Online Appendix for “The Informed Principal with Agent Moral Hazard”

Daniel Clark
November 14, 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

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

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 $\mathbb{P}_p[(s, e) = (s^*, 4s^*)|\theta = 4] = 1$.

**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 $\theta$ from the conditional distribution of $p$ given 4 is $\mathbb{E}_p[U(\theta, s, t, e)|\theta = 4] = \mathbb{E}_p[s(1-s)|\theta = 4] - \mathbb{E}_p[t|\theta = 4]$, 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

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

\[
t = -4s^2 + 4s(1 - s) - 1, \\
U(2) = 4s^2 + 8s(1 - s) + 1.
\]

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

\[
U(2) = -4s^2 + 8s + 1, \\
U(4) = -8s^2 + 12s + 1.
\]

Solving for \( U(4) \) in terms of \( U(2) \) then gives \( U(4) = -U(2)^2/2 + 4U(2) - 5/2 \).
Proposition OA 2. The upper envelope of the firm payoffs sustainable in contracting equilibrium with explicit contracts is
\[
\{ (U(2), U(4)) \in \mathbb{R}^2 : (U(2), U(4)) = (2, 4), \ 2 < U(2) \leq \frac{21}{8} \text{ and } U(4) = U(2) + \frac{3}{2}, \text{ or } \]
\[
\frac{21}{8} \leq U(2) \leq \frac{9}{2} \text{ and } U(4) = 5U(2) - 6\sqrt{36 - 6U(2) - 36} \}.
\]

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

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 \(\alpha\), 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 \(t + (2 - \tilde{\lambda}(2))^2s^2/2 \geq 0\). Thus, the payoff of the type 4 firm is bounded from above.
Observe that, for any \( \bar{\lambda} > 1/2 \), decreasing \( \bar{\lambda} \) to 1/2 and increasing \( t \) by \( 4(\bar{\lambda}(2) - 1/2)s(1-s) \) 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 \( \bar{\lambda}(2) = 1/2 \) 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 \((s,t,\bar{\lambda}(2),\alpha)\in[0,1] \times [\frac{1}{2},1] \times (0,1]\) such that \( U(2) = \alpha(4(2 - \bar{\lambda}(2))s(1-s) - t) \) and \( t + (2 - \bar{\lambda}(2))^2s^2/2 \geq 0 \). Thus, we have \( U(2) \leq \alpha(4(2 - \bar{\lambda}(2))s(1-s) + (2 - \bar{\lambda}(2))^2s^2/2) \). Standard computations show that \( \max_{(s,t,\bar{\lambda}(2),\alpha)} \alpha(4(2 - \bar{\lambda}(2))s(1-s) + (2 - \bar{\lambda}(2))^2s^2/2) = 9/2 \). ■

**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

\[
t = -\frac{9}{2} s^2, \\
U(2) = \frac{9}{2} s^2 + 6s(1-s).
\]

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

\[
U(2) = -\frac{3}{2} s^2 + 6s, \quad U(4) = -\frac{15}{2} s^2 + 12s.
\]

Solving for \( U(4) \) in terms of \( U(2) \) then gives \( U(4) = 5U(2) - 6\sqrt{36 - 6U(2)} - 36 \). ■
OAx.3  Plausible Payoffs with Flexible Contracts

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

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

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

\[
\max_{(s,t) \in [0,1] \times \mathbb{R}} 16s(1 - s) - t
\]

s.t. AIR: \( 8s^2 + t \geq 0 \),

\[ \text{PIC: } 8s(1 - s) - t \leq U(2) . \]

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.

\[ \blacksquare \]

**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 \( \theta \). We first note that \( \mathcal{M}_{safe} \) is non-empty since every direct mechanism in which each principal type commits to \( x_0 \), i.e. \( q(\theta)[x_0] = 1 \) 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 \(\{q_j(\theta)\}_{\theta \in \Theta} \) 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 \(\theta \). Then there is some closed set \(\tilde{X} \subseteq X \) and agent action \(\tilde{y} \in Y \) such that \(\mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\tilde{X}}(x)V(\theta, x, \tilde{y})] > \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\tilde{X}}(x)V(\theta, x, y)] \). For every \(\varepsilon > 0 \), let \(\tilde{X}_{\leq \varepsilon} = \{x \in X : d(x, \tilde{X}) < \varepsilon\} \), \(\tilde{X}_{= \varepsilon} = \{x \in X : d(x, \tilde{X}) = \varepsilon\} \), and \(\tilde{X}_{> \varepsilon} = \{x \in X : d(x, \tilde{X}) > \varepsilon\} \). Additionally, let \(V(\theta) = \min_{(x,y) \in X \times Y} V(\theta, x, y) \). By continuity, there exists some \(\tilde{\varepsilon} > 0 \) such that

\[
\mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\tilde{X}_{\leq \varepsilon}}(x)(V(\theta, x, \tilde{y}) - V(\theta))] + \mathbb{E}_{q^*(\theta)}[\mathbb{1}_{\tilde{X}_{= \varepsilon}}(x)(V(\theta, x, y) - V(\theta))] \\
> \mathbb{E}_{q^*(\theta)}[V(\theta, x, y)] - V(\theta).
\]

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

\[
\mathbb{E}_{q_j(\theta)}[\mathbb{1}_{\tilde{X}_{\leq \varepsilon}}(x)(V(\theta, x, \tilde{y}) - V(\theta))] + \mathbb{E}_{q_j(\theta)}[\mathbb{1}_{\tilde{X}_{> \varepsilon}}(x)(V(\theta, x, y) - V(\theta))] \\
> \mathbb{E}_{q_j(\theta)}[V(\theta, x, y)] - V(\theta).
\]

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

Let \(\overline{U}_{safe}(\theta) \equiv \sup_{(q(\theta))' \in \Theta_{safe}} U(\theta, q(\theta)) \) denote the supremum of the type \(\theta\) principal’s payoff over all safe mechanisms. Let \(\{q_j(\theta(\theta))\}_{\theta \in \Theta} \) 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 \( \theta \) 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.

\[ \square \]

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

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

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

\[ q^*(\theta_n) \in \arg \max_{q \in \Delta(X \times T \times Y)} \mathbb{E}_q[u(\theta_n, x, y) - t] \]

s.t.  \( AIC: \mathbb{P}_q[y = y^*(\theta_n, x)|x \neq x_o] = 1, \)

\( AIR: \mathbb{E}_q[v(\theta_n, x, y) + g(t)] \geq 0, \)

\( PIC: \mathbb{E}_q[u(\theta_{n'}, x, y) - t] \leq \mathbb{E}_{q^*(\theta_{n'})}[u(\theta_{n'}, x, y) - t] \forall n' < n, \)

for all \( n \in \{1, \ldots, N\}. \)

**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 \( \theta_n \). 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. \( \blacksquare \)
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

\[
q^*(\theta_n) = \arg \max_{q \in \Delta(X \times T \times Y)} \mathbb{E}_q[u(\theta_n, x, y^*(\theta_n, x)) - t]
\]

s.t. \( AIC: \mathbb{P}_q[y = y^*(\theta_n, x)|x \neq x_o] = 1 \)

\( AIR: \mathbb{E}_q[v(\theta_n, x, y) + g(t)] \geq 0 \)

\( PIC: \mathbb{P}_q[u(\theta_{n'}, x, y) - t \leq U'(\theta_n, q^*(\theta_{n'}))] = 1 \) \( \forall n' < n \),

for all \( n \in \{1, \ldots, N\} \).

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.

\[\square\]

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 \( \theta \). 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 \( x_{\theta}^{CI} \) such that

\[
u(\theta, x_i, y^*(\theta, x_i)) - u(\theta, x_{\theta}^{CI}, y^*(\theta, x_{\theta}^{CI})) > \nu(\theta', x_i, y^*(\theta, x_i)) - u(\theta', x_{\theta}^{CI}, y^*(\theta, x_{\theta}^{CI})) \]

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 \( \theta \) 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 $\theta$, that gives the agent a strictly positive expected utility. Let $t_{\theta}$ be the expected transfer played by $\theta$ in the ex-ante mechanism design benchmark. For each $i \in \mathbb{N}$, consider the transfer given by $t_i = t_{\theta} + u(\theta, x_i, y^*(\theta, x_i)) - u(\bar{\theta}, x_{CI}, y^*(\bar{\theta}, x_{CI})).$

By construction, the type $\theta$ would obtain the same payoff from $(x_i, t_i)$ and the agent responding with $y^*(\theta, x_i)$ 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 $\varepsilon$ such that $(x_i, t_i - \varepsilon)$ and the agent responding with $y^*(\theta, x_i)$ gives the type $\theta$ a strictly higher payoff than the ex-ante mechanism design benchmark, all types lower than $\theta$ a strictly lower payoff than the ex-ante mechanism design benchmark, and the agent a strictly positive expected utility when the type is $\theta$, which means that the payoff of the type $\theta$ 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 $y^*(\lambda, x)$, quasi-strictness holds at each $x_i$, and either one of the following conditions hold:

1. \begin{enumerate}
   \item $(a) \ u(\theta, x, y^*(\tilde{\lambda}, x))$ is constant in $\theta$.
   \item $(b) \ u(\theta, x_i, y^*(\tilde{\lambda}, x_i)) > u(\theta, x, y^*(\tilde{\lambda}, x))$ for all $i$.
   \item $(c) \ v(\theta, x_i, y^*(\tilde{\lambda}, x_i)) > v(\theta, x, y^*(\tilde{\lambda}, x))$ for all $i$.
\end{enumerate}

2. \begin{enumerate}
   \item $(a) \ u(\theta, x, y^*(\tilde{\lambda}, x))$ is constant in $\theta$.
   \item $(b) \ v(\theta, x, y^*(\tilde{\lambda}, x))$ is strictly increasing in $\theta$.
\end{enumerate}

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 \( \bar{\theta} \in \Theta, x \in X, t \in T, \bar{\lambda} \in \Delta(\Theta) \), and \( \alpha \in (0, 1) \) such that (1) \( \alpha(u(\bar{\theta}, x, y^*(\bar{\lambda}, x)) - t) = U(\bar{\theta}, p) \), (2) \( \alpha(u(\theta, x, y) - t) \leq U(\theta, p) \) for all \( \theta \neq \bar{\theta} \), (3) \( \bar{\lambda} \) is weakly below \( \delta_{\bar{\theta}} \) under FOSD, and (4) \( v(\bar{\theta}, x, y^*(\bar{\theta}, x')) + g(t) \geq 0 \). Consider \( (x', t') \) such that \( t' = \alpha t + u(\bar{\theta}, x', y^*(\bar{\theta}, x')) - \alpha u(\bar{\theta}, x, y^*(\bar{\lambda}, x)) \). By construction, this \( (x', t') \) is such that, when the agent responds with \( y^*(\bar{\theta}, x') \), the type \( \bar{\theta} \) 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 \( (x', t', y^*(\bar{\theta}, x')) \) than \( p \) and the agent gets a strictly higher utility from \( \bar{\theta} \) playing \( (x', t') \) than their outside option. Thus, for sufficiently small \( \varepsilon > 0 \), \( (x', t' - \varepsilon) \) would satisfy the constraints of the type \( \bar{\theta} \) optimization problem in (2) and give \( \bar{\theta} \) 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 \( \theta \) such that \( \mathbb{P}[v(\theta, x, y) + g(t) > 0|\theta] > 0 \), and suppose that \( \bar{\theta} \) is the highest type for which this is true. Then there is some \( x \in X, t \in T, \) and \( \bar{\lambda} \in \Delta(\Theta) \) such that (1) \( u(\bar{\theta}, x, y^*(\bar{\lambda}, x)) - t = U(\bar{\theta}, p) \), (2) \( u(\theta, x, y) - t \leq U(\theta, p) \) for all \( \theta \neq \bar{\theta} \), (3) \( \bar{\lambda} \) is weakly below \( \delta_{\bar{\theta}} \) under FOSD, and (4) \( v(\bar{\theta}, x, y^*(\bar{\theta}, x)) + g(t) > 0 \). Consider \( (x', t') \) such that \( t' = t + u(\bar{\theta}, x', y^*(\bar{\theta}, x')) - u(\bar{\theta}, x, y^*(\bar{\lambda}, x)) \). By construction, this \( (x', t') \) is such that, when the agent responds with \( y^*(\bar{\theta}, x') \), the type \( \bar{\theta} \) 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 \( (x', t', y^*(\bar{\theta}, x')) \) than \( p \) and the agent gets a strictly higher utility from \( \bar{\theta} \) playing \( (x', t') \) than their outside option. Thus, for sufficiently small \( \varepsilon > 0 \), \( (x', t' - \varepsilon) \) would satisfy the constraints of the type \( \bar{\theta} \) optimization problem in (2) and give \( \bar{\theta} \) 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 $(\mu_j, M_P)$ indexed by $j \in \mathbb{N}$, where $M_P = \{m_{P,1}, m_{P,2}, m_{P,3}, m_{P,4}\}$ and

$$
\mu_j(m_P) = \begin{cases}
\delta((1/2, 0), 0) & \text{if } m_P = m_{P,1} \\
\delta((-1/2, 0), 0) & \text{if } m_P = m_{P,2} \\
\delta((1, 1), 1) & \text{if } m_P = m_{P,3} \\
\delta((-1, 1), -1) & \text{if } m_P = m_{P,4}
\end{cases}
$$

