Contracting with Private Information, Moral Hazard, and Limited Commitment[*]

Daniel Clark[†]

August 4, 2022

Abstract

We study principal-agent settings with moral hazard where the principal has private information. In contrast to past work, we assume that the principal can propose contracts that give them flexibility in their choice of future actions, but cannot commit to arbitrary stochastic randomizations over these actions. We focus on environments where the principal's type and agent's action are complements. The *principal-optimal safe outcomes*, which are analogs of the least-cost separating outcomes of signaling games, are key to the analysis: They are always equilibrium outcomes, and they give lower bounds on the payoffs to the principal types that must be met in every equilibrium satisfying the refinement of *payoff-plausibility.* Moreover, if there are complementarities between the principal's type and their action, payoff-plausibility selects the principal-optimal safe outcomes. Otherwise, pooling between principal types can survive payoff-plausibility, and is more prevalent than would be predicted with the frequently assumed explicit contracts.

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

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

[†] Department of Economics, Yale University. Email: daniel.clark@yale.edu.

1 Introduction

Many important economic interactions are principal-agent problems where the principal possesses private information. Despite this, there has been relatively little work studying the informed principal problem, and what work has been done has made one of two extreme assumptions. One is that the principal can propose contracts that precisely pin down the actions they will take should a relationship with the agent form; this enables the interaction to be treated as a standard signaling game. The other is that the principal can propose an arbitrary mechanism with unlimited commitment power, which enables the use of familiar mechanism design techniques.

This paper instead will study the informed principal problem under the more realistic assumption that the principal can propose a contract under which they retain some degree of flexibility in their choice of future actions, but does not require unlimited commitment power. In particular, principals can propose contracts that are menus over the future actions that they can take, and, should a contract be accepted, the principal will be required to choose an action from the corresponding menu.[1] However, the principal cannot commit to an arbitrary randomization over these actions, as would be the case with the usual mechanism design approach.[2]

To fix ideas, consider an informed-principal version of a canonical firm and worker problem, where the firm is more informed about how the employee's effort will translate into profit. Here, when the compensation specified by the contract depends on the firm's profit, the potential employee's perception of what the firm knows will be important for their decisions of whether to accept the employment offer and, if they do, how hard to work. Our approach allows the firm to potentially propose a contract in which they restrict the compensation schemes they will eventually use, but does not

[1] As noted by Segal and Whinston [2003], publishers often use contracts with multiple options concerning publication and copyrights of books. Similarly, a firm may offer a contract to a prospective employee that places some constraints on the possible tasks the firm could assign or the exact nature of how the firm will compensate the employee, but does not completely narrow down the firm's possible actions.

[2] It is still possible that non-degenerate distributions over principal actions prevail, but these rely on incentive compatibility rather than exogenous commitment power.

restrict them to a specific compensation scheme at the time of contracting. (Of course, the firm can choose to propose a perfectly explicit contract that does restrict them to a single scheme.) However, the firm cannot commit to a contract in which one of the firm's options is a non-degenerate random distribution over payment schemes. The paper will later formalize this example and use it to illustrate the main findings.

We focus on a natural class of environments with complementarity between the principal's type and the agent's action. This complementarity holds in many settings of interest, and it leads to a tendency for higher principal types to separate from lower principal types. Indeed, the principal-optimal safe outcomes, which are the analog of the least-cost separating outcomes from signaling games, are always perfect Bayesian equilibrium (PBE, Fudenberg and Tirole [1991]) outcomes. They also always satisfy the refinement of payoff-plausibility (Clark [2022]), a refinement similar to but stronger than the Intuitive Criterion (Cho and Kreps [1987]), and they give a lower bound on the payoffs of the principal types in every payoff-plausible PBE outcome.

We compare our findings to those that emerge under the alternate assumptions that either the principal can only propose explicit contracts that precisely pin down their future actions or that the principal can propose contracts committing to arbitrary mechanisms. Under either of these alternate assumptions, the principal-optimal safe outcomes are payoff-plausible PBE outcomes and give payoff lower bounds for all other payoff-plausible PBE outcomes. (However, the principal-optimal safe outcomes when arbitrary mechanisms can be proposed are different than those with limited commitment.) Unlike the case where only explicit contracts can be proposed, with flexible contracts and limited commitment, payoff-plausibility does not generally select the principal-optimal safe outcomes. This avoids a long-standing concern about the tendency of refinements to select only the least-cost separating equilibria in signaling games.

While the general-mechanism approach of allowing the principal to implement arbitrary mechanisms is consistent with the standard mechanism design literature, and affords useful analytical tools, such as the Inscrutability Principle (Myerson [1983]), the approach assuming limited commitment often leads to narrower predictions and more striking results. We illustrate this in a special class of environments in which there are additional complementarities between the principal's type and their action. Here payoff-plausibility with limited commitment selects precisely the principal-optimal safe outcomes, whereas many other outcomes can survive payoff-plausibility when arbitrary mechanisms can be proposed.

