}, \{0\}) \in \mathcal{M}^c_{j,\theta}$ if either $u(\theta', x, y^*(\theta, x)) - t \geq \widehat{U}_j(\theta', \zeta) + 1/j$ for some $\theta' \neq \theta$ or $\widehat{V}_j(\theta, \zeta) \leq V_k^\dagger - 1/j$. On the other hand, if $u(\theta', x, y^*(\theta, x)) - t \leq \widehat{U}_j(\theta', \zeta)$ holds for all $\theta' \neq \theta$ and $\widehat{V}_j(\theta, \zeta) \geq V_k^\dagger$, then the artificial cost from proposing $(\delta_{((x,t),\theta)}, \{0\})$ is 0 for θ. In this case, θ would want to propose such a mechanism (if the agent responded according to $y^*(\theta, x)$) whenever they would get a higher payoff from it than from the outcome under ζ. ■

Construction of Limit Outcomes and Distributions over Principal Types. Fixing $k \in \mathbb{N}$, standard arguments show that the j-th game has a Nash equilibrium. Let $\lambda_{j,k}$ be the distribution over the principal's type induced by a Nash equilibrium of the j-th game. For the same Nash equilibrium, let $p_{j,k} \in \Delta(\Theta \times \mathcal{M}_j \times [0,1] \times X \times T \times Y)$ be the outcome induced by the corresponding mechanism proposal strategies used by the principal types and the following continuation play for each mechanism: For any $(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta\in\Theta} \mathcal{M}^c_{j,\theta})$, the principal types and agent play as they do in the

Nash equilibrium, i.e. θ plays according to $\boldsymbol{\pi}_\theta(\mu, M_P)$ while the agent accepts the mechanism with probability $\boldsymbol{\alpha}(\mu, M_P)$ and then plays according to $\boldsymbol{\beta}_A(\mu, M_P)$; for any $(\delta_{((x,t),\theta)}, \{0\}) \in \mathcal{M}^c_{j,\theta}$, the agent accepts with probability 1 and then plays $y^*(\theta, x)$ when they observe x. Suppose (by restricting attention to a convergent subsequence if necessary) that $\lim_{j\to\infty} p_{j,k} = p_k$ and $\lim_{j\to\infty} \lambda_{j,k} = \lambda_k$. Since $\text{marg}_\Theta p_{j,k} = \lambda_{j,k}$ and $\lambda_{j,k}(\theta) \geq 1/k$ hold for each $\theta \in \Theta$ and $j \in \mathbb{N}$, we have that $\text{marg}_\Theta p_k = \lambda_k$ and $\lambda_k(\theta) \geq 1/k$ for all $\theta \in \Theta$. ■

Proof of Lemma 6. Precisely the same arguments as in the proof of Lemma 5 shows that Conditions 1, 2, and 5 hold. The remainder of this proof shows that the other two conditions are satisfied.

To see that Condition 3 holds, observe that whenever $\theta' \neq \theta$ is willing to play a mechanism in $\mathcal{M}^c_{j,\theta'}$, θ must (weakly) prefer to not play said mechanism given the prevailing outcome. Moreover, since θ can always mimic the play of θ' in mechanisms in $\mathcal{M}_j \setminus \mathcal{M}^c_{j,\theta'}$, it follows that, for all $\varepsilon > 0$, θ must get a weakly higher payoff from their conditional outcome under $p_{j,k}$ than their payoff from the conditional outcome given θ', x, t, and α for almost all $(x,t) \in X_j \times T_j$ and $\alpha \in [0,1]$ in the $j \to \infty$ limit, which gives Condition 3.

We establish Condition 4 by induction over θ_n, beginning with θ_N as the base case. Suppose towards a contradiction that $\mathbb{P}_{p_k}[x \neq x_o|\theta_N] > 0$ and $\mathbb{P}_{p_k}[y = y^*(\theta_N, x)|\theta_N, x \neq x_o] < 1$. Then, since it is never optimal for an agent to play any action strictly greater than $y^*(\theta_N, x)$ given an $x \neq x_o$, it must be that $\mathbb{P}_{p_k}[y > y^*(\theta_N, x)|\theta_N, x \neq x_o] = 0$ and $\mathbb{P}_{p_k}[y < y^*(\theta_N, x)|\theta_N, x \neq x_o] > 0$. Take some $(x,t) \in X_j \times T_j$ and $\alpha \in [0,1]$ such that $\mathbb{E}_{p_k}[\alpha(u(\theta_N, x, y) - t)|\theta_N, x, t, \alpha] = U(\theta_N, p_k)$, $\mathbb{E}_{p_k}[\alpha(u(\theta, x, y) - t)|\theta_N, x, t, \alpha] \leq U(\theta, p_k)$ for all $\theta \neq \theta_N$, and $\mathbb{E}_{p_k}[\alpha(v(\theta_N, x, y) + g(t))|\theta_N, x, t, \alpha] \geq V_k^\dagger$. (This is possible because $\mathbb{E}_{p_k}[v(\theta, x, y) + g(t)|\theta] \geq V_k^\dagger$ holds for all θ and the $V_k^\dagger$ defined in (OA 7), which can be shown as in the proof of Lemma 5.) Consider shifting t up to $\tilde{t} = \alpha t + u(\theta_N, x, y^*(\theta_N, x)) - \mathbb{E}_{p_k}[\alpha u(\theta_N, x, y)|\theta_N, x, t, \alpha]$. When the agent accepts a mechanism committing to $(x, \tilde{t})$ and plays $y^*(\theta_N, x)$ in response, θ_N obtains the same expected

utility as they do under p_k while every other type obtains a weakly lower expected utility than under p_k. Moreover, the agent would obtain an expected utility that is weakly greater than $V_k^\dagger$ from accepting a proposal of $(x, \tilde{t})$ by θ_N. So for sufficiently high j, the type θ_N principal can achieve a payoff in the j-th game that is uniformly bounded above their payoff from p_k by proposing some mechanism $(\delta_{((x',t'),\theta_N)}, \{0\})$ where $(x', t') \in X_j \times T_j$ sufficiently closely approximates $(x, \tilde{t})$, but this contradicts the fact that their payoff should be no more than that under p_k in the $j \to \infty$ limit.