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

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

The unique sequential continuation equilibrium after mechanism $(\mu_j, M_P)$ 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$. Consequently, the agent’s expected utility from accepting this mechanism is strictly negative, so mechanism ($\mu_j, M_P$) must be rejected in any contracting equilibrium.

In every sequential continuation equilibrium where mechanism ($\mu, M_P$) 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 $\{P_j\}_{j \in \mathbb{N}}$ that converges to the original primitives $P$. 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 ($\mu, M_P$) 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 $f_{m_P} : \cup_{m'_P, \text{supp}(\mu(m'_P))} \rightarrow [0, |M_P|]$ be the Radon-Nikodym derivative of the $\mu(m_P)$ distribution with respect to $\nu$. 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 $M(x, r) = \{(m_P) \in M_P : f_{(m_P)}(x, r) > 0\}$ 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 ($\mu, M_P$), and the
\( \tilde{\mu} \) is induced from \( \mu \) by replacing each principal-action recommendation pair \((x, r) ∈ \cup m'_p, \text{supp}(\mu(m'_p)) \) with \((x, (r, M(x, r))) \). Let \( \tilde{\nu} = \sum_{m_p} \tilde{\mu}(m_p) / |M_p| \) and \( \tilde{f}_{(m_p)} : \cup m'_p, \text{supp}(\mu(m'_p)) → [0, |M_p|] \) be the Radon-Nikodym derivative of the \( \tilde{\mu}(m_p) \) distribution with respect to \( \tilde{\nu} \). Then, by construction, \( \tilde{f}_{(m_p)}(x, (r, M(x, r))) = f_{(m_p)}(x, r) \) for all \( m_p ∈ M_p \) and \((x, r) ∈ X × R \), while \( \tilde{f}_{m_p}(x, (r, M)) = 0 \) for all \( m_p ∈ M_p \), \((x, r) ∈ X × R \), and \( M ≠ M(x, r) \). Moreover, both \((\mu, M_p)\) and \((\tilde{\mu}, M_p)\) have the same sequential continuation equilibrium outcomes. \( \blacksquare \)

Construction of Sequences of Mechanisms. Suppose, by restricting attention to a subsequence if necessary, that there is a finite \( \tilde{X}_j ⊆ X_j \) such that, for all \( j ∈ N \) and \( x ∈ X \), there is an \( x' ∈ \tilde{X}_j \) satisfying \(|x - x'| ≤ 1/j\). Additionally, suppose that there is some finite \( \tilde{R}_j ⊂ R \) such that \(|\tilde{R}_j| ≤ |R_j|/2^{M_p}|\) and, for all \( r ∈ R \), there is an \( r' ∈ \tilde{R}_j \) satisfying \(|r - r'| ≤ 1/j\). The requirement on the relative sizes of \( \tilde{R}_j \) and \( R_j \) means that, for each \( r' ∈ \tilde{R}_j \) and \( M' ∈ P_+(M_p) \), 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 \( x ∈ \tilde{X}_j, r ∈ \tilde{R}_j, \) and \( M ∈ P_+(M_p) \), let

$$
\tilde{\mu}_j(m_p)[x', (r', M')] = \mathbb{E}_{(x, r) ∼ \mu(m_p)} \left[ \mathbb{I}(M(x, r) = M') \mathbb{I} \left( |x - x'| ≤ \frac{1}{j} \right) \mathbb{I} \left( |r - r'| ≤ \frac{1}{j} \right) \right]
$$

and \( \tilde{\mu}_j(m_p)[x', (r', M'')] = 0 \) for all \( M'' ≠ M' \). By construction, \( \tilde{\mu}_j(m_p)[x', (r', M')] ≥ 0 \) for all \( x ∈ \tilde{X}_j, r ∈ \tilde{R}_j, \) and \( M ∈ P_+(M_p) \), and \( \sum_{x', r', M'} \tilde{\mu}_j(m_p)[x', (r', M')] = 1 \). Therefore, \( \tilde{\mu}_j(m_p) ∈ Δ(\tilde{X}_j × \tilde{R}_j × P_+(M_p)) \). Moreover, \( \tilde{\mu}_j(m_p)[M'] = \mathbb{E}_{(x, r) ∼ \mu(m_p)}[\mathbb{I}(M(x, r) = M')] \) is the probability of realizing some \((x, r)\) for which \( M(x, r) = M' \) under \( \mu(m_p) \). \( \blacksquare \)

**Lemma OA 7.** For all \( m_p ∈ M_p \), \( \lim_{j → ∞} \tilde{\mu}_j(m_p) = \tilde{\mu}(m_p) \).

**Proof.** Let \( O \) be an arbitrary open subset of \( X × R \) and \( M \) be an arbitrary element of \( P_+(M_p) \). We need to show that \( \liminf_{j → ∞} \tilde{\mu}_j(m_p)[O × \{M\}] ≥ \tilde{\mu}(m_p)[O × \{M\}] \). For any \( ε > 0 \), let \( ((X × \tilde{R}) \setminus O)_{≥ ε} = \{(x, r) ∈ O : \forall (x', r') \notin O, |x - x'| ≥ ε \text{ or } |r - r'| ≥ ε\} \) be the subset of points in \( O \) that are of distance at least \( ε \) from \((X × R) \setminus O\). By
construction, for every $j > 1/\varepsilon$, we have $\mu_j(m_P)[\mathcal{O} \times \{M\}] \geq \mu(m_P)[((X \times \tilde{R}) \setminus \mathcal{O})_{\geq \varepsilon} \times \{M\}]$. Since $\lim_{\varepsilon \to 0} \mu(m_P)[((X \times \tilde{R}) \setminus \mathcal{O})_{\geq \varepsilon} \times \{M\}] = \mu(m_P)[\mathcal{O} \times \{M\}]$, the claim follows.

**Lemma OA 8.** Fix an $M \in P_+(M_P)$ and $m_P, m'_P \in M_P$ such that $m_P, m'_P \in M$. For any $j \in \mathbb{N}$, let $q_j \in \Delta(\tilde{X}_j \times \tilde{R}_j \times Y_j)$ and $q'_j \in \Delta(\tilde{X}_j \times \tilde{R}_j \times Y_j)$ be the distributions induced by the conditional distributions given $M$ of $\mu_j(m_P)$ and $\mu_j(m'_P)$, 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 $C \subseteq X \times \tilde{R}$, closed $\tilde{Y} \subseteq Y$, and $\kappa > 0$ such that either 

(1) $\mathbb{P}_q[C] > 0$ and $\mathbb{P}_q[\tilde{Y}|x, r] > (1 + \kappa)\mathbb{P}_q[\tilde{Y}|x, r]$ for all $(x, r) \in C$, or

(2) $\mathbb{P}_q[C] > 0$ and $\mathbb{P}_{q'}[\tilde{Y}|x, r] > (1 + \kappa)\mathbb{P}_q[\tilde{Y}|x, r]$ for all $(x, r) \in C$. 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 $\tilde{C} \subseteq C$ satisfying $\mathbb{P}_q[\tilde{C}] > 0$ and on which $h$ is continuous.

Fix $\eta > 0$. Since $h$ is continuous and strictly positive on $\tilde{C}$, for any $(x, r) \in \tilde{C}$, 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 $\tilde{C}$ given by \(\{B_{\delta(x, r)}(x, r)\}_{(x, r) \in \tilde{C}}\), where, for any $(x, r) \in \tilde{C}$ and $\delta > 0$, $B_{\delta}(x, r) = \{(x', r') \in \tilde{C} : |x' - x|, |r' - r| < \delta\}$ is the set of points in $\tilde{C}$ of distance less than $\delta$ to $(x, r)$. As $\tilde{C}$ 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 $\tilde{C} = B_{\delta(x_k, r_k)}(x_k, r_k)$. Note that, by construction, $(1 - \eta)\rho < \tilde{f}_{m_P}(x, r)/\tilde{f}_{m'_P}(x, r) < (1 + \eta)\rho$
for all \((x, r) \in \tilde{\mathcal{C}}\) where \(\rho = \tilde{f}_{m, p}(x_k, r_k)/\tilde{f}_{m, p'}(x_k, r_k) > 0\). From this, it follows that

\[
\mathbb{P}_q[\tilde{\mathcal{C}} \times \tilde{Y}] = \mathbb{E}_q[\mathbf{1}_{\tilde{\mathcal{C}}}(x, r) \mathbb{P}_q[\tilde{Y} | x, r]]
\]

\[
= \mathbb{E}_q \left[ \frac{\tilde{f}_{m, p}(x, r)}{\tilde{f}_{m, p'}(x, r)} \mathbf{1}_{\tilde{\mathcal{C}}}(x, r) \mathbb{P}_q[\tilde{Y} | x, r] \right]
\]

\[
\leq \left( \frac{1 + \eta}{1 + \kappa} \right) \rho \mathbb{E}_q[\mathbf{1}_{\tilde{\mathcal{C}}}(x, r) \mathbb{P}_q[\tilde{Y} | x, r]]
\]

\[
= \left( \frac{1 + \eta}{1 + \kappa} \right) \rho \mathbb{P}_q[\tilde{\mathcal{C}} \times \tilde{Y}],
\]

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

\[
\mathbb{P}_q[\tilde{\mathcal{C}}] = \mathbb{E}_q[\mathbf{1}_{\tilde{\mathcal{C}}}(x, r)]
\]

\[
= \mathbb{E}_q \left[ \frac{\tilde{f}_{m, p}(x, r)}{\tilde{f}_{m, p'}(x, r)} \mathbf{1}_{\tilde{\mathcal{C}}}(x, r) \right]
\]

\[
\geq (1 - \eta) \rho \mathbb{E}_q[\mathbf{1}_{\tilde{\mathcal{C}}}(x, r)]
\]

\[
= (1 - \eta) \rho \mathbb{P}_q[\tilde{\mathcal{C}}].
\]