2 Related Literature

Beaudry [1994], Inderst [2001], Chade and Silvers [2002], Bénabou and Tirole [2003], Martimort and Sand-Zantman [2006], and Sun [2021] (in a dynamic setting) studied informed principals with explicit contracts that commit the principal to a single action. Beaudry [1994] and Inderst [2001] in particular studied settings like the example presented in Section 3.

The study of informed principals with unrestricted contracts began with Myerson [1983], which analyzed a general setting in which the principal and agents can all posses asymmetric information and the agents' actions may be subject to moral hazard. The subsequent literature studying the design of general mechanisms by informed principals has largely focused on settings without moral hazard; it includes Maskin and Tirole [1990, 1992], Inderst [2005], Cella [2008], Severinov [2008], Mylovanov and Tröger [2012, 2014], Balkenborg and Makris [2015], Koessler and Skreta [2016], Bedard [2017b], DeMarzo and Frankel [2020], DeMarzo et al. [2020], and Dosis [2022].

Similar to this paper, Clark [2022] focuses on informed principal settings with agent moral hazard. However, it takes a mechanism design perspective and allows for unrestricted contracts. It also develops payoff-plausibility, and shows that it is a consequence of two signaling game refinements, *robust neologism proofness (RNP)* (Clark [2021]) and *strongly justified communication equilibrium (SJCE)* (Clark and Fudenberg [2021]), when they are extended to certain informed principal environments. Other papers studying informed principals with agent moral hazard are Wagner et al. [2015], Bedard [2017a], and Mekonnen [2021], which limited attention to very special environments.[3] Wagner et al. [2015] and Mekonnen [2021] allow for unrestricted mechanisms, while Bedard [2017a] implicitly focused on mechanisms that rule out stochastic randomizations, though this is not its focus and it does not perform equilibrium analysis.

3 Firm and Employee Example

3.1 Setup

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

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

The standard approach, seen for instance in Beaudry [1994] and Inderst [2001], requires that the firm's contract commit to a single action, which in this case is a payment scheme, so that the contracts correspond to (s, t) pairs. With these *explicit* contracts, the agent knows precisely what share of profits and transfer the firm will

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

implement should the agent accept. This does not allow *flexible* contracts, which are both plausible and observed in the real world. In the present example, the firm might want to retain some flexibility, e.g. about how much of the employee's compensation will be governed by profit sharing or transfers, rather than completely pinning down their future actions.[4]

The flexible contracts we study correspond to menus of (s, t) pairs. The interpretation is that, should the employee accept the contract, the firm will be bound to choose one of the payment schemes allowed by the contract. While significant, the allowed flexibility is not unlimited. In particular, the firm is not able to commit to an arbitrary stochastic randomization over (s, t) pairs.

3.2 Equilibria with Explicit Contracts

We first consider equilibria when the firm can only propose explicit contracts. Essentially, this amounts to a standard signaling game with a slightly more convoluted timeline. First, the firm observes θ and then proposes a contract corresponding to a (s, t) pair. The employee observes the chosen (s, t) and either accepts or rejects the offer. If the employee rejects, both parties get a payoff of 0. If instead the employee accepts, the employee will then exert some effort e, after which profits and payoffs are realized.

Before analyzing the equilibria of our contracting game, we discuss a benchmark solution for contracting with symmetric information. The *complete-information benchmark* is the outcome that would occur if the firm's type were commonly known to be θ. Here the standard solution is that the employee receives all of the profits ($s = 1$), the employee exerts first-best effort level ($e = \theta$), and the firm extracts all of the surplus ($t = -\theta^2/2$). This results in payoffs of 2 to the type 2 firm, 8 to the type 4 firm, and 0 to the employee regardless of the firm's type.

This outcome is not possible with asymmetric information, because the type 2 firm

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

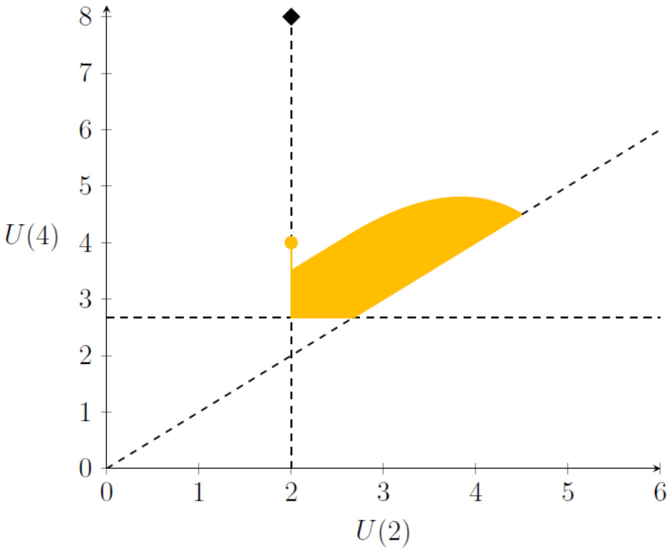

Figure 1: The yellow region depicts the possible equilibrium payoff pairs. The diamond at $(2, 8)$ denotes the payoffs of the firm types in the complete-information benchmark.