Since $\mathbb{P}_{p_k}[y = y^*(\theta_N, x)|\theta_N, x \neq x_o] = 1$, it follows from the fact that λ_k puts strictly positive probability on θ_{N-1} that $\mathbb{P}_{p_k}[y > y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] = 0$ (assuming $\mathbb{P}_{p_k}[x \neq x_o|\theta_{N-1}] > 0$ so that this conditional probability is even relevant). Therefore, if $\mathbb{P}_{p_k}[y = y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] < 1$, it must be that $\mathbb{P}_{p_k}[y < y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] > 0$. The same argument as for the θ_N case shows that this is not possible. Proceeding with this argument inductively by moving down the θ_n establishes that Condition 4 holds for all θ. ■

OA.10 Generalization of Proposition 5

Proposition OA 4. *Suppose the environment is MCS with definite gains and that, for every $\tilde{\lambda} \in \Delta(\Theta)$ and $x \neq x_o$, either quasi-strictness holds at x, or there exists a sequence $\{x_i\}$ converging to x such that $y^*(\tilde{\lambda}, x_i)$ converges to $y^*(\tilde{\lambda}, x)$, quasi-strictness holds at each x_i, and either one of the following conditions hold:*

1. *(a) $u(\theta, x, y^*(\tilde{\lambda}, x))$ is constant in θ.*
   (b) $u(\theta, x_i, y^*(\tilde{\lambda}, x_i)) > u(\theta, x, y^*(\tilde{\lambda}, x))$ for all i.
   (c) $v(\theta, x_i, y^*(\tilde{\lambda}, x_i)) > v(\theta, x, y^*(\tilde{\lambda}, x))$ for all i.
2. *(a) $u(\theta, x, y^*(\tilde{\lambda}, x))$ is constant in θ.*
   (b) $v(\theta, x, y^*(\tilde{\lambda}, x))$ is strictly increasing in θ.

Then payoff-plausibility selects the least-cost separating outcomes when contracts must be explicit.

Observe that the sufficient conditions cover the firm-employee example (the issues with $s = 0$ are handled by Condition 3′ in particular, while Condition 3″ takes care of $s = 1$), as well as the quasi-strict environments of Definition 10.

Proof. We first establish that every contracting equilibrium outcome that is payoff-plausible must be separating. Let p be a contracting equilibrium outcome with pooling, and let $\overline{\theta}$ be the highest type that does not fully separate. There must be some $x \in X$, $t \in \mathbb{R}$, $\tilde{\lambda} \in \Delta(\Theta)$, and $\alpha \in [0, 1]$ such that $U(\overline{\theta}, p) = \alpha(u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) - t)$, $U(\theta, p) \leq \alpha(u(\theta, x, y^*(\tilde{\lambda}, x)) - t)$ for all $\theta \neq \overline{\theta}$, acceptance probability α is optimal for an agent with belief $\tilde{\lambda}$ facing a contract committing to (x, t), and $\tilde{\lambda}$ is strictly lower than $\delta_{\overline{\theta}}$ under FOSD. Since there are definite gains, $U(\overline{\theta}, p) > 0$, so $u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) - t > 0$ and $\alpha > 0$. Because $\alpha > 0$, we have that $v(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) + g(t) \geq 0$.

We now analyze two cases depending on whether $(x, \tilde{\lambda})$ satisfies Condition 3 or it satisfies either of the 3′ or 3″ conditions.

Case 1: Condition 3 holds for $(x, \tilde{\lambda})$. Consider $t' = \alpha t + u(\overline{\theta}, x, y^*(\overline{\theta}, x)) - \alpha u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) > t$. Observe that $u(\overline{\theta}, x, y^*(\overline{\theta}, x)) - t' = U(\overline{\theta}, p)$, $u(\theta, x, y^*(\overline{\theta}, x)) - t' > U(\theta, p)$ for all $\theta < \overline{\theta}$, and $v(\overline{\theta}, x, y^*(\overline{\theta}, x)) + g(t') > 0$. Thus, (x, t') strictly satisfies the constraints in the type $\overline{\theta}$ principal's plausibility threshold problem given by (2). Moreover, when the agent responds to a contract proposing (x, t') under the belief that $\theta = \overline{\theta}$, the type $\overline{\theta}$ principal obtains a payoff equal to that they get from p. The constraints would continue to be satisfied if t' were decreased slightly, and type $\overline{\theta}$ would get a strictly higher payoff than $U(\overline{\theta}, p)$, which means that p is not payoff-plausible.

Case 2: Condition 3′ or 3″ holds for $(x, \tilde{\lambda})$. Consider $t'_i = \alpha t + u(\overline{\theta}, x_i, y^*(\overline{\theta}, x_i)) - \alpha u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) > t$. By construction, $u(\overline{\theta}, x_i, y^*(\overline{\theta}, x_i)) - t'_i = U(\overline{\theta}, p)$, $u(\theta, x_i, y^*(\overline{\theta}, x_i)) - t'_i < U(\theta, p)$ for all $\theta < \overline{\theta}$, and $v(\overline{\theta}, x, y^*(\overline{\theta}, x)) + g(t'_i) > 0$. A similar argument to that in Case 1 then shows that p can not be payoff-plausible.

Having shown that every payoff-plausible contracting equilibrium is separating, we conclude the proof by observing that payoff-plausibility requires that every principal type obtain a weakly higher payoff than their least-cost separating payoff. It follows

that payoff-plausibility selects the least-cost separating outcomes. ■

OA.11 Proof of Proposition 7.1

Proof. Consider an arbitrary mechanism (μ, M_P, M_A). Throughout the proof, let $U(\theta, m_P, m_A) \equiv \mathbb{E}_{\mu(m_P,m_A)}[U(\theta, x)]$ and $V(\theta, m_P, m_A) \equiv \mathbb{E}_{\mu(m_P,m_A)}[V(\theta, x)]$ denote the expected utility of the principal and agent, respectively, when the principal's type is θ and x is drawn according to $\mu(m_P, m_A)$.

Let $\Psi : \Delta(M_P)^\Theta \times \Delta(M_A) \rightrightarrows \Delta(\Theta)$ be the correspondence given by

$$\Psi(\pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A) = \Delta(\arg\min_{\theta \in \Theta} \mathbb{E}_{\pi_\theta \times \pi_A}[V(\theta, m_P, m_A)]).$$

Ψ maps profiles of principal and agent behavior strategies into beliefs that put support only on the principal types that minimize the agent's conditional expected utility.