Combining these two inequalities gives

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

For any \(\varepsilon > 0\), let \(\tilde{\mathcal{C}}_{\leq \varepsilon} = \{(x, r) \in X \times \tilde{R} : \exists (x', r') \in \tilde{\mathcal{C}} \text{ s.t. } |x - x'|, |r - r'| \leq \varepsilon\}\) be the set of points in \(X \times \tilde{R}\) that are of distance no more than \(\varepsilon\) from \(\tilde{\mathcal{C}}\). Likewise, let \(\tilde{Y}_{\leq \varepsilon} = \{y \in Y : \exists \hat{y} \in \tilde{Y} \text{ s.t. } |y - \hat{y}| \leq \varepsilon\}\) be the set of points in \(Y\) that are of distance no more than \(\varepsilon\) from \(\tilde{Y}\). Note that \(\mathbb{P}_q[\tilde{\mathcal{C}} \times \tilde{Y}] = \lim_{\varepsilon \to 0} \liminf_{j \to \infty} \mathbb{P}_{q_j}[\tilde{\mathcal{C}}_{\leq \varepsilon} \times \tilde{Y}_{\leq \varepsilon}]\). For
any $j > 1/\varepsilon$,

$$\mathbb{P}_{q_j}[\hat{C}_{\leq \varepsilon} \times \hat{Y}_{\leq \varepsilon}]$$

$$= \sum_{(x, r) \in (X_j \times \hat{R}_j) \cap \hat{C}_{\leq \varepsilon}} \frac{\tilde{\mu}_j(m_P)[x, (r, M)]}{\sum_{(x', r') \in \tilde{X}_j \times \tilde{R}_j} \tilde{\mu}_j(m_P)[x', (r', M)]} \beta_{j,A}(x, r)[\hat{Y}_{\leq \varepsilon}]$$

$$\leq \mathbb{E}_q \left[ \mathbb{I}((x, r) \in \hat{C}) \sum_{(x', r') \in (\tilde{X}_j \times \tilde{R}_j) \cap \hat{C}_{\leq \varepsilon}} \frac{1}{\sum_{(x''', r''') \in \tilde{X}_j \times \tilde{R}_j} \frac{1}{\frac{1}{2} \sum_{(x''', r''') \in \tilde{X}_j \times \tilde{R}_j} \frac{1}{\frac{1}{2} \beta_{j,A}(x', r')[\hat{Y}_{\leq \varepsilon}]} \frac{1}{\frac{1}{2} \beta_{j,A}(x', r')[\hat{Y}_{\leq \varepsilon}]} \frac{1}{\frac{1}{2} \beta_{j,A}(x', r')[\hat{Y}_{\leq \varepsilon}]} \right]$$

$$+ \mathbb{P}_{q_j}[\hat{C}_{\leq 2\varepsilon} \setminus \hat{C}],$$

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 $2\varepsilon$ from $\hat{C}$ contributes positive probability to any $(x', r') \in (\tilde{X}_j \times \tilde{R}_j) \cap \hat{C}_{\leq \varepsilon}$. Similarly,

$$\mathbb{P}_{q_j}[\hat{C}_{\leq \varepsilon} \times \hat{Y}_{\leq \varepsilon}]$$

$$= \sum_{(x, r) \in (X_j \times \hat{R}_j) \cap \hat{C}_{\leq \varepsilon}} \frac{\tilde{\mu}_j(m_P)[x, (r, M)]}{\sum_{(x', r') \in \tilde{X}_j \times \tilde{R}_j} \tilde{\mu}_j(m_P)[x', (r', M)]} \beta_{j,A}(x, r)[\hat{Y}_{\leq \varepsilon}]$$

$$\geq \mathbb{E}_q \left[ \mathbb{I}((x, r) \in \hat{C}) \sum_{(x', r') \in (\tilde{X}_j \times \tilde{R}_j) \cap \hat{C}_{\leq \varepsilon}} \frac{1}{\sum_{(x''', r''') \in \tilde{X}_j \times \tilde{R}_j} \frac{1}{\frac{1}{2} \sum_{(x''', r''') \in \tilde{X}_j \times \tilde{R}_j} \frac{1}{\frac{1}{2} \beta_{j,A}(x', r')[\hat{Y}_{\leq \varepsilon}]} \frac{1}{\frac{1}{2} \beta_{j,A}(x', r')[\hat{Y}_{\leq \varepsilon}]} \frac{1}{\frac{1}{2} \beta_{j,A}(x', r')[\hat{Y}_{\leq \varepsilon}]} \right]$$

$$- \mathbb{P}_{q_j}[\hat{C}_{\leq 2\varepsilon} \setminus \hat{C}],$$

where the last inequality comes from the previously established inequality for $\mathbb{P}_{q_j}[\hat{C}_{\leq \varepsilon} \times \hat{Y}_{\leq \varepsilon}]$. Since $\lim_{\varepsilon \to 0} \mathbb{P}_{q_j}[\hat{C}_{\leq 2\varepsilon} \setminus \hat{C}] = 0$, we thus have $\mathbb{P}_{q_j}[\hat{C} \times \hat{Y}] = \lim_{\varepsilon \to 0} \lim \inf_{j \to \infty} \mathbb{P}_{q_j}[\hat{C}_{\leq \varepsilon} \times \hat{Y}_{\leq \varepsilon}]$. 

21
\( \hat{Y}_{\leq \varepsilon} \geq (1 - \eta) \rho \lim_{\varepsilon \to 0} \lim \inf_{j \to \infty} \mathbb{P}_{q_{j}}[\hat{C} \times \hat{Y}] = (1 - \eta) \rho \mathbb{P}_{q_{j}}[\hat{C} \times \hat{Y}] \). Moreover, since

\[
\mathbb{P}_{q'}[\hat{C}] = \mathbb{E}_{q'}[\mathbb{I}_{\hat{C}}(x, r)]
\]

\[
= \mathbb{E}_{q} \left[ \frac{\tilde{f}_{m_{p}}(x, r)}{f_{m_{p}}(x, r)} \mathbb{I}_{\hat{C}}(x, r) \right]
\]

\[
\leq (1 + \eta) \rho \mathbb{E}_{q}[\mathbb{I}_{\hat{C}}(x, r)]
\]

\[
= (1 + \eta) \rho \mathbb{P}_{q}[\hat{C}],
\]

we obtain

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

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

\[\blacksquare\]

**Proof of Lemma 2.** The mechanism \((\tilde{\mu}, M_{P})\) has the same sequential continuation equilibrium outcomes as \((\mu, M_{P})\). 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 \((\tilde{\lambda}_{j}, \pi_{j, \theta_{1}}, \ldots, \pi_{j, \theta_{N}}, A_{j}, \beta_{j, \lambda})\) 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 \((\tilde{\lambda}^{*}, \pi_{\theta_{1}}^{*}, \ldots, \pi_{\theta_{N}}^{*}, A^{*}, \beta_{\lambda}^{*})\) 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
\[ \pi^*_\theta = \lim_{j \to \infty} \pi_{j,\theta} \text{ for all } \theta \in \Theta. \] For any \( M \in P_+ (M_P) \) such that \( M(m_P) = M \) for some \( m_P \in M_P \), let \( p_j(M) \in \Delta(\Theta \times X \times \hat{R} \times Y) \) be the conditional distribution given \( M \) that is induced by this assessment, and let \( p^*(M) = \lim_{j \to \infty} p_j(M) \). The agent’s belief updating rule \( \Lambda^* \) is such that \( \Lambda^*(x, (r, M(x, r))) \) equals the conditional distribution of \( \theta \) given \( (x, r) \) under \( p^*(M(x, r)) \) for all \( (x, r) \in \bigcup_{m_P \in M_P} \text{supp}(\mu(m_P)) \). Likewise, the agent’s action rule \( \beta^*_A \) is such that \( \beta^*_A(x, (r, M(x, r))) \) equals the conditional distribution of \( y \) given \( (x, r) \) under \( p^*(M(x, r)) \) for all \( (x, r) \in \bigcup_{m_P \in M_P} \text{supp}(\mu(m_P)) \). The construction of \( (\hat{\lambda}^*, \pi^*_\theta_1, \ldots, \pi^*_\theta_N, \Lambda^*, \beta^*_A) \) guarantees that it is consistent.

We now argue that \( (\hat{\lambda}^*, \pi^*_\theta_1, \ldots, \pi^*_\theta_N, \Lambda^*, \beta^*_A) \) constitutes a contracting equilibrium after \( (\hat{\mu}, M_P) \) 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 \( \hat{\mu}_j(m_P) \) and the agent responding according to \( \beta^*_A \). Likewise, let \( q^*(m_P) \in \Delta(X \times Y) \) be the distribution that results from \( \hat{\mu}(m_P) \) and the agent responding according to \( \beta^*_A \). 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 \( (\pi_{j,\theta_1}, \ldots, \pi_{j,\theta_N}) \) are optimal given the other’s play, it follows that the message choices prescribed in \( (\pi^*_\theta_1, \ldots, \pi^*_\theta_N) \) 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 \( \beta^*_A \) assigns probability 1 to best responses to their posterior beliefs about the principal’s type.

\[ \blacksquare \]

OA.8 Proof of Lemma 4

Lemma 4. There is a profile of mechanism proposal distributions \( \{\mu^*_\theta\}_\theta \in \Theta \) 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 \( (\alpha, x, y) \) for each principal type that corresponds to a single sequential continuation equilibrium after \( (\mu, M_P) \) is proposed such that

1. There is a regular conditional probability distribution obtained from \( \lambda \) and \( \{\mu^*_\theta\}_\theta \) 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 \((\mu, M_P) \in M\),

2. \(U(\theta, \tau^*(\mu, M_P)) \leq U(\theta, p)\) for all \(\theta \in \Theta\) and \((\mu, M_P) \in M\), and

3. \(\{\nu_{j,\theta}\}_{\theta \in \Theta}\) combined with \(\tau^*(\mu, M_P)\) following the proposal of each \((\mu, M_P) \in M\) induces the same distribution over \((\theta, \alpha, x, y)\) 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 \((\theta, \alpha, x, y)\) as in \(p\) occurring in a contracting equilibrium outcome. For any \((x_1, x_2) \in X \times X\), let \(\mu_{(x_1, x_2)} : \{1, 2\} \to \Delta(X \times N)\) 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 \(\Sigma = \{(x_1, x_2, \tilde{\lambda}, \pi_\theta_1, ..., \pi_\theta_N, \alpha, \beta) \in X^2 \times \Delta(\Theta) \times \Delta(\{1, 2\}) \times [0, 1] \times \Delta(Y)^2 : (\tilde{\lambda}, \pi_\theta_1, ..., \pi_\theta_N, \alpha, \beta) \in \sigma(x_1, x_2)\}\). Observe that \(\Sigma\) is a compact subset of \(X^2 \times \Delta(\Theta) \times \Delta(\{1, 2\}) \times [0, 1] \times \Delta(Y)^2\).

Let

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

We will say that there is obedient play following the proposal of the mechanism in \(M_{bin}^*\) corresponding to \((x_1, x_2, \tilde{\lambda}, \pi_\theta_1, ..., \pi_\theta_N, \alpha, \beta) \in \Sigma\) if each principal type \(\theta\) plays according to \(\pi_\theta\) and the agent plays according to \((\alpha, \beta)\). For every \(\theta\) and \((\mu, \{1, 2\}) \in M_{bin}^*\), 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 \((\alpha, x, y)\) 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 \(\{\nu_{j,\theta}\}_{\theta \in \Theta} \subset \Delta(M_{bin}^*)\) such that

1. There is a regular conditional probability distribution obtained from \(\lambda\) and \(\{\nu_{j,\theta}\}_{\theta \in \Theta}\) 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 $\mathcal{M}^{\text{bin}^*}$ corresponding to $(x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, \ldots, \pi_{\theta_N}, \alpha, \beta)$.