would strictly prefer to mimic the type 4 firm, which would let them extract a higher fee from the employee. Under perfect Bayesian equilibrium (Fudenberg and Tirole [1991]), the possible pairs of firm-type equilibrium payoffs, where $U(\theta)$ denotes the equilibrium payoff of type θ, are given in Figure 1.

To understand the possible equilibrium payoff pairs, observe that the type 2 firm can never get a lower payoff than 2, their complete-information benchmark. The reason is the firm can offer a contract corresponding to $(s, t) = (1, 2-\varepsilon)$ for some $\varepsilon > 0$, which amounts to a perturbation of their optimal contract with complete information. Such a proposal is guaranteed to be accepted and result in a payoff of $2-\varepsilon$ to the firm. This holds for all $\varepsilon > 0$, so the firm can always get arbitrarily close to a payoff of 2. Moreover, the lowest payoff that the type 4 firm can be held to is $8/3$, which comes from having the employee believe $\theta = 2$ following any off-path contract proposal. Additionally, the high-type firm can never get a lower equilibrium payoff than the low-type firm.

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

All points to right of $U(2) = 2$ involve pooling. The reason is that the payoff of the type 2 firm in all separating equilibria is 2. Thus, in a pooling equilibrium where $U(2) > 2$, there must be some (s, t) played with positive probability by both firm types where the employee's posterior puts at least probability 1/2 on $\theta = 2$. This fact enables the formulation of a constrained optimization problem that maximizes the payoff of the type 4 firm subject to the type 2 firm's payoff equaling $U(2)$, employee incentive compatibility, and an individual rationality constraint that averages across both $\theta = 2$ and $\theta = 4$. The solution to this problem, the analysis of which is given in Section OA.1.2, characterizes the upper envelope in the $U(2) > 2$ region.

3.3 Equilibria with Flexible Contracts

The timing of the game with flexible contracts is similar to that when only explicit contracts can be proposed, with the following differences: Contracts do not necessarily commit to single (s, t) pairs, and, should the employee accept the firm's contract offer, the firm then chooses some (s, t) permitted by the contract. After this, the employee observes (s, t) and then exerts some effort level e, following which profits and payoffs are realized. Figure 2 depicts the PBE payoffs with flexible contracts as well as those possible when only explicit contracts can be proposed.

Observe that the type 4 firm cannot be held to same minimum payoff with flexible contracts as with explicit contracts. The reason is that the type 4 firm can always get payoffs strictly higher than 8/3 because of the richer space of deviations. In particular, there are contracts in which all the continuation equilibria following their proposal give a higher payoff than 8/3 to the type 4 firm. For example, consider a contract with

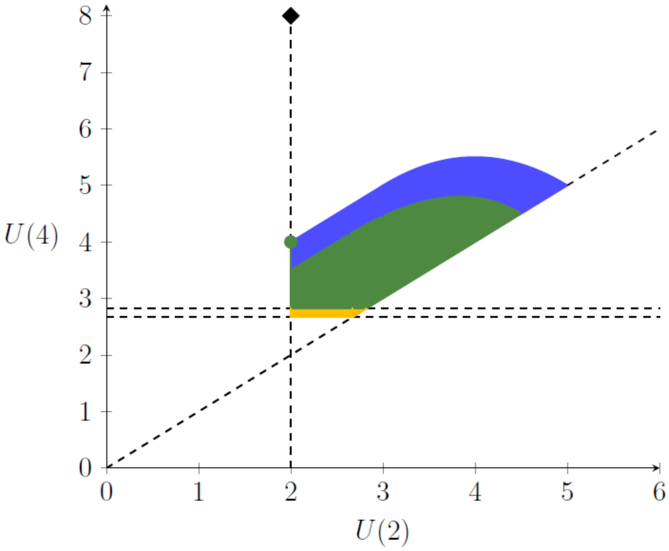

Figure 2: The blue region depicts the equilibrium payoffs that can only be sustained with flexible contracts, while the yellow region consists of equilibrium payoffs that can be sustained only with the restriction to explicit contracts. Equilibrium payoffs that can be sustained with both classes of contracts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