For every $\theta \in \Theta$, let $\Pi_\theta : \Delta(M_A) \rightrightarrows \Delta(M_P)$ be the correspondence given by

$$\Pi_\theta(\pi_A) = \Delta(\arg\max_{m_P \in M_P} \mathbb{E}_{\pi_A}[U(\theta, m_P, m_A)]).$$

Π_θ maps agent behavior strategies into the corresponding optimal behavior strategies for the type θ principal in the subgame in which (μ, M_P, M_A) has been accepted.

Let $\Pi_A : \Delta(\Theta) \times \Delta(M_P)^\Theta \rightrightarrows \Delta(M_A)$ be the correspondence given by

$$\Pi_A(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}) = \Delta(\arg\max_{m_A \in M_A} \mathbb{E}_{\tilde{\lambda}}[\mathbb{E}_{\pi_\theta}[V(\theta, m_P, m_A)]]).$$

Π_A maps profiles of beliefs over the principal's type and behavior strategies into the corresponding optimal behavior strategies for the agent in the subgame in which (μ, M_P, M_A) has been accepted.

For every $j \in \mathbb{N}$, let $\Phi_j : \Delta(\Theta) \times \Delta(M_P)^\Theta \times \Delta(M_A) \rightrightarrows \Delta(\Theta) \times \Delta(M_P)^\Theta \times \Delta(M_A)$

be the correspondence given by

$$
\begin{aligned}
&\Phi_j(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A) \\
=&\{(\lambda', \pi'_{\theta_1}, ..., \pi'_{\theta_N}, \pi'_A) \in \Delta(\Theta) \times \Delta(M_P)^\Theta \times \Delta(M_A) : \\
&(1)\ \exists \lambda'' \in \Psi(\pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A) \text{ s.t. } \lambda'(\theta) = \frac{1}{j+1}\frac{1}{|\Theta|} + \frac{j}{j+1}\lambda''(\theta)\ \forall \theta \in \Theta, \\
&(2)\ \pi'_\theta \in \Pi_\theta(\pi_A)\ \forall \theta \in \Theta, \text{ and} \\
&(3)\ \exists \pi'_A \in \Pi_A(\tilde{\lambda}, \pi_{\theta_1}, ..., \pi_{\theta_N})\}.
\end{aligned}
$$

By construction, Φ_j is everywhere non-empty-valued, compact-valued, convex-valued, and upper hemicontinuous, and $\Delta(\Theta) \times \Delta(M_P)^\Theta \times \Delta(M_A)$ is a compact and convex subset of a Euclidean space. Thus, by Kakutani's fixed point theorem, some $(\lambda_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A})$ satisfies $(\lambda_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A}) \in \Phi_j(\lambda_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A})$. Since $\Delta(\Theta)\times\Delta(M_P)^\Theta\times\Delta(M_A)$ is sequentially compact, there is a limit point $(\lambda^*, \pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \pi^*_A)$ of the sequence $\{(\lambda_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A})\}_{j\in\mathbb{N}}$. Suppose (by restricting attention to a convergent subsequence if necessary) that $\lim_{j\to\infty}(\lambda_j, \pi_{j,\theta_1}, ..., \pi_{j,\theta_N}, \pi_{j,A}) = (\lambda^*, \pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \pi^*_A)$. Standard arguments show that $(\pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \pi^*_A)$ is a sequential continuation equilibrium when mechanism (μ, M_P, M_A) is accepted given belief λ^*.

We conclude by arguing that, when the principal types receive their principal-optimal safe payoffs, there is a sequential continuation equilibrium that deters every principal type from proposing (μ, M_P, M_A). Suppose first that $\mathbb{E}_{\pi^*_\theta \times \pi^*_A}[V(\theta, m_P, m_A)] \leq 0$ for some θ. Then it must be that λ^* puts positive probability only on those types for which the conditional expected utility of the agent is weakly less than their outside option utility. This means that $\mathbb{E}_{\lambda^*}[\mathbb{E}_{\pi_\theta \times \pi_A}[V(\theta, m_P, m_A)]] \leq 0$, so it is a sequential continuation equilibrium outcome for the agent to reject (μ, M_P, M_A) when offered. Such an outcome deters every principal type from proposing (μ, M_P, M_A). Now suppose that $\mathbb{E}_{\pi^*_\theta \times \pi^*_A}[V(\theta, m_P, m_A)] > 0$ for all $\theta \in \Theta$. Since $(\pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \pi^*_A)$ is a sequential continuation equilibrium when mechanism (μ, M_P, M_A) is accepted, incentive compatibility of the principal types implies that $\mathbb{E}_{\pi^*_\theta \times \pi^*_A}[U(\theta, m_P, m_A)] \geq \mathbb{E}_{\pi^*_{\theta'} \times \pi^*_A}[U(\theta, m_P, m_A)]$ for all

$\theta, \theta' \in \Theta$. Thus, $(\pi^*_{\theta_1}, ..., \pi^*_{\theta_N}, \pi^*_A)$ induces an safe allocation, which means that every principal type obtains an expected utility from proposing (μ, M_P, M_A) that is weakly lower than their principal-optimal safe payoff. ■

OA.12 Contracting Equilibrium Payoffs Outside of MCS Environments

Outside of MCS environments, payoff-plausibility is not defined, and so does not eliminate contracting equilibria that fail to principal-payoff-dominate the principal-optimal safe outcomes in non-MCS environments. Despite this, it may still be reasonable to expect the principal types to achieve at least their payoffs from the principal-optimal safe outcomes, especially when a principal-optimal safe outcome can be approximated by strictly safe outcomes. Indeed, suppose the principal proposed a direct mechanism corresponding to a strictly safe outcome and told the agent that they would report their type truthfully should the mechanism be accepted. Then it would be optimal for the agent, assuming they believed the principal's claim, to accept the offer and obediently follow any action recommendation regardless of their beliefs about the principal's type. To the extent that such communication is focal, equilibria in which any principal type receives a lower payoff than in the principal-optimal safe outcome seem unlikely to arise. The following proposition shows that there are always equilibria in which every principal type achieves a weakly higher payoff than in the principal-optimal safe outcomes.

Proposition OA 5. *With or without moral hazard, there are always contracting equilibrium outcomes that principal-payoff-dominate the principal-optimal safe outcomes in both the general-mechanism and deterministic-mechanism games.*

We handle the proof for the general-mechanism game. An analogous argument proves it for the deterministic-mechanism game.