2. $\{\rho_{j,\theta}^*\}_{\theta \in \Theta}$ combined with the principal and agent playing obediently for each mechanism in $\mathcal{M}^{\text{bin}^*}$ induces the same distribution over $(\theta, \alpha, x, y)$ as outcome $p_j$.

3. $U(\theta, \gamma_{\text{obed}}(\mu, \{1, 2\})) \leq U(\theta, p_j)$ for all $\theta \in \Theta$ and $(\mu, \{1, 2\}) \in \cup_{\theta' \in \Theta} \text{supp}(\rho_{j,\theta'})$.

We prove Lemma OA 9 in the following way. For any mechanism $(\mu, M_P)$ that is proposed with positive probability under $p_j$, we construct a joint distribution over principal types and mechanisms in $\mathcal{M}^{\text{bin}^*}$ that, when coupled with obedient play, leads to the same distribution over $(\theta, \alpha, x, y)$ as the conditional distribution of $p_j$ given the proposal of $(\mu, M_P)$. The key will be to pair off the various actions that occur with positive probability in $p_j$ after the acceptance of $(\mu, M_P)$ 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 $(\mu, M_P)$ 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 $\mathcal{M}^{\text{bin}^*}$ identified separately for each on-path mechanism in $\mathcal{M}_j$ to obtain a profile of proposal distributions over mechanisms in $\mathcal{M}^{\text{bin}^*}$ that results in an outcome of $p_j$.

Proof. Consider the equilibrium $((\rho_{j,k,\theta_1}^*, \pi_{j,k,\theta_1}^*), \ldots, (\rho_{j,k,\theta_N}^*, \pi_{j,k,\theta_N}^*), (\alpha_{j,k}^* (\cdot), \beta_{j,k}^* (\cdot)))$ of the $(j, k)$ game, and, restricting attention to a convergent subsequence if necessary, let

$$(\rho_{j,\theta_1}^*, \pi_{j,\theta_1}^*), \ldots, (\rho_{j,\theta_N}^*, \pi_{j,\theta_N}^*), (\alpha_{j,k}^* (\cdot), \beta_{j,k}^* (\cdot))) = \lim_{k \to \infty} ((\rho_{j,k,\theta_1}^*, \pi_{j,k,\theta_1}^*), \ldots, (\rho_{j,k,\theta_N}^*, \pi_{j,k,\theta_N}^*), (\alpha_{j,k}^* (\cdot), \beta_{j,k}^* (\cdot))).$$

Fix an arbitrary $(\mu, M_P) \in \mathcal{M}_j$ that is proposed with positive probability under $(\rho_{j,\theta_1}^*, \ldots, \rho_{j,\theta_N}^*)$. Let $\tilde{\lambda}_j^* (\mu, M_P) \in \Delta(\Theta)$ be the posterior distribution over the principal’s type conditional on the proposal of $(\mu, M_P)$. Further, let $(x_1, y_1), \ldots, (x_M, y_M) \in \mathcal{X}_j$.
$X_j \times \Delta(Y)$ be the pairs of principal actions and agent action distributions that occur with positive probability under $(\pi^*_{j,\theta_1}(\mu, M_P), \ldots, \pi^*_{j,\theta_N}(\mu, M_P), \beta^*_{j}(\cdot))$ when $(\mu, M_P)$ is accepted. For every $\theta \in \Theta$ and $m \in \{1, \ldots, M\}$, we use $q_{j,(x_m,y_m)}(\theta)$ to denote the probability of $(x_m, y_m)$ conditional on type $\theta$ and $(\mu, M_P)$ being accepted under $(\pi^*_{j,\theta_1}(\mu, M_P), \ldots, \pi^*_{j,\theta_N}(\mu, M_P), \beta^*_{j}(\cdot))$.

The $k \to \infty$ limit of the expected utility of the agent from accepting $(\mu, M_P)$ in the $(j, k)$ equilibrium is thus $V_j(\mu, M_P) = \sum_{m \in \{1, \ldots, M\}} \sum_{\theta \in \Theta} \lambda^*_{j}(\theta|\mu, M_P)q_{j,(x_m,y_m)}(\theta)\mathbb{E}_{y_m}[V(\theta, x_m, y)]$.

Observe that there are collections of pairs $\{(m_{l,1}, m_{l,2})\}_{l \in \{1, \ldots, L\}}$ and $\{(s_{l,1}, s_{l,2})\}_{l \in \{1, \ldots, L\}}$ for some $L \in \mathbb{N}$ such that

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

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

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

$$\text{sign}\left(s_{l,1} \sum_{\theta \in \Theta} \lambda^*_{j}(\theta|\mu, M_P)q_{j,l,1}(\theta)\mathbb{E}_{y_{l,1}}[V(\theta, x_{l,1}, y)] + s_{l,2} \sum_{\theta \in \Theta} \lambda^*_{j}(\theta|\mu, M_P)q_{j,l,2}(\theta)\mathbb{E}_{y_{l,2}}[V(\theta, x_{l,2}, y)]\right)$$

$$= \text{sign}(V_j(\mu, M_P)),$$

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

For each $l \in \{1, \ldots, L\}$, we create a mechanism $(\mu_l, \{1, 2\}) \in \mathcal{M}^b_{\text{bins}}$ in which $m_P = 1$ maps to $x_{l,1}$ and a recommended action distribution of $y_{l,1}$, and $m_P = 2$ maps to $x_{l,2}$ and a recommended action distribution of $y_{l,2}$. We will have each type $\theta$ 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 $\theta$ 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 $\theta$ play
$m_P = 1$ whenever they weakly prefer $(x_{l,1}, y_{l,1})$ to $(x_{l,2}, y_{l,2})$ 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 $(\mu, M_P)$ 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 $\alpha_j^*(\mu, M_P)$, and indeed we will have them do so. This means that the conditional distribution of $(\alpha, x, y)$ given type $\theta$ is exactly as when $(\mu, M_P)$ 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

$$\text{supp}(\mu(1)) = (x_{l,1}, (\tilde{\lambda}_{j,l}^*, \pi_{j,l,1}^*, \pi_{j,l,2}^*, \alpha_{j,l}^*(\mu, M_P), y_{l,1})), \text{ and }$$

$$\text{supp}(\mu(2)) = (x_{l,2}, (\tilde{\lambda}_{j,l}^*, \pi_{j,l,1}^*, \pi_{j,l,2}^*, \alpha_{j,l}^*(\mu, M_P), y_{l,2})), \text{ where, for each } \theta \in \Theta,$$

$$\tilde{\lambda}_{j,l}^*(\theta) = \frac{\tilde{\lambda}_{j,l}^*(\theta \mu, M_P)(s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta))}{\sum_{\theta' \in \Theta} \tilde{\lambda}_{j,l}^*(\theta' \mu, M_P)(s_{l,1}q_{j,l,1}(\theta') + s_{l,2}q_{j,l,2}(\theta'))},$$

$$\pi_{j,l,1}^*(\theta) = \begin{cases} \frac{s_{l,1}q_{j,l,1}(\theta)}{s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta)} & \text{if } s_{l,1}q_{j,l,1}(\theta) + s_{l,2}q_{j,l,2}(\theta) > 0 \\ 0 & \text{if } E_{y_{l,1}}[U(\theta, x_{l,1}, y)] < E_{y_{l,2}}[U(\theta, x_{l,2}, y)] \\ 1 & \text{otherwise} \end{cases},$$

$$\pi_{j,l,2}^*(\theta) = 1 - \pi_{j,l,1}^*(\theta).$$

Note that $\tilde{\lambda}_{j,l}^*$ is the posterior distribution induced by having each type $\theta$ 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 $\tilde{\lambda}_{j,l}^*(\mu, M_P)$, while $\pi_{j,l,\theta}^*$ gives the mechanism selection probabilities for each type $\theta$ as discussed above.

We have already established that if each principal type $\theta$ proposes the binary mec-
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 \((\mu, M_P)\). We now argue that obedient play is consistent with these proposal probabilities and sequential continuation equilibrium. To see this, observe that, the only \((x, y)\) pairs that occur with positive probability conditional on type \(\theta\) are those which occurred with positive probability conditional on \(\theta\) under the \(k \to \infty\) limit of the \((j,k)\) equilibrium outcome following the acceptance of \((\mu, M_P)\). Because the principal’s trembles vanish in the \(k \to \infty\) limit, these must be the optimal \((x, y)\) 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 \((x_{l,1}, y_{l,2})\) is the same as the posterior after \((x_{m,1}, y_{l,2})\) in the \(k \to \infty\) of the \((j,k)\) equilibrium. Likewise, for \((x_{l,1}, y_{l,2})\). 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 \(\alpha^*_j(\mu, M_P)\) is optimal for the agent.

We have thus obtained a profile of mechanism proposal distributions \(\{m_{j,\mu, M_P, \theta}\}_{\theta \in \Theta}\) 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 \((\mu, M_P)\) being proposed. This can be done for every mechanism in \(\mathcal{M}_j^{on} = \{(\mu, M_P) \in \mathcal{M} : \exists \theta \in \Theta \text{ s.t. } r_{j,\theta}^*(\mu, M_P) > 0\}\), 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 \(\theta\) weighted by the probability of proposing each \((\mu, M_P)\) then gives \(r_{j,\theta} = \sum_{(\mu, M_P) \in \mathcal{M}_j} r_{j,\theta}^*(\mu, M_P) r_{j,\mu, M_P, \theta}\). The profile of these mechanism proposal distributions satisfies all the conditions of Lemma OA 9.

\[\square\]

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

1. There is a regular conditional probability distribution obtained from \(\lambda\) and \(\{r_{j,\theta}\}_{\theta \in \Theta}\)
that, for every \((x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, \ldots, \pi_{\theta_N}, \alpha, \beta) \in \Sigma\), induces \(\tilde{\lambda}\) as the belief over the principal’s type following the proposal of the mechanism in \(\mathcal{M}^{\text{bin}*}\) corresponding to \((x_1, x_2, \tilde{\lambda}, \pi_{\theta_1}, \ldots, \pi_{\theta_N}, \alpha, \beta)\).

2. \(\{\rho_{\theta}\}_{\theta \in \Theta}\) combined with the principal and agent playing obediently for each mechanism in \(\mathcal{M}^{\text{bin}*}\) induces the same distribution over \((\theta, \alpha, x, y)\) as outcome \(p\).

3. \(U(\theta, \tau_{\text{obed}}(\mu, \{1, 2\})) \leq U(\theta, p)\) for all \(\theta \in \Theta\) and \((\mu, \{1, 2\}) \in \cup_{\theta \in \Theta} \text{supp}(\rho_{\theta'})\).

Proof. For each \(\theta \in \Theta\), let \(\rho_{\theta} \in \Delta(M)\) be a limit point of the sequence \(\{\rho_{j,\theta}\}_{j \in \mathbb{N}}\).

Without loss, suppose that \(\lim_{j \to \infty} \rho_{j,\theta} = \rho_{\theta}\) for all \(\theta \in \Theta\). Since \(\Delta(M^{\text{bin}*})\) is closed, it follows that each \(\rho_{\theta} \in \Delta(M^{\text{bin}*})\).