Proof of Proposition OA 5 for the General-Mechanism Game. Consider an alternate principal-agent game where the principal has the option to forgo proposing any of the usual mechanisms and can instead unilaterally implement an alternative "outside option" x'_o that results in the same payoffs as a principal-optimal safe outcome. Formally, this game proceeds as follows: The principal observes their type θ, and either chooses x'_o or proposes a mechanism (μ, M_P) to the agent. If the principal chooses x'_o, both the principal and agent receive their conditional expected utility from a principal-optimal safe outcome given the principal's type. If the principal proposes a mechanism to the agent, the game proceeds and the payoffs of the principal and agent are the same as in the standard principal-agent game.

Arguments that are almost identical to the proof of Theorem 1 imply the existence of a contracting equilibrium in this environment, which we denote by $p \in \Delta(\Theta \times (X \cup \{x'_o\}) \times Y)$. Standard arguments show that this outcome is incentive compatible in the original principal-agent game.

Let $q^* = \Delta(\Theta \times X \times Y)$ be a principal-optimal safe outcome, and for each $\theta \in \Theta$, let $q^*(\theta) \in \Delta(X \times Y)$ be conditional distribution of q^* given type θ. Consider the direct mechanism (μ^*, Θ) where μ^* is given by $\mu^*(\theta) = \mathbb{P}_p[\{\theta\} \times X \times Y]\text{marg}_{\{\theta\}\times X\times Y} p + \mathbb{P}_p[\{\theta\} \times \{x'_o\} \times Y]q^*(\theta)$. This mechanism maps the principal type into the distributions over principal action and recommendation pairs that are identical to outcome p, except that instances of x'_o are replaced by the q^* allocation corresponding to the principal's type. By construction, this mechanism is incentive compatible, individually rational, and results in each type obtaining a weakly higher expected utility than their principal-optimal safe payoff. Additionally, because p is a contracting equilibrium outcome in the alternate principal-agent game defined earlier, there is a sequential continuation equilibrium after any mechanism is proposed that gives each principal type a weakly lower payoff than they obtain from proposing (μ^*, Θ). We conclude that (μ^*, Θ) corresponds to a contracting equilibrium outcome that principal-payoff-dominates the principal-optimal safe mechanism. ■

OA.13 Communication-Based Refinements

OA.13.1 Definitions for the General-Mechanism Game

Robust Neologism Proofness: We now formally develop an adaptation of RNP for our informed principal setting. For every $\widetilde{\Theta} \subseteq \Theta$, let $B(\widetilde{\Theta})$ be the set of agent action rules taking principal actions into agent responses that are best responses to some fixed belief supported on $\widetilde{\Theta}$:

$$B(\widetilde{\Theta}) = \{\beta \in \Delta(Y)^X : \beta \text{ is measurable, and } \exists \tilde{\lambda} \in \Delta(\widetilde{\Theta}) \text{ s.t. } \beta(x) \in \Delta(\arg\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]) \ \forall x \in X\}.$$

Also, let $U(\theta, p) \equiv \mathbb{E}_p[U(\theta, x, y)|\theta]$ denote the expected utility of the type $\theta \in \Theta$ principal from outcome $p \in \Delta(\Theta \times X \times Y)$.

Definition OA 1. *Contracting equilibrium outcome p has a* **credible robust neologism** *if there exists some $\chi \in \Delta(X)$ and non-empty subset of principal types $\widetilde{\Theta} \subseteq \Theta$ such that*

1. $\min_{\tilde{\lambda} \in \Delta(\widetilde{\Theta})} \mathbb{E}_\chi[\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]] > 0$,
2. $\min_{\beta \in B(\widetilde{\Theta})} \mathbb{E}_\chi[\mathbb{E}_{\beta(x)}[U(\theta, x, y)]] > U(\theta, p)$ *for some* $\theta \in \widetilde{\Theta}$, *and*
3. $\max_{\beta \in B(\widetilde{\Theta})} \mathbb{E}_\chi[\mathbb{E}_{\beta(x)}[U(\theta', x, y)]] < U(\theta', p)$ *for all* $\theta' \notin \widetilde{\Theta}$.

The first condition says that the agent strictly prefers to accept a contract proposal in which the principal commits to χ for any belief about the principal's type supported on $\widetilde{\Theta}$.[1] The second condition says that there is some principal type in $\widetilde{\Theta}$ that would obtain a strictly higher payoff than they do from p by proposing χ, as long as the agent believes the principal's type belongs to $\widetilde{\Theta}$, while the third condition says that every type outside of $\widetilde{\Theta}$ would do strictly worse by proposing χ given such agent beliefs.

[1] In our formalism, the principal cannot directly propose a $\chi \in \Delta(X)$. However, they can propose a mechanism $(\mu_\chi, \{0\})$ in which the message space of both the principal and agent is empty, and the resulting distribution over principal actions is χ, i.e. $\text{marg}_X \mu_\chi = \chi$.

Definition OA 2. *A contracting equilibrium outcome is* **robust neologism proof (RNP)** *if it does not have a credible robust neologism.*

Strongly Justified Communication Equilibrium: For every $\widetilde{\Theta} \subseteq \Theta$ and distribution over principal actions $\chi \in \Delta(X)$, let $C(\widetilde{\Theta}, \chi)$ be the set of agent responses to a proposal of χ given by

$$\begin{aligned} C(\widetilde{\Theta}, \chi) = \{(\alpha, \beta) \in [0,1] \times B(\Theta) : \exists \tilde{\lambda} \in \Delta(\widetilde{\Theta}) \text{ s.t.} \\ (1)\ \beta(x) \in \Delta(\arg\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)])\ \forall x \in X, \\ (2)\ \alpha = 0 \text{ if } \mathbb{E}_{\chi}[\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]] < 0, \text{ and} \\ (3)\ \alpha = 1 \text{ if } \mathbb{E}_{\chi}[\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]] > 0\}. \end{aligned}$$

A given agent response consists of a probability $\alpha \in [0,1]$ of accepting the proposal and an action rule $\beta \in B(\Theta)$ governing the agent's play should they accept the proposal. Condition 1 ensures that β is optimal for some fixed belief $\tilde{\lambda}$ with support on $\widetilde{\Theta}$, while Conditions 2 and 3 say that the agent's decision of whether to accept the proposal is also optimal given belief $\tilde{\lambda}$. We let $\Gamma(\widetilde{\Theta}, \chi) \equiv \Delta(C(\widetilde{\Theta}, \chi))$ be the set of distributions over all such agent responses.

Fixing $\chi \in \Delta(X)$ and outcome $p \in \Delta(\Theta \times X \times Y)$, consider the following procedure for computing sets of principal types. Initialize $\overline{\Theta}^{-1}(\chi, p) = \Theta$. For $k \in \mathbb{N}$, let

$$\begin{aligned} \widetilde{D}^k_\theta(\chi, p) &= \{\gamma \in \Gamma(\overline{\Theta}^{k-1}(\chi, p), \chi) : \mathbb{E}_\gamma[\alpha \mathbb{E}_\chi[\mathbb{E}_{\beta(x)}[U(\theta, x, y)]]] > U(\theta, p)\}, \\ \widetilde{D}^{0,k}_\theta(\chi, p) &= \{\gamma \in \Gamma(\overline{\Theta}^{k-1}(\chi, p), \chi) : \mathbb{E}_\gamma[\alpha \mathbb{E}_\chi[\mathbb{E}_{\beta(x)}[U(\theta, x, y)]]] = U(\theta, p)\}, \\ \Theta^{\dagger,k}(\chi, p) &= \{\theta \in \Theta : \widetilde{D}^k_\theta(\chi, p) \cup \widetilde{D}^{0,k}_\theta(\chi, p) \not\subseteq \cup_{\theta' \neq \theta} \widetilde{D}_{\theta'}(\chi, p)\}, \\ \overline{\Theta}^k(\chi, p) &= \begin{cases} \Theta^{\dagger,k}(\chi, p) & \text{if } \Theta^{\dagger,k}(\chi, p) \neq \emptyset \\ \overline{\Theta}^{k-1}(\chi, p) & \text{if } \Theta^{\dagger,k}(\chi, p) = \emptyset \end{cases}, \text{ and then set} \\ \overline{\Theta}^\infty(\chi, p) &= \cap_{k \in \mathbb{N}} \overline{\Theta}^k(\chi, p). \end{aligned}$$

$\widetilde{D}^k_\theta(\chi, p)$ gives the set of distributions over agent best responses to a belief supported in $\overline{\Theta}^{k-1}(\chi, p)$ that would make type θ strictly better off by proposing χ than sticking with the outcome p. $\widetilde{D}^{0,k}_\theta(\chi, p)$ gives the analogous set of distributions that make type θ indifferent between proposing χ and sticking with p. $\Theta^{\dagger,k}(\chi, p)$ is the set of principal types for which there is some mixture over agent best responses to the proposal of χ and beliefs supported on $\overline{\Theta}^{k-1}(\chi, p)$ that makes that type (weakly) prefer to propose such a mechanism than stick with p and makes every other type (weakly) prefer sticking with p. $\overline{\Theta}^k(\chi, p)$ equals $\Theta^{\dagger,k}(\chi, p)$ if $\Theta^{\dagger,k}(\chi, p)$ is non-empty and otherwise equals $\overline{\Theta}^{k-1}(\chi, p)$, and $\overline{\Theta}^\infty(\chi, p)$ is the limit of $\overline{\Theta}^k(\chi, p)$ as $k \to \infty$.

Let $\Theta^{SJ,\dagger}(\chi, p) = \{\theta \in \overline{\Theta}^\infty(\chi, p) : \exists (1, \beta) \in C(\overline{\Theta}^\infty(\chi, p), \chi) \text{ s.t. } \mathbb{E}_\chi[\mathbb{E}_{\beta(x)}[U(\theta, x, y)]] \geq U(\theta, p)\}$ be the set of principal types in $\overline{\Theta}^\infty(\chi, p)$ for which there is some agent best response to the proposal of χ and beliefs supported on $\overline{\Theta}^\infty(\chi, p)$ that accepts the proposal and makes that type (weakly) prefer to propose such a mechanism than stick with p. Then let

$$\Theta^{SJ}(\chi, p) = \begin{cases} \Theta^{SJ,\dagger}(\chi, p) & \text{if } \Theta^{SJ,\dagger}(\chi, p) \neq \emptyset \\ \overline{\Theta}^\infty(\chi, p) & \text{if } \Theta^{SJ,\dagger}(\chi, p) = \emptyset \end{cases}.$$

Definition OA 3. *The set of* **strongly justified types** *for χ given outcome p is $\Theta^{SJ}(\chi, p)$.*

Definition OA 4. *Outcome p is a* **strongly justified communication equilibrium (SJCE)** *if it is incentive compatible and, for every $\chi \in \Delta(X)$, there is some $\gamma \in \Gamma(\Theta^{SJ}(\chi, p), \chi)$ such that $\mathbb{E}_\gamma[\alpha \mathbb{E}_\chi[\mathbb{E}_{\beta(x)}[U(\theta, x, y)]]] \leq U(\theta, p)$ for all $\theta \in \Theta$.*

OA.13.2 Definitions for the Deterministic-Mechanism Game

Robust Neologism Proofness:

Definition OA 5. *Contracting equilibrium outcome p has a* **credible robust neologism** *if there exists some $x \in X$ and non-empty subset of principal types $\widetilde{\Theta} \subseteq \Theta$ such*

that

1. $\min_{\tilde{\lambda}\in\Delta(\widetilde{\Theta})} \max_{y\in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta,x,y)] > 0$,
2. $\min_{y\in BR(\widetilde{\Theta},x)} U(\theta,x,y)] > U(\theta,p)$ *for all* $\theta \in \widetilde{\Theta}$, *and*
3. $\max_{y\in BR(\widetilde{\Theta},x)} U(\theta',x,y)] < U(\theta',p)$ *for all* $\theta' \notin \widetilde{\Theta}$.

Definition OA 6. *A contracting equilibrium outcome is* **robust neologism proof (RNP)** *if it does not have a credible robust neologism.*

Strongly Justified Communication Equilibrium: For every $\widetilde{\Theta} \subseteq \Theta$ and $x \in X$, let $C(\widetilde{\Theta}, x)$ be the set of agent responses to a proposal of x given by

$$\begin{aligned} C(\widetilde{\Theta},x) = \{(\alpha,\beta) \in [0,1]\times\Delta(Y) : \exists \tilde{\lambda} \in \Delta(\widetilde{\Theta}) \text{ s.t.} \\ (1)\ \beta \in \Delta(\arg\max_{y\in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta,x,y)]), \\ (2)\ \alpha = 0 \text{ if } \max_{y\in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta,x,y)] < 0, \text{ and} \\ (3)\ \alpha = 1 \text{ if } \max_{y\in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta,x,y)] > 0\}. \end{aligned}$$

We let $\Gamma(\widetilde{\Theta},x) \equiv \Delta(C(\widetilde{\Theta},x))$ be the set of distributions over all such agent responses.