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

\[
\mathbb{P}_{p_{\Theta \times \mathcal{M}^{\text{bin}*}}}[\{\theta\} \times \overline{\mathcal{M}}] = \mathbb{E}_{\rho}[\tilde{\lambda}(\theta | (M, \mu_P)) \mathbb{I}_{\overline{\mathcal{M}}}(M, \mu_P)] \\
\geq \limsup_{j \to \infty} \mathbb{E}_{\rho_j}[\tilde{\lambda}(\theta | (M, \mu_P)) \mathbb{I}_{\overline{\mathcal{M}}}(M, \mu_P)] \\
= \limsup_{j \to \infty} \mathbb{P}_{p_{j,\Theta \times \mathcal{M}^{\text{bin}*}}}[\{\theta\} \times \overline{\mathcal{M}}].
\]

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

Now we show that Condition 2 holds. Fix \(\theta \in \Theta\) and let \(\overline{X}, \overline{A}, \text{ and } \overline{Y}\) be arbitrary closed subsets of \(X, [0, 1],\) and \(Y\), respectively. Since \(\tilde{\lambda}(\theta | (\mu, \{1, 2\})) \tau_{\text{obed}}(\overline{X} \times \overline{A} \times \overline{Y} | \theta, \mu, \{1, 2\})\) is a continuous function of \((\mu, \{1, 2\}) \in \mathcal{M}^{\text{bin}*}\), it follows that
\[
E_\eta[\tilde{\lambda}(\theta)(\mu, \{1, 2\}))\tau^{\text{obed}}(X \times A \times Y|\theta, \mu, \{1, 2\})] = \lim_{j \to \infty} E_\eta[\tilde{\lambda}(\theta)(\mu, \{1, 2\}))\tau^{\text{obed}}(X \times A \times Y|\theta, \mu, \{1, 2\})] = \lim_{j \to \infty} P \circ\pi \circ \theta \in \Theta, \supp(m) \in \Theta, \tau^{\text{obed}}(X \times A \times Y|\theta, \mu, \{1, 2\})].
\]

Because \(\theta, X, A, \) and \(Y\) are arbitrary, we conclude that \(\{m_\theta\}_{\theta \in \Theta}\), together with obedient play, induces \(p\).

Finally, since \(U(\theta, \tau^{\text{obed}}(\mu, \{1, 2\})) \leq U(\theta, p)\) for all \(\theta \in \Theta\) and \((\mu, \{1, 2\}) \in \bigcup_{\theta' = \Theta} \supp(n_{\theta'\theta})\), standard continuity arguments show that \(U(\theta, \tau^{\text{obed}}(\mu, \{1, 2\})) \leq U(\theta, p)\) for all \(\theta \in \Theta\) and \((\mu, \{1, 2\}) \in \bigcup_{\theta' = \Theta} \supp(n_{\theta'\theta})\).

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

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

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 \((\mu, \{1, ..., M\}) \in M^{\text{rev}, M}\), 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 \((\mu, \{1, ..., M\}) \in M^{\text{rev}, M}\) can be approximated to arbitrary accuracy by some sequence of mechanisms \(\{(\mu_j, \{1, ..., M\})\}_{j \in \mathbb{N}}\), where \((\mu_j, \{1, ..., M\}) \in M_j\) 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 \(M_j\), 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 $\tau^{M*} : \mathcal{M}^{rev,M} \to \Delta(\Theta) \times \Delta(\{1,\ldots,M\})^\Theta \times \Delta([0,1]) \times \Delta(Y)^M$ such that, for every $(\mu,\{1,\ldots,M\}) \in \mathcal{M}^{rev,M}$, $\tau^{M*}(\mu,\{1,\ldots,M\})$ is a sequential continuation equilibrium after $(\mu,\{1,\ldots,M\})$ is proposed that gives every principal type a weakly lower payoff than $U(\theta,p)$.

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

By Lemma OA 12, $\sigma^{M*}(\mu,\{1,\ldots,M\})$ is non-empty for all $(\mu,\{1,\ldots,M\}) \in \mathcal{M}^{rev,M}$. Additionally, standard arguments show that $\sigma^{M*} : \mathcal{M}^{rev,M} \to \Delta(\Theta) \times \Delta(\{1,\ldots,M\})^\Theta \times \Delta([0,1]) \times \Delta(Y)^M$ 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 $\sigma^{M*}$ and hence the desired $\tau^{M*}$.

**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 $(\alpha,x,y)$ for each principal type that corresponds to a single sequential continuation equilibrium outcome after $(\mu,M_P)$ 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 $\tau^{M*} : \mathcal{M}^{rev,M} \to \Delta(\Theta) \times \Delta(\{1,\ldots,M\})^\Theta \times \Delta([0,1]) \times \Delta(Y)^M$ 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 $(\alpha,x,y)$ corresponding to $\tau^{M*}(\mu,\{1,\ldots,M\})$ for each $(\mu,\{1,\ldots,M\})$. 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 $\mathcal{X}^{M}(\mu,M_P) = (x_1,\ldots,x_M)$ 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 \( f(\mu, M_P) = X^M(\mu, M_P) \) 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 \((\mu, M_P)\) 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 \((\alpha, x, y)\) for each principal type that corresponds to a single sequential continuation equilibrium outcome after \((\mu, M_P)\) 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 \( \bigcup_{M \in \mathbb{N}} \mathcal{M}^{eff,M} = \mathcal{M} \), 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 \( \{\rho^\theta\}_{\theta \in \Theta} \) 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

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

By construction, \( \tau^* \) is measurable. Additionally, Lemma OA 10 guarantees that Conditions 1 and 3 of Lemma 4 hold with this \( \{\rho^\theta\}_{\theta \in \Theta} \) and \( \tau^* \), 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. \( \text{marg}_\Theta p_k = \lambda_k \) 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 \( (\mu, M_P) \) is proposed that gives every principal type a weakly lower payoff than \( p_k \).

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

\[
V^\dagger_k = -((|\Theta|V + 1)/(k - |\Theta|)).
\]  

(OA 7)

Note that \( V^\dagger_k \) is such that \( (1 - (|\Theta| - 1)/k)V^\dagger_k + (|\Theta| - 1)V/k < -1/k \). This means that, if the agent's conditional expected utility given some principal type is weakly lower than \( V^\dagger_k \) 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

\[
\mathcal{M}_j = \left\{ (\mu, M_P) \in \mathcal{M} : \begin{array}{l}
(1) |M_P| \leq j, \\
(2) \forall m_P \in M_P, x \in X_j, t \in T_j, r \in R_j, \exists k \in \mathbb{N} \text{ s.t. } \mu((x, t), r | m_P) = \frac{k}{j|X_j||T_j||R_j|} \end{array} \right\}
\]

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 $\theta$ principal in the $j$-th game. Part of this player’s choice is over which mechanisms to propose. We force $\theta$ to propose almost all mechanisms with positive probability. The exceptions are mechanisms which commit to some $\chi$ as the distribution over principal actions and some $\theta' \neq \theta$ as the recommendation received by the agent; $\theta$ 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 $\mathcal{M}_{j, \theta'}^c = \{(\mu_{\chi, \theta'}, \emptyset) : \chi \in \Delta_j(X_j \times T_j), \theta' \neq \theta\}$ be the set of mechanisms in the $j$-th game that commit to some $\chi$ as the distribution over principal actions and $\theta'$ as the recommendation received by the agent. Additionally, let $\mathcal{M}_{j, \theta}^0 = \bigcup_{\theta' \neq \theta} \mathcal{M}_{j, \theta'}^c$. The distribution over mechanism proposals used by $\theta$ must belong to

$$\Delta_{j, \theta}(\mathcal{M}_j) = \left\{ \pi \in \Delta(\mathcal{M}_j) : \begin{array}{l}
(1) \pi[(\mu, M_P)] \geq \frac{1}{j|M_j|} \forall (\mu, M_P) \in \mathcal{M}_j \setminus \mathcal{M}_{j, \theta}^0, \\
(2) \pi[(\mu, M_P)] = 0 \forall (\mu, M_P) \in \mathcal{M}_{j, \theta}^0 \end{array} \right\}.$$

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

$$\Pi_{j, P}(\mu, M_P) = \left\{ \pi_P \in \Delta(M_P) : \pi_P[m_P] \geq \frac{1}{j|M_P|} \forall m_P \in M_P \right\}.$$
A valid strategy for \( \theta \) in the \( j \)-th game is any pair \((\pi_\theta, \pi_\theta(\cdot))\) consisting of a \(\pi_\theta \in \Delta_{j,\theta}(M_j)\) and a rule \(\pi_\theta(\cdot)\) for how to play when an arbitrary mechanism is accepted that satisfies \(\pi_\theta(\mu, M_P) \in \Pi_{j,P}(\mu, M_P)\).

The strategy space of the agent is unaltered from the principal-agent game, aside from the addition of trembles. For every mechanism \((\mu, M_P)\), we require the probability \(\alpha\) 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 \((\mu, M_P)\) and principal action-recommendation pair \((x, r)\), the agent’s choice of action must be a distribution belonging to

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

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

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 \(\zeta\). For any \(\theta \in \Theta\), let \(\tilde{U}_j(\theta, \mu, M_P, \alpha, \pi_M, \beta_A)\) and \(\tilde{V}_j(\theta, \mu, M_P, \alpha, \pi_M, \beta_A)\) be the unmodified expected payoffs to the principal and agent, respectively, when the principal’s type is \(\theta\), the mechanism \((\mu, M_P) \in M_j\) is proposed, the agent uses the acceptance probability rule \(\alpha \in [0, 1]^{M_j}\), and subsequent play is governed by the rules \(\pi_M\) and \(\beta_A\).

The agent’s payoff is

\[
V_j(\zeta) = \sum_{\theta \in \Theta} \lambda'(\theta) \left[ \sum_{(\mu, M_P) \in M_j \setminus \cup_{\mu' \in \Theta} M_j^{\mu',\theta}} n^\alpha(\mu, M_P) \tilde{V}_j(\theta, \mu, M_P, \alpha, \pi_M, \beta_A) \right].
\]
This is precisely the agent’s total expected utility from play over mechanisms in $\mathcal{M} \setminus (\cup_{\theta \in \Theta} \mathcal{M}_{j,\theta}^c)$. 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 $\mathcal{M} \setminus (\cup_{\theta \in \Theta} \mathcal{M}_{j,\theta}^c)$.

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

$$
\hat{U}_j(\theta, \zeta) = \sum_{(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}_{j,\theta'}^c)} r_{\theta}^{M_P}[(\mu, M_P)] \hat{U}_j(\theta, \mu, M_P, \alpha, \pi_\theta, \beta_A)
$$

and

$$
\hat{V}_j(\theta, \zeta) = \sum_{(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}_{j,\theta'}^c)} r_{\theta}^{M_P}[(\mu, M_P)] \hat{V}_j(\theta, \mu, M_P, \alpha, \pi_\theta, \beta_A)
$$

would be the total expected utilities of the principal and agent, respectively, when the principal’s type is $\theta$, the principal follows the mechanism proposal rule $r_{\theta}$, and the play that follows a mechanism proposal of $(\mu, M_P) \in \mathcal{M}_j$ proceeds as follows: For $(\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta} \mathcal{M}_{j,\theta'}^c)$, play proceeds according to the rules $\alpha, \pi_P$, and $\beta_A$; for $(\mu, \theta', \{0\}) \in \mathcal{M}_{j,\theta'}^c$, 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 $\theta$ to propose any $(\mu, \theta', \{0\}) \in \mathcal{M}_{j,\theta'}^c$ whenever either some principal type $\theta' \neq \theta$ would prefer to propose $(\mu, \theta', \{0\})$ (and have the agent respond according to $y^*(\theta', x)$) to their outcome or the agent gets a low expected utility conditional on $\theta$. Let $A > 2 \max_{(\theta, t, y)} |u(\theta, x, y) - t|$, 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 $c_{j,\theta,\zeta} : \mathcal{M}_{j,\theta}^c \to \mathbb{R}_+$
be the “cost” function given by

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

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

\[ U_j(\theta, \zeta) = \hat{U}_j(\theta, \zeta) - \sum_{(\mu_{X,\theta}, \{0\}) \in M^c_{j,\theta}} p_M^{\mu_{X,\theta}}(\{0\}) \left( c_{j,\theta,\zeta}(\mu_{X,\theta}, \{0\}) - \frac{1}{k} \right). \]

The important feature of the cost terms is that \( \theta \) would never want to propose a \((\mu_{X,\theta}, \{0\}) \in M^c_{j,\theta} \) if either \( \sum_{x,t} \chi[x,t](u(\theta', x, y^*(\theta, x)) - t) \geq \hat{U}_j(\theta', \zeta) + 1/j \) for some \( \theta' \neq \theta \) or \( \hat{V}_j(\theta, \zeta) \leq V_k^+ - 1/j \). On the other hand, if \( \sum_{x,t} \chi[x,t](u(\theta', x, y^*(\theta, x)) - t) \leq \hat{U}_j(\theta', \zeta) \) holds for all \( \theta' \neq \theta \) and \( \hat{V}_j(\theta, \zeta) \geq V_k^+ \), then the artificial cost from proposing \((\mu_{X,\theta}, \{0\})\) is 0 for \( \theta \). In this case, \( \theta \) 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 \( \zeta \).