Fixing $x \in X$ and outcome $p \in \Delta(\Theta\times X\times Y)$, consider the following procedure for computing sets of principal types. Initialize $\overline{\Theta}^{-1}(x,p) = \Theta$. For $k \in \mathbb{N}$, let

$$\begin{aligned} \widetilde{D}^k_\theta(x,p) &= \{\gamma \in \Gamma(\overline{\Theta}^{k-1}(x,p),x) : \mathbb{E}_\gamma[\alpha\mathbb{E}_\beta[U(\theta,x,y)]] > U(\theta,p)\}, \\ \widetilde{D}^{0,k}_\theta(x,p) &= \{\gamma \in \Gamma(\overline{\Theta}^{k-1}(x,p),x) : \mathbb{E}_\gamma[\alpha\mathbb{E}_\beta[U(\theta,x,y)]] = U(\theta,p)\}, \\ \Theta^{\dagger,k}(x,p) &= \{\theta\in\Theta : \widetilde{D}^k_\theta(x,p) \cup \widetilde{D}^{0,k}_\theta(x,p) \not\subseteq \cup_{\theta'\neq\theta}\widetilde{D}_{\theta'}(x,p)\}, \\ \overline{\Theta}^k(x,p) &= \begin{cases} \Theta^{\dagger,k}(x,p) & \text{if } \Theta^{\dagger,k}(x,p)\neq\emptyset \\ \overline{\Theta}^{k-1}(x,p) & \text{if } \Theta^{\dagger,k}(x,p)=\emptyset \end{cases}, \text{ and then set} \\ \overline{\Theta}^\infty(x,p) &= \cap_{k\in\mathbb{N}}\overline{\Theta}^k(x,p). \end{aligned}$$

Let $\Theta^{SJ,\dagger}(x,p) = \{\theta \in \overline{\Theta}^\infty(x,p) : \exists(1,\beta) \in C(\overline{\Theta}^\infty(x,p),x) \text{ s.t. } \mathbb{E}_\beta[U(\theta,x,y)] \geq U(\theta,p)\}$ be the set of principal types in $\overline{\Theta}^\infty(x,p)$ for which there is some agent best

response to the proposal of x and beliefs supported on $\overline{\Theta}^{\infty}(x,p)$ that accepts the proposal and makes that type (weakly) prefer to propose such a mechanism than stick with p. Then let

$$\Theta^{SJ}(x,p) = \begin{cases} \Theta^{SJ,\dagger}(x,p) & \text{if } \Theta^{SJ,\dagger}(x,p) \neq \emptyset \\ \overline{\Theta}^{\infty}(x,p) & \text{if } \Theta^{SJ,\dagger}(x,p) = \emptyset \end{cases}.$$

Definition OA 7. *The set of* **strongly justified types** *for x given outcome p is $\Theta^{SJ}(x,p)$.*

Definition OA 8. *Outcome p is a* **strongly justified communication equilibrium (SJCE)** *if it is incentive compatible and, for every $x \in X$, there is some $\gamma \in \Gamma(\overline{\Theta}^{\infty}(x,p),x)$ such that $\mathbb{E}_{\gamma}[\alpha \mathbb{E}_{\beta}[U(\theta,x,y)]] \leq U(\theta,p)$ for all $\theta \in \Theta$.*

OA.14 Payoff-Plausibility Characterizes RNP and SJCE

Proposition OA 6. *Suppose the environment is MCS. In both the general-mechanism and deterministic-mechanism games, any RNP or SJCE outcome must be payoff-plausible, and every payoff-plausible outcome is both RNP and SJCE.*

Here we give the proof for the general-mechanism game. The proof for the deterministic-mechanism game is analogous.

Lemma OA 14. *Suppose the environment is MCS. In the general-mechanism game, any RNP or SJCE outcome must be payoff-plausible.*

Proof of Lemma OA 14 for RNP. Let p be an RNP outcome. We proceed by induction on the type index n beginning with the base case $n = N$. Take any $\chi \in \Delta(X \times T)$ that solves the type θ_N optimization problem in (1). For every $\varepsilon > 0$, let $\chi_{\varepsilon} \in \Delta(X \times T)$ be the distribution obtained from χ by shifting every t to $t + \varepsilon$. Then the constraints

in (1) are strictly satisfied by q_ε. Robust neologism proofness demands that the type θ_N principal obtain a payoff at least $\mathbb{E}_\chi[u(\theta_N, x, y^*(\theta_N, x)) - t] - \varepsilon$. Since this holds for all $\varepsilon > 0$, it follows that $U(\theta_N, p) \geq \mathbb{E}_\chi[u(\theta_N, x, y^*(\theta_N, x)) - t]$.

Now suppose that payoff-plausibility holds for all $n'' > n$ but not for n itself. Take any $\chi \in \Delta(X \times T)$ that solves the type θ_n optimization problem in (1), and let $q \in \Delta(X \times T \times Y)$ be the distribution obtained from χ by setting $y = y^*(\theta_n, x)$ and shifting every t to $t + \kappa$, where $\kappa > 0$ is chosen so that $\mathbb{E}_q[u(\theta_n, x, y^*(\theta_n, x)) - t] = U(\theta_n, p)$. Additionally, let $\chi' = \text{marg}_{X \times T} q$ and, for every $\varepsilon > 0$, let $\chi'_\varepsilon \in \Delta(X \times T)$ be the distribution obtained from χ' by shifting every t to $t - \varepsilon$. Every type below θ_n gets a strictly lower payoff from q than p. Moreover, since payoff-plausibility holds for all $n'' > n$, all types above θ_n must get a weakly lower payoff from q than p. If additionally every type above θ_n were to get a strictly lower payoff from q than p, then there would be a credible robust neologism corresponding to χ'_ε and θ_n for some sufficiently small $\varepsilon > 0$, a contradiction. Suppose instead that there are types above θ_n that would be indifferent between q and p, and let $\overline{\Theta}_n$ be the set of such types with $\theta_{n''}$ being the maximum of $\overline{\Theta}_n$. Then either (1) there is a credible robust neologism corresponding to χ'_ε and $\{\theta_n\} \cup \overline{\Theta}_n$ for some sufficiently small $\varepsilon > 0$, or (2) there is a type outside of $\{\theta_n\} \cup \overline{\Theta}_n$, say $\tilde{\theta}_n$ that would weakly prefer playing χ' when the agent responds under the belief that $\theta = \theta_{n''}$ over their outcome in p. Case (1) contradicts p being RNP. In Case (2), it must be that $\tilde{\theta}_n$ obtains a strictly higher payoff from playing χ' when the agent responds under the belief that $\theta = \theta_{n''}$ than when the agent responds under the belief that $\theta = \theta_n$. This implies that $\mathbb{E}_{\chi'}[u(\theta_{n''}, x, y^*(\theta_{n''}, x)) - u(\theta_{n''}, x, y^*(\theta_n, x))] > \mathbb{E}_{\chi'}[u(\theta_{n'}, x, y^*(\theta_{n''}, x)) - u(\theta_{n'}, x, y^*(\theta_n, x))]$ for all $n' < n''$. Consider the $\chi'' \in \Delta(X \times T)$ obtained from χ' by shifting every t up to $t + \mathbb{E}_{\chi'}[u(\theta_{n''}, x, y^*(\theta_{n''}, x)) - u(\theta_{n''}, x, y^*(\theta_n, x))]$. This χ'' strictly satisfies the constraints in the plausibility threshold problem of type $\theta_{n''}$ given in (1) and gives $\theta_{n''}$ the same payoff as p. This means that payoff-plausibility does not hold for $\theta_{n''}$, contradicting our inductive assumption. ■

Proof of Lemma OA 14 for SJCE. Let p be an SJCE outcome. We again proceed

by induction, beginning with the base case $n = N$. Take any $\chi \in \Delta(X \times T)$ that solves the type θ_N optimization problem in (1), and let $\chi_\varepsilon \in \Delta(X \times T)$ be the distribution obtained from taking χ and shifting every t to $t + \varepsilon$. Suppose that $U(\theta_N, p) < \mathbb{E}_\chi[u(\theta_N, x, y^*(\theta_N, x)) - t]$. Then θ_N is the unique strongly justified type for χ_ε for all sufficiently small $\varepsilon > 0$. Consequently, SJCE demands that the type θ_N principal obtain a payoff of at least $\mathbb{E}_\chi[u(\theta_N, x, y^*(\theta_N, x)) - t] - \varepsilon$, and since this holds for all $\varepsilon > 0$, a payoff of at least $\mathbb{E}_\chi[u(\theta_N, x, y^*(\theta_N, x)) - t]$.

Now suppose that payoff-plausibility holds for all $n'' > n$ but not for n itself. Take any $\chi \in \Delta(X \times T)$ that solves the type θ_n optimization problem in (1), and let $\chi_\varepsilon \in \Delta(X \times T)$ be the distribution obtained from taking χ and shifting every t to $t + \varepsilon$. Let $\varepsilon > 0$ be sufficiently small so that $U(\theta_n, p) < \mathbb{E}_\chi[u(\theta_N, x, y^*(\theta_N, x)) - t] - \varepsilon$. If all strongly justified types for χ_ε are above θ_n, a similar argument to the case for θ_N above then implies that $U(\theta_n, p) \geq \mathbb{E}_\chi[u(\theta_n, x, y^*(\theta_n, x)) - t] - \varepsilon$, a contradiction. Suppose instead that there is a strongly justified type for χ_ε below θ_n. There must be some $n'' > n$ such that type $\theta_{n''}$ is also strongly justified. Without loss of generality, assume that n'' is the highest such value. Consequently, there must be some $q' \in \Delta(X \times T \times Y)$ and $\alpha \in (0, 1]$ such that $\text{marg}_{X \times T} q' = \chi_\varepsilon$, $\mathbb{P}_{q'}[y \leq y^*(\theta_{n''}, x)] = 1$, either $\alpha < 1$ or $\mathbb{P}_{q'}[u(\theta_{n''}, x, y) < u(\theta_{n''}, x, y^*(\theta_{n''}, x))] > 0$, and $(1 - \alpha)\mathbb{E}_{q'}[u(\theta_{n''}, x, y) - t] = U(\theta_{n''}, p)$ and $(1 - \alpha)\mathbb{E}_{q'}[u(\theta_{n'}, x, y) - t] \leq U(\theta_{n'}, p)$ for all $n' < n''$. Consider now the allocation $q'' \in \Delta(X \times T \times Y)$ obtained from q' by shifting every y to $y^*(\theta_{n''}, x)$ and shifting every t up to $t + \mathbb{E}_{q'}[u(\theta_{n''}, x, y^*(\theta_{n''}, x))] - \alpha\mathbb{E}_{q'}[u(\theta_{n''}, x, y)]$. The allocation given by q'' strictly satisfies the constraints in the type $\theta_{n''}$ optimization problem given in (1) and gives $\theta_{n''}$ a payoff of $U(\theta_{n''}, p)$. This means that the payoff-plausibility threshold for $\theta_{n''}$ is strictly higher than $U(\theta_{n''}, p)$, contradicting our inductive assumption. ■

Lemma OA 15. *Suppose the environment is MCS. In the general-mechanism game, any payoff-plausible outcome is both RNP and SJCE.*

Proof of Lemma OA 15 for RNP. Suppose towards a contradiction that there is a credible robust neologism corresponding to χ and some non-empty $\widetilde{\Theta}$. Let $\underline{\theta} = \min(\widetilde{\Theta})$, and

consider the conditional distribution $q^*(\underline{\theta}) \in \Delta(X \times T \times Y)$ where $\text{marg}_{X\times T} q^*(\underline{\theta}) = \chi$ and $y = y^*(\underline{\theta}, x)$ for all $x \neq x_o$. By the definition of a credible robust neologism, $\mathbb{E}_{q^*(\underline{\theta})}[v(\underline{\theta}, x, y) + g(t)] > 0$ and $\mathbb{E}_{q^*(\underline{\theta})}[u(\theta, x, y) - t] < U^*(\theta)$ for all $\theta \notin \widetilde{\Theta}$, so $q^*(\underline{\theta})$ satisfies the constraints in (1) for type $\underline{\theta}$. Consequently, $\underline{\theta}$'s payoff must be weakly greater that from $q^*(\underline{\theta})$, but this contradicts there being a credible robust neologism corresponding to χ and $\widetilde{\Theta}$. ■

Proof of Lemma OA 15 for SJCE. Fix some $\chi \in \Delta(X \times T)$ and let p^* denote the outcome of the contracting equilibrium. We will show by induction that, for all $k \in \mathbb{N}$, there is a best response $\gamma \in \Gamma(\overline{\Theta}^k(\chi, p^*), \chi)$ that deters all principal types from proposing χ.