**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 \((\mu, M_P) \in M_j \setminus (\cup_{\theta \in \Theta} M^c_{j,\theta})\), the principal types and agent play as they do in the Nash equilibrium, i.e. \( \theta \) plays according to \( \pi_\theta(\mu, M_P) \) while the agent accepts the mechanism with probability \( \alpha(\mu, M_P) \) and then plays according to \( \beta_A(\mu, M_P) \); for any \((\mu_{X,\theta}, \{0\}) \in M^c_{j,\theta} \), 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 \( \theta \), for \( V_k^\dagger \) defined in (OA 7), since \( \lim_{k \to \infty} V_k^\dagger = 0 \). To see that \( \mathbb{E} p_k[ v(\theta, x, y) + g(t) | \theta ] \geq V_k^\dagger \) 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 \( \theta \) proposes mechanisms in \( \mathcal{M}_{j,\theta}^\epsilon \) 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 \( \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta \mathcal{M}_{j,\theta'}}) \) 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 \( \mathcal{M}_j \setminus (\cup_{\theta' \in \Theta \mathcal{M}_{j,\theta'}}) \) 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 \( \mathcal{M}_{j,\theta'}^\epsilon \), \( \theta \) must (weakly) prefer to not play said mechanism given the
prevailing outcome. Moreover, since \( \theta \) can always mimic the play of \( \theta' \) in mechanisms in \( \mathcal{M}_j \setminus \mathcal{M}^c_{j,\theta'} \), it follows that, \( \theta \) must weakly prefer their conditional outcome under \( p_{j,k} \) to that of \( \theta' \) in the \( j \rightarrow \infty \) limit, which gives Condition 3.

We establish Condition 4 by induction over \( \theta_n \), beginning with \( \theta_N \) 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 \( x \neq x_o \), 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 \( \theta_N \) 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 \( \chi \) and plays \( y^*(\theta_N, x) \) in response to any \( x \), \( \theta_N \) 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 \( \theta_N \) is no less than \( V^\dagger_k \). Thus, the agent would obtain an expected utility that is weakly greater than \( V^\dagger_k \) from accepting a proposal of \( \chi \) by \( \theta_N \). So for sufficiently high \( j \), the type \( \theta_N \) principal can achieve a payoff in the \( j \)-th game that is uniformly bounded above their payoff from \( p_k \) by proposing some mechanism \( (\mu_{\chi',\theta}, \{0\}) \) where \( \chi'_j \in \Delta_j(X_j \times T_j) \) sufficiently closely approximates \( \chi' \), but this contradicts the fact that their payoff should be no more than that under \( p_k \) in the \( j \rightarrow \infty \) limit.

Since \( \mathbb{P}_{p_k}[y = y^*(\theta_N, x)|\theta_N, x \neq x_o] = 1 \), it follows from the fact that \( \lambda_k \) puts strictly positive probability on \( \theta_{N-1} \) 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 \( \theta_N \) case shows that this is not possible. Proceeding with this argument inductively by moving down the \( \theta_n \) establishes that Condition 4 holds for all \( \theta \).

To see why Condition 5 holds, note that, in the \( j \rightarrow \infty \) limit, every \( \theta \) must get a weakly higher payoff from \( p_{j,k} \) than \( 1/k \) less the payoff they would get from proposing any \( (\mu, M_F) \in \mathcal{M}_j \setminus \mathcal{M}^0_{j,\theta} \). Fix some \( r \in R_j \setminus \Theta \). For any \( \chi \in \Delta_j(X_j \times T_j) \),
every mechanism of the form $(\mu_{x,\theta}, \{0\})$ for some $\theta' \in \Theta$ can be identified with $(\mu_{x,r}, \{0\}) \in M_j \setminus (\cup_{\theta' \in \Theta} M_{j,\theta'})$. This means that, for every mechanism in $M_j$, 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 $\theta'$ 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 $M_j$ 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 $(\mu, M_P) \in M$ 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. $\text{marg}_{\Theta} p_k = \lambda_k$ for all $k \in \mathbb{N}$,
2. $\liminf_{k \to \infty} E_{p_k}[\alpha(v(\theta, x, y) + g(t))|\theta] \geq 0$ for all $\theta \in \Theta$,
3. $P_{p_k}[U(\theta, p_k) \geq \alpha(u(\theta, x, y^*(\theta', x)) - t)|\theta', x, t, \alpha] = 1$ for all $\theta, \theta' \in \Theta$ and $k \in \mathbb{N}$,
4. $P_{p_k}[y = y^*(\theta, x)|\theta, x \neq x_o] = 1$ for all $\theta \in \Theta$ and $k \in \mathbb{N}$, and
5. For each mechanism $(\mu, M_P) \in M$ and $k \in \mathbb{N}$, there is a sequential continuation equilibrium after $(\mu, M_P)$ 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

\[ M_j = \left\{ (\mu, M_P) \in \mathcal{M} : \begin{array}{l}
(1) \ |M_P| \leq |X_j||T_j||R_j|, \\
(2) \ \forall m_P \in M_P, \ \exists x \in X_j, t \in T_j, r \in R_j \text{ s.t. } \mu((x, t), r|m_P) = 1
\end{array} \right\} \]

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 \(\theta\) principal in the \(j\)-th game. Part of this player’s choice is over which mechanisms to propose. We force \(\theta\) 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; \(\theta\) is required to propose these mechanisms with 0 probability. Formally, let \(M^c_{j,\theta'} = \{((\delta((x, t), \theta'), \{0\}) : (x, t) \in X_j \times T_j, \theta' \neq \theta\} \) 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 \(\theta'\) as the recommendation received by the agent. Additionally, let \(M^0_{j,\theta} = \cup_{\theta' \neq \theta} M^c_{j,\theta'}\). The distribution over mechanism proposals used by \(\theta\) must belong to

\[ \Delta_{j,\theta}(\mathcal{M}_j) = \left\{ r_\theta \in \Delta(\mathcal{M}_j) : \begin{array}{l}
(1) \ r_\theta[(\mu, M_P)] \geq \frac{1}{|M_j|} \forall (\mu, M_P) \in \mathcal{M}_j \setminus M^0_{j,\theta}, \\
(2) \ r_\theta[(\mu, M_P)] = 0 \forall (\mu, M_P) \in M^0_{j,\theta}
\end{array} \right\}. \]

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

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

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

The strategy space of the agent is unaltered from the principal-agent game, aside from the addition of trembles. For every mechanism $(\mu, M_P)$, we require the probability $\alpha$ 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 $(\mu, M_P)$ and principal action-recommendation pair $(x, r)$, the agent’s choice of action must be a distribution belonging to

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

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

We now develop the payoffs of the various players for an arbitrary strategy profile $\zeta$. For any $\theta \in \Theta$, let $\tilde{U}_j(\theta, \mu, M_P, \alpha, \pi_P, \beta_A)$ and $\tilde{V}_j(\theta, \mu, M_P, \alpha, \pi_P, \beta_A)$ be the unmodified expected payoffs to the principal and agent, respectively, when the principal’s type is $\theta$, the mechanism $(\mu, M_P) \in \mathcal{M}_j$ is proposed, the agent uses the acceptance probability rule $\alpha \in [0, 1]^{\mathcal{M}_j}$, and subsequent play is governed by the rules $\pi_P$ and $\beta_A$. 
The agent’s payoff is

\[ V_j(\zeta) = \sum_{\theta \in \Theta} \lambda'(\theta) \left[ \sum_{(\mu, M_P) \in M_j \setminus (\cup_{\theta' \in \Theta} M_{j,\theta'}^c)} n_{\theta'}[(\mu, M_P)] \tilde{V}_j(\theta, \mu, M_P, \alpha, \pi_{\theta}, \beta_A) \right]. \]

This is precisely the agent’s total expected utility from play over mechanisms in \( M \setminus (\cup_{\theta \in \Theta} M_{j,\theta}^c) \). 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 \( M \setminus (\cup_{\theta \in \Theta} M_{j,\theta}^c) \).

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

\[ \hat{U}_j(\theta, \zeta) = \sum_{(\mu, M_P) \in M_j \setminus (\cup_{\theta' \in \Theta} M_{j,\theta'}^c)} n_{\theta'}[(\mu, M_P)] \hat{U}_j(\theta, \mu, M_P, \alpha, \pi_{\theta}, \beta_A) \]

\[ + \sum_{(\delta((x,t),\theta'),\{0\}) \in M_{j,\theta'}^c} n_{\theta'}[(\delta((x,t),\theta'),\{0\})](u(\theta, x, y^*(\theta, x)) - t) \]

\[ \hat{V}_j(\theta, \zeta) = \sum_{(\mu, M_P) \in M_j \setminus (\cup_{\theta' \in \Theta} M_{j,\theta'}^c)} n_{\theta'}[(\mu, M_P)] \hat{V}_j(\theta, \mu, M_P, \alpha, \pi_{\theta}, \beta_A) \]

\[ + \sum_{(\delta((x,t),\theta'),\{0\}) \in M_{j,\theta'}^c} n_{\theta'}[(\delta((x,t),\theta'),\{0\})](v(\theta, x, y^*(\theta, x)) + g(t)) \]

would be the total expected utilities of the principal and agent, respectively, when the principal’s type is \( \theta \), the principal follows the mechanism proposal rule \( n_{\theta'} \), and the play that follows a mechanism proposal of \( (\mu, M_P) \in M_j \) proceeds as follows: For \( (\mu, M_P) \in M_j \setminus (\cup_{\theta' \in \Theta} M_{j,\theta'}^c) \), play proceeds according to the rules \( \alpha, \pi_P \), and \( \beta_A \); for \( (\delta((x,t),\theta'),\{0\}) \in M_{j,\theta'}^c \), 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 \( \theta \) to propose any \((\delta((x,t),\theta'),\{0\}) \in M_{j,\theta}^c\) 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 \( \theta \).

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 \( c_{j,\theta,\zeta} : \mathcal{M}_{j,\theta}^c \to \mathbb{R}_+ \) be the “cost” function given by

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

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

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

The important feature of the cost terms is that \( \theta \) would never want to propose a \( (\delta((x,t), \{0\}) \in \mathcal{M}_{j,\theta}^c \) 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 \( (\delta((x,t), \{0\}) \) is 0 for \( \theta \). In this case, \( \theta \) 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 \( \zeta \).

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 \( (\mu, M_P) \in \mathcal{M}_j \setminus (\cup_{\theta \in \Theta} \mathcal{M}_{j,\theta}^c) \), the principal types and agent play as they do in the
Nash equilibrium, i.e. $\theta$ plays according to $\pi(\mu, M_P)$ while the agent accepts the mechanism with probability $\alpha(\mu, M_P)$ and then plays according to $\beta(\mu, M_P)$; for any $(\delta((x,t),\theta), \{0\}) \in \mathcal{M}_{j,\theta}^c$, 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 $\mathcal{M}_{j,\theta'}$, $\theta$ must (weakly) prefer to not play said mechanism given the prevailing outcome. Moreover, since $\theta$ can always mimic the play of $\theta'$ in mechanisms in $\mathcal{M}_j \setminus \mathcal{M}_{j,\theta}$, it follows that, for all $\varepsilon > 0$, $\theta$ must get a weakly higher payoff from their conditional outcome under $p_{j,k}$ than their payoff from the conditional outcome given $\theta'$, $x$, $t$, and $\alpha$ 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 $\theta_n$, beginning with $\theta_0$ as the base case. Suppose towards a contradiction that $\mathbb{P}_{p_k}[x \neq x_o | \theta_N] > 0$ and $\mathbb{P}_{p_k}[y = y^*(\theta_N, x) | \theta_N, x \neq x_o] = 0$. 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 $\theta$ 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, $\theta_N$ 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 \( \theta_N \). So for sufficiently high \( j \), the type \( \theta_N \) 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 \( P_{p_k} [y = y^*(\theta_N, x)|\theta_N, x \neq x_o] = 1 \), it follows from the fact that \( \lambda_k \) puts strictly positive probability on \( \theta_{N-1} \) that \( P_{p_k} [y > y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] = 0 \) (assuming \( P_{p_k} [x \neq x_o|\theta_{N-1}] > 0 \) so that this conditional probability is even relevant). Therefore, if \( P_{p_k} [y = y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] < 1 \), it must be that \( P_{p_k} [y < y^*(\theta_{N-1}, x)|\theta_{N-1}, x \neq x_o] > 0 \). The same argument as for the \( \theta_N \) case shows that this is not possible. Proceeding with this argument inductively by moving down the \( \theta_n \) establishes that Condition 4 holds for all \( \theta \).

\[ \square \]

**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 \( \theta \).
   (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 \( \theta \).
   (b) \( v(\theta, x, y^*(\tilde{\lambda}, x)) \) is strictly increasing in \( \theta \).

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

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

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

**Case 2:** Condition 3’ or 3’’ holds for \((x, \bar{\lambda})\). Consider \( t'_i = \alpha t + u(\bar{\theta}, x_i, y^*(\bar{\theta}, x_i)) - \alpha u(\bar{\theta}, x_i, y^*(\bar{\lambda}, x)) > t \). By construction, \( u(\bar{\theta}, x_i, y^*(\bar{\theta}, x_i)) - t'_i = U(\bar{\theta}, p), u(\theta, x_i, y^*(\bar{\theta}, x_i)) - t'_i < U(\theta, p) \) for all \( \theta < \bar{\theta} \), and \( v(\bar{\theta}, x, y^*(\bar{\theta}, x)) + g(t'_i) > 0 \). 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 \((\mu, M_P, M_A)\). 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 \(\theta\) and \(x\) is drawn according to \(\mu(m_P, m_A)\).

Let \(\Psi : \Delta(M_P) \times \Delta(M_A) \Rightarrow \Delta(\Theta)\) be the correspondence given by
\[
\Psi(\pi_{\theta_1}, \ldots, \pi_{\theta_N}, \pi_A) = \Delta(\arg \min_{\theta \in \Theta} \mathbb{E}_{\pi_A \times \pi_A}[V(\theta, m_P, m_A)]).
\]
\(\Psi\) 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 \(\Pi_{\theta} : \Delta(M_A) \Rightarrow \Delta(M_P)\) be the correspondence given by
\[
\Pi_{\theta}(\pi_A) = \Delta(\arg \max_{m_P \in M_P} \mathbb{E}_{\pi_A}[U(\theta, m_P, m_A)]).
\]
\(\Pi_{\theta}\) maps agent behavior strategies into the corresponding optimal behavior strategies for the type \(\theta\) principal in the subgame in which \((\mu, M_P, M_A)\) has been accepted.

Let \(\Pi_A : \Delta(\Theta) \times \Delta(M_P) \times \Delta(M_A) \Rightarrow \Delta(M_A)\) be the correspondence given by
\[
\Pi_A(\widetilde{\lambda}, \pi_{\theta_1}, \ldots, \pi_{\theta_N}) = \Delta(\arg \max_{m_A \in M_A} \mathbb{E}_{\widetilde{\lambda}}[\mathbb{E}_{\pi_A}[V(\theta, m_P, m_A)]]).
\]
\(\Pi_A\) 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 \((\mu, M_P, M_A)\) has been accepted.

For every \(j \in \mathbb{N}\), let \(\Phi_j : \Delta(\Theta) \times \Delta(M_P) \times \Delta(M_A) \Rightarrow \Delta(\Theta) \times \Delta(M_P) \times \Delta(M_A)\)
be the correspondence given by

\[ \Phi_j(\lambda, \pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A) = \{(\lambda', \pi'_{\theta_1}, ..., \pi'_{\theta_N}, \pi'_A) \in \Delta(\Theta) \times \Delta(M_P)^\Theta \times \Delta(M_A) : \]

1. \( \exists \lambda'' \in \Psi(\pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A) \) s.t. \( \lambda'(\theta) = \frac{1}{j+1} + \frac{j}{j+1} \lambda''(\theta) \) \( \forall \theta \in \Theta, \)

2. \( \pi'_{\theta} \in \Pi_\theta(\pi_A) \) \( \forall \theta \in \Theta, \)

3. \( \exists \pi'_A \in \Pi_A(\lambda, \pi_{\theta_1}, ..., \pi_{\theta_N}) \}\].

By construction, \( \Phi_j \) 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 \((\lambda^*_j, \pi^*_1, ..., \pi^*_N, \pi^*_A)\) satisfies \((\lambda^*_j, \pi^*_1, ..., \pi^*_N, \pi^*_A) \in \Phi_j(\lambda^*_j, \pi^*_1, ..., \pi^*_N, \pi^*_A).\) Since \( \Delta(\Theta) \times \Delta(M_P)^\Theta \times \Delta(M_A) \) is sequentially compact, there is a limit point \((\lambda^*, \pi^*_1, ..., \pi^*_N, \pi^*_A)\) of the sequence \{\((\lambda_j, \pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A)\)\}_{j \in \mathbb{N}}. Suppose (by restricting attention to a convergent subsequence if necessary) that \( \lim_{j \to \infty} (\lambda_j, \pi_{\theta_1}, ..., \pi_{\theta_N}, \pi_A) = (\lambda^*, \pi^*_1, ..., \pi^*_N, \pi^*_A) .\)

Standard arguments show that \((\pi^*_1, ..., \pi^*_N, \pi^*_A)\) is a sequential continuation equilibrium when mechanism \((\mu, M_P, M_A)\) is accepted given belief \( \lambda^* .\)

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 \((\mu, M_P, M_A)\). Suppose first that \( \mathbb{E}_{\pi_A^*}[V(\theta, m_P, m_A)] \leq 0 \) for some \( \theta \). Then it must be that \( \lambda^* \) 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 \( \mathbb{E}_{\lambda^*}[\mathbb{E}_{\pi_A}[V(\theta, m_P, m_A)]] \leq 0 \), so it is a sequential continuation equilibrium outcome for the agent to reject \((\mu, M_P, M_A)\) when offered. Such an outcome deters every principal type from proposing \((\mu, M_P, M_A)\). Now suppose that \( \mathbb{E}_{\pi_A^*}[V(\theta, m_P, m_A)] > 0 \) for all \( \theta \in \Theta \). Since \((\pi^*_1, ..., \pi^*_N, \pi^*_A)\) is a sequential continuation equilibrium when mechanism \((\mu, M_P, M_A)\) is accepted, incentive compatibility of the principal types implies that \( \mathbb{E}_{\pi_A^*}[U(\theta, m_P, m_A)] \geq \mathbb{E}_{\pi_A^*}[U(\theta, m_P, m_A)] \) 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 $(\mu, M_P, M_A)$ 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 $\theta$, and either chooses $x'_o$ or proposes a mechanism $(\mu, MP)$ 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 $\theta$. Consider the direct mechanism $(\mu^*, \Theta)$ where $\mu^*$ 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 $(\mu^*, \Theta)$. We conclude that $(\mu^*, \Theta)$ 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 $\tilde{\Theta} \subseteq \Theta$, let $B(\tilde{\Theta})$ be the set of agent action rules taking principal actions into agent responses that are best responses to some fixed belief supported on $\tilde{\Theta}$:

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

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 $\tilde{\Theta} \subseteq \Theta$ such that

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

The first condition says that the agent strictly prefers to accept a contract proposal in which the principal commits to $\chi$ for any belief about the principal’s type supported on $\tilde{\Theta}$.\(^1\) The second condition says that there is some principal type in $\tilde{\Theta}$ that would obtain a strictly higher payoff than they do from $p$ by proposing $\chi$, as long as the agent believes the principal’s type belongs to $\tilde{\Theta}$, while the third condition says that every type outside of $\tilde{\Theta}$ would do strictly worse by proposing $\chi$ 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 $\chi$, i.e. $\text{marg}_X \mu_\chi = \chi$. 

52
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 \( \tilde{\Theta} \subseteq \Theta \) and distribution over principal actions \( \chi \in \Delta(X) \), let \( C(\tilde{\Theta}, \chi) \) be the set of agent responses to a proposal of \( \chi \) given by

\[
C(\tilde{\Theta}, \chi) = \{ (\alpha, \beta) \in [0, 1] \times B(\Theta) : \exists \tilde{\lambda} \in \Delta(\tilde{\Theta}) \text{ s.t.} \}
\]

1. \( \beta(x) \in \Delta(\arg\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]) \forall x \in X, \)

2. \( \alpha = 0 \) if \( \mathbb{E}_{\chi}[\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]] < 0 \), and

3. \( \alpha = 1 \) if \( \mathbb{E}_{\chi}[\max_{y \in Y} \mathbb{E}_{\tilde{\lambda}}[V(\theta, x, y)]] > 0 \).

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 \( \beta \) is optimal for some fixed belief \( \tilde{\lambda} \) with support on \( \tilde{\Theta} \), 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 \( \Gamma(\tilde{\Theta}, \chi) \equiv \Delta(C(\tilde{\Theta}, \chi)) \) 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

\[
\overline{D}^0(\chi, p) = \{ \gamma \in \Delta(\arg\max_{y \in Y} \mathbb{E}_{\gamma}[U(\theta, x, y)]) \}
\]

\[
\overline{D}^k(\chi, p) = \{ \gamma \in \Delta(\arg\max_{y \in Y} \mathbb{E}_{\gamma}[U(\theta, x, y)]) \}
\]

\[
\Theta^\dagger(\chi, p) = \{ \theta \in \Theta : \overline{D}^k(\chi, p) \subseteq \bigcup_{\theta \neq \theta} \overline{D}^k(\chi, p) \}
\]

\[
\overline{\Theta}^k(\chi, p) = \begin{cases} 
\Theta^\dagger(\chi, p) & \text{if } \Theta^\dagger(\chi, p) \neq \emptyset \\
\overline{\Theta}^{k-1}(\chi, p) & \text{if } \Theta^\dagger(\chi, p) = \emptyset
\end{cases}
\]

and then set

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

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

\[
\Theta_{SJ}^+(\chi, p) = \begin{cases} 
\Theta_{SJ}^+(\chi, p) & \text{if } \Theta_{SJ}^+(\chi, p) \neq \emptyset \\
\Theta^\infty(\chi, p) & \text{if } \Theta_{SJ}^+(\chi, p) = \emptyset 
\end{cases}
\]

**Definition OA 3.** The set of strongly justified types for \( \chi \) 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 \( E_\gamma[\alpha E_\chi[E_\beta(x)[U(\theta, x, y)]]] \leq U(\theta, p) \) 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 \( \tilde{\Theta} \subseteq \Theta \) such
that