We begin with the base case $k = 0$. If every agent best response to χ makes every principal type no better off than in p^*, then we are done. Suppose instead that there is some agent best response to χ that makes some principal type strictly better off than in p^*. To obtain some $\gamma \in \Gamma(\overline{\Theta}^0(\chi, p^*), \chi)$ that deters the principal types, it will be sufficient to consider the family of agent posterior beliefs $\mathbf{\Lambda}_0 = \{\tilde{\lambda} \in \Delta(\Theta) : \exists n \in \{1, ..., N\} \text{ s.t. } \tilde{\lambda}(\theta_m) = 0 \text{ if } m < n \text{ or } m > n + 1\}$ that put positive probability on at most two principal types, which must be adjacent. Note that FOSD gives a complete ordering over $\mathbf{\Lambda}_0$. Since the mapping from agent beliefs to agent best responses is upper hemicontinuous, there is some smallest (according to FOSD) $\underline{\lambda} \in \mathbf{\Lambda}_0$ for which there is an agent best response that makes some principal type in $\overline{\Theta}^0(\chi, p^*)$ weakly better off than in p^*. If the agent strictly prefers to either accept or reject χ under belief $\underline{\lambda}$, the associated best response is pinned down. If instead the agent is precisely indifferent between accepting or rejecting χ, fix the agent best response to $\underline{\lambda}$ that accepts the proposal with the smallest probability among the best responses for which some principal type in $\overline{\Theta}^0(\chi, p^*)$ weakly prefers χ to p^*. Let $q \in \Delta(X \times T \times Y)$ be the distribution obtained from χ under this agent best response, and let $\theta_{\underline{n}}$ be the smallest type in $\overline{\Theta}^0(\chi, p^*)$ which weakly prefers q to p^*. We handle three cases: (1) $\underline{\lambda}(\theta_{\underline{n}}) = 1$, (2) $\underline{\lambda}(\theta) > 0$ for some $\theta > \theta_{\underline{n}}$, and (3) $\underline{\lambda}(\theta) > 0$ for some $\theta < \theta_{\underline{n}}$. In Cases (1) and (2),

there is an agent best response to a belief fully supported on $\theta_{\underline{n}} \in \overline{\Theta}^0(\chi, p^*)$ that deters all principal types. We now establish that this is also true for Case (3). If there is an agent best response to χ and a belief fully supported on $\theta_{\underline{n}}$ that rejects χ, then we are done. Otherwise, let $q' \in \Delta(X \times T \times Y)$ be the distribution obtained from χ and the agent best response to a belief fully supported on $\theta_{\underline{n}}$. If $\theta_{\underline{n}}$ were to get a strictly higher payoff from q' than p^*, then, for sufficiently small $\varepsilon > 0$, the $\chi'_\varepsilon \in \Delta(X \times T)$ that results from taking χ and shifting every t to $t + \mathbb{E}_{q'}[u(\theta_{\underline{n}}, x, y)] - \mathbb{E}_q[u(\theta_{\underline{n}}, x, y)] - \varepsilon$, satisfies the constraints in (1) and gives type $\theta_{\underline{n}}$ a strictly higher payoff than p^*, which violates payoff-plausibility. Since $\theta_{\underline{n}}$ gets a weakly lower payoff from q' than p^*, this must hold for all lower types as well. Suppose that some higher type θ'' would get a strictly higher payoff from q' than p^*, and suppose without loss of generality that θ' is the lowest such type. Then the χ'' which results from taking χ and shifting every t to $t + \mathbb{E}_\chi[u(\theta'', x, y^*(\theta'', x)) - u(\theta'', x, y^*(\theta_{\underline{n}}, x))]$ would satisfy the constraints in (1) and give type θ'' a strictly higher payoff than p^*, violating payoff-plausibility.

We now establish the claim for arbitrary $K \in \mathbb{N}$ assuming that it is true for all $k < K$. Since $\overline{\Theta}^K(\chi, p^*) \subseteq \overline{\Theta}^{K-1}(\chi, p^*)$, if every $\gamma \in \Gamma(\overline{\Theta}^{K-1}(\chi, p^*), \chi)$ makes every principal type no better than in p^*, then we are done. Suppose instead that there is some $\gamma \in \Gamma(\overline{\Theta}^{K-1}(\chi, p^*), \chi)$ that makes some principal type strictly better off than in p^*. Consider the family of agent posterior beliefs $\mathbf{\Lambda}_K$ that are supported on $\overline{\Theta}^{K-1}(\chi, p^*)$ and put positive probability on at most two principal types, which must be adjacent. A similar argument to the $K = 0$ case shows that there is some smallest (according to FOSD) $\underline{\lambda} \in \mathbf{\Lambda}_K$ for which there is an agent best response that makes some principal type in $\overline{\Theta}^K(\chi, p^*)$ weakly better off than in p^*. As before, if the agent is precisely indifferent between accepting or rejecting χ under belief $\underline{\lambda}$, fix the agent best response to $\underline{\lambda}$ that accepts the proposal with the smallest probability among the best responses for which some principal type in $\overline{\Theta}^K(\chi, p^*)$ weakly prefers χ to p^*. Let $q \in \Delta(X \times T \times Y)$ be the distribution obtained from χ under this agent best response, and let $\theta_{\underline{n}}$ be the smallest type in $\overline{\Theta}^K(\chi, p^*)$ that weakly prefers q to p^*. A similar argument to the base

case above then shows that there must be some agent best response to a belief fully supported on $\theta_{\underline{n}}$ which deters all principal types from proposing χ.

Since there is some $K \in \mathbb{N}$ such that $\overline{\Theta}^k(\chi, p^*) = \overline{\Theta}^\infty(\chi, p^*)$ for all $k > K$, it follows that there is a best response $\gamma \in \Gamma(\overline{\Theta}^\infty(\chi, p^*), \chi)$ that deters all principal types from proposing χ. Using this fact, a similar argument to those above then shows that there is a best response $\gamma \in \Gamma(\Theta^{SJ}(\chi, p^*), \chi)$ that deters all principal types from proposing χ, which means that p^* is an SJCE outcome. ■