1. \( \min_{\tilde{\lambda} \in \Delta(\tilde{\Theta})} \max_{y \in Y} E_{\tilde{\lambda}}[V(\theta, x, y)] > 0, \)
2. \( \min_{y \in BR(\tilde{\Theta}, x)} U(\theta, x, y) > U(\theta, p) \) for all \( \theta \in \tilde{\Theta}, \) and
3. \( \max_{y \in BR(\tilde{\Theta}, x)} U(\theta', x, y) < U(\theta', p) \) for all \( \theta' \not\in \tilde{\Theta}. \)

**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 \( \tilde{\Theta} \subseteq \Theta \) and \( x \in X, \) let \( C(\tilde{\Theta}, x) \) be the set of agent responses to a proposal of \( x \) given by

\[
C(\tilde{\Theta}, x) = \{ (\alpha, \beta) \in [0, 1] \times \Delta(Y) : \exists \tilde{\lambda} \in \Delta(\tilde{\Theta}) \text{ s.t.} \}
\]

\[
(1) \beta \in \Delta(\arg \max_{y \in Y} E_{\tilde{\lambda}}[V(\theta, x, y)]),
\]

\[
(2) \alpha = 0 \text{ if } \max_{y \in Y} E_{\tilde{\lambda}}[V(\theta, x, y)] < 0, \text{ and}
\]

\[
(3) \alpha = 1 \text{ if } \max_{y \in Y} E_{\tilde{\lambda}}[V(\theta, x, y)] > 0. \}
\]

We let \( \Gamma(\tilde{\Theta}, x) \equiv \Delta(C(\tilde{\Theta}, x)) \) 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

\[
\tilde{D}_{\theta}^k(x, p) = \{ \gamma \in \Gamma(\overline{\Theta}^{k-1}(x, p), x) : E_{\gamma}[\alpha E_{\beta}[U(\theta, x, y)]] > U(\theta, p) \},
\]

\[
\tilde{D}_{\theta}^{0,k}(x, p) = \{ \gamma \in \Gamma(\overline{\Theta}^{k-1}(x, p), x) : E_{\gamma}[\alpha E_{\beta}[U(\theta, x, y)]] = U(\theta, p) \},
\]

\[
\Theta^{\dagger,k}(x, p) = \{ \theta \in \Theta : \tilde{D}_{\theta}^k(x, p) \cup \tilde{D}_{\theta}^{0,k}(x, p) \not\subseteq \cup_{y \neq \theta} \tilde{D}_{\theta'}(x, p) \},
\]

\[
\overline{\Theta}^k(x, p) = \begin{cases} 
\Theta^{\dagger,k}(x, p) & \text{if } \Theta^{\dagger,k}(x, p) \neq \emptyset \text{, and then set} \\
\overline{\Theta}^{k-1}(x, p) & \text{if } \Theta^{\dagger,k}(x, p) = \emptyset
\end{cases}
\]

\[
\overline{\Theta}^{\infty}(x, p) = \bigcap_{k \in \mathbb{N}} \overline{\Theta}^k(x, p).
\]

Let \( \Theta^{S,J^{\dagger}}(x, p) = \{ \theta \in \overline{\Theta}^{\infty}(x, p) : \exists (1, \beta) \in C(\overline{\Theta}^{\infty}(x, p), x) \text{ s.t. } E_{\beta}[U(\theta, x, y)] \geq U(\theta, p) \} \) 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

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

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 $E_\gamma[\alpha E_\beta[U(\theta,x,y)]] \leq U(\theta,p)$ 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 $\theta_N$ optimization problem in (1). For every $\varepsilon > 0$, let $\chi_\varepsilon \in \Delta(X \times T)$ be the distribution obtained from $\chi$ by shifting every $t$ to $t + \varepsilon$. Then the constraints
in (1) are strictly satisfied by \( q_\varepsilon \). Robust neologism proofness demands that the type \( \theta_N \) 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 \( \theta_n \) optimization problem in (1), and let \( q \in \Delta(X \times T \times Y) \) be the distribution obtained from \( \chi \) 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 \( \chi' \) by shifting every \( t \) to \( t - \varepsilon \). Every type below \( \theta_n \) gets a strictly lower payoff from \( q \) than \( p \). Moreover, since payoff-plausibility holds for all \( n'' > n \), all types above \( \theta_n \) must get a weakly lower payoff from \( q \) than \( p \). If additionally every type above \( \theta_n \) were to get a strictly lower payoff from \( q \) than \( p \), then there would be a credible robust neologism corresponding to \( \chi'_\varepsilon \) and \( \theta_n \) for some sufficiently small \( \varepsilon > 0 \), a contradiction. Suppose instead that there are types above \( \theta_n \) 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 \( \chi'_\varepsilon \) 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 \( \chi' \) 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 \( \chi' \) 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 \( \chi' \) 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 \( \chi'' \) 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. \( \blacksquare \)

**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 \( \theta_N \) optimization problem in (1), and let \( \chi_\varepsilon \in \Delta(X \times T) \) be the distribution obtained from taking \( \chi \) 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 \( \theta_N \) is the unique strongly justified type for \( \chi_\varepsilon \) for all sufficiently small \( \varepsilon > 0 \). Consequently, SJCE demands that the type \( \theta_N \) 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 \( \theta_n \) optimization problem in (1), and let \( \chi_\varepsilon \in \Delta(X \times T) \) be the distribution obtained from taking \( \chi \) 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 \( \chi_\varepsilon \) are above \( \theta_n \), a similar argument to the case for \( \theta_N \) 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 \( \chi_\varepsilon \) below \( \theta_n \). There must be some \( n'' > n \) such that type \( \theta_{n''} \) is also strongly justified. Without loss of generality, assume that \( n'' > n \) 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.

\[ \blacksquare \]

**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 \( \chi \) and some non-empty \( \widetilde{\Theta} \). Let \( \widetilde{\theta} = \min(\widetilde{\Theta}) \), and
consider the conditional distribution \( q^*(\theta) \in \Delta(X \times T \times Y) \) where \( \text{marg}_{X \times T} q^*(\theta) = \chi \) and \( y = y^*(\theta, x) \) for all \( x \neq x_0 \). By the definition of a credible robust neologism, \( \mathbb{E}_{q^*(\theta)}[v(\theta, x, y) + g(t)] > 0 \) and \( \mathbb{E}_{q^*(\theta)}[u(\theta, x, y) - t] < U^*(\theta) \) for all \( \theta \notin \tilde{\Theta} \), so \( q^*(\theta) \) satisfies the constraints in (1) for type \( \theta \). Consequently, \( \theta \)'s payoff must be weakly greater that from \( q^*(\theta) \), but this contradicts there being a credible robust neologism corresponding to \( \chi \) and \( \tilde{\Theta} \).

**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 \( \gamma \in \Gamma(\Theta^k(\chi, p^*), \chi) \) that deters all principal types from proposing \( \chi \).

We begin with the base case \( k = 0 \). If every agent best response to \( \chi \) makes every principal type no better off than in \( p^* \), then we are done. Suppose instead that there is some agent best response to \( \chi \) that makes some principal type strictly better off than in \( p^* \). To obtain some \( \gamma \in \Gamma(\Theta^0(\chi, p^*), \chi) \) that deters the principal types, it will be sufficient to consider the family of agent posterior beliefs \( \Lambda_0 = \{ \tilde{\lambda} \in \Delta(\Theta) : \exists n \in \{1, \ldots, N\} \text{ s.t. } \tilde{\lambda}(\theta_m) = 0 \text{ if } m < n \text{ or } m > n + 1 \} \) that put positive probability on at most two principal types, which must be adjacent. Note that FOSD gives a complete ordering over \( \Lambda_0 \). Since the mapping from agent beliefs to agent best responses is upper hemicontinuous, there is some smallest (according to FOSD) \( \lambda \in \Lambda_0 \) for which there is an agent best response that makes some principal type in \( \Theta^0(\chi, p^*) \) weakly better off than in \( p^* \). If the agent strictly prefers to either accept or reject \( \chi \) under belief \( \lambda \), the associated best response is pinned down. If instead the agent is precisely indifferent between accepting or rejecting \( \chi \), fix the agent best response to \( \lambda \) that accepts the proposal with the smallest probability among the best responses for which some principal type in \( \Theta^0(\chi, p^*) \) weakly prefers \( \chi \) to \( p^* \). Let \( q \in \Delta(X \times T \times Y) \) be the distribution obtained from \( \chi \) under this agent best response, and let \( \theta_2 \) be the smallest type in \( \Theta^0(\chi, p^*) \) which weakly prefers \( q \) to \( p^* \). We handle three cases: (1) \( \lambda(\theta_2) = 1 \), (2) \( \lambda(\theta) > 0 \) for some \( \theta > \theta_2 \), and (3) \( \lambda(\theta) > 0 \) for some \( \theta < \theta_2 \). In Cases (1) and (2),
there is an agent best response to a belief fully supported on $\theta_n \in \Theta^0(\chi, p^*)$ that deters all principal types. We now establish that this is also true for Case (3). If there is an agent best response to $\chi$ and a belief fully supported on $\theta_n$ that rejects $\chi$, then we are done. Otherwise, let $q' \in \Delta(X \times T \times Y)$ be the distribution obtained from $\chi$ and the agent best response to a belief fully supported on $\theta_n$. If $\theta_n$ 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 $\chi$ and shifting every $t$ to $t + \varepsilon$, satisfies the constraints in (1) and gives type $\theta_n$ a strictly higher payoff than $p^*$, which violates payoff-plausibility. Since $\theta_n$ gets a weakly lower payoff from $q'$ than $p^*$, this must hold for all lower types as well. Suppose that some higher type $\theta''$ would get a strictly higher payoff from $q'$ than $p^*$, and suppose without loss of generality that $\theta'$ is the lowest such type. Then the $\chi''$ which results from taking $\chi$ and shifting every $t$ to $t + \varepsilon$ satisfies the constraints in (1) and gives type $\theta''$ 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 $\Theta^K(\chi, p^*) \subseteq \Theta^{K-1}(\chi, p^*)$, if every $\gamma \in \Gamma(\Theta^{K-1}(\chi, p^*), \chi)$ makes every principal type no better than in $p^*$, then we are done. Suppose instead that there is some $\gamma \in \Gamma(\Theta^{K-1}(\chi, p^*), \chi)$ that makes some principal type strictly better off than in $p^*$. Consider the family of agent posterior beliefs $\Lambda_K$ that are supported on $\Theta^{K-1}(\chi, p^*)$ 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) $\lambda \in \Lambda_K$ for which there is an agent best response that makes some principal type in $\Theta^K(\chi, p^*)$ weakly better off than in $p^*$. As before, if the agent is precisely indifferent between accepting or rejecting $\chi$ under belief $\lambda$, fix the agent best response to $\lambda$ that accepts the proposal with the smallest probability among the best responses for which some principal type in $\Theta^K(\chi, p^*)$ weakly prefers $\chi$ to $p^*$. Let $q \in \Delta(X \times T \times Y)$ be the distribution obtained from $\chi$ under this agent best response, and let $\theta_n$ be the smallest type in $\Theta^K(\chi, p^*)$ 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 \( \theta_n \) which deters all principal types from proposing \( \chi \).

Since there is some \( K \in \mathbb{N} \) such that \( \bar{\Theta}^k(\chi, p^*) = \bar{\Theta}^\infty(\chi, p^*) \) for all \( k > K \), it follows that there is a best response \( \gamma \in \Gamma(\bar{\Theta}^\infty(\chi, p^*), \chi) \) that deters all principal types from proposing \( \chi \). 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 \( \chi \), which means that \( p^* \) is an SJCE outcome. 

\[\blacksquare\]

References