two options, $(s_1, t_1) = (1, -199/100)$ and $(s_2, t_2) = (2/3, -1)$. If the contract were proposed and accepted, then the type 4 firm would always select (s_2, t_2), and obtain a payoff of at least 25/9. The type 2 firm would only select (s_2, t_2) when it induces the employee to exert effort at least $e = 297/200$. Given (s_2, t_2) and any belief that would induce the employee to exert effort higher than $e = 297/200$, the employee's conditional expected utility must be at least $(297/200)^2/2 - 1 > 0$. Moreover, the employee's expected utility conditional on (s_1, t_1) is always strictly positive. Thus, this contract gives the employee a strictly positive expected utility in every sequential continuation equilibrium, so the type 4 firm's payoff from its proposal must at least be 25/9.[5]

Additionally, with flexible contracts, the upper envelope is higher and smooth. It also can be found through a constrained optimization problem, details of which are in Section OA.1.1. However, unlike the case with explicit contracts, all the points on the

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

upper envelope with flexible contracts correspond to outcomes where the agent correctly anticipates the principal's type when they choose their effort. In particular, any payoff on the upper envelope can be realized in an outcome where, conditional on the low type $\theta = 2$, the employee receives the full profit share $s = 1$ and exerts efficient effort level $e = 2$, and conditional on the high type $\theta = 4$, the employee exerts optimal effort $e = 4s$ for the corresponding profit share s. Intuitively, if this were violated, the payoffs of both the firm and the worker when $\theta = 2$ could be weakly increased by increasing the surplus to its maximum value of 2 and appropriately dividing it. Moreover, the payoff of the high type $\theta = 4$ could only improve from not being mistaken for the low type. The reason why these outcomes are possible with flexible contracts is that they can be achieved with both firm types proposing the same contract. This leads the employee to be willing to accept a relationship with a type 2 firm despite regretting it later.

3.4 Plausible Equilibria

There are many equilibria both when flexible contracts can be proposed and when only explicit contracts can be proposed, but not all the equilibria are reasonable. Consider for instance equilibria with flexible contracts in which both firm types obtain a payoff of 5. (Graphically, these equilibria correspond to the star in the right-hand plot of Figure 3.) We should expect the high-type firm to obtain a strictly higher payoff than the low-type firm, because the high type should be able to credibly signal their identity to the employee when the prevailing equilibrium has both types receiving the same payoff. For example, suppose the type 4 firm proposed a contract committing to $(s, t) = (1/2, -1.5)$. Every undominated response of the employee to such a contract would involve effort levels less than 2 and thus give the type 2 firm a strictly lower payoff than 5; however, the employee accepting and exerting effort 2, as they would if they knew $\theta = 4$, would give the type 4 firm a strictly higher payoff of 5.5. Because of this, *payoff-plausibility*, which is formally defined in Section 4.4.2, rules out the equilibria

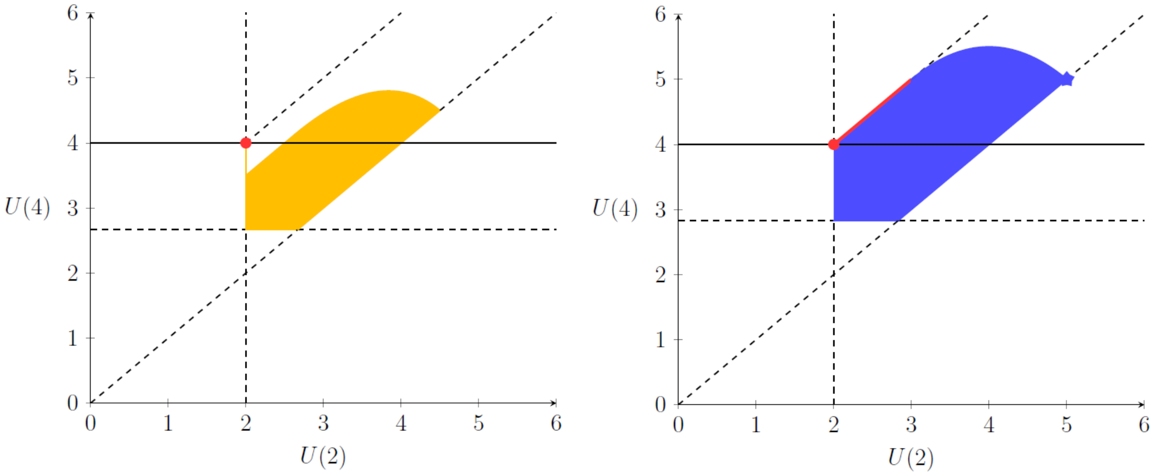

Figure 3: The left-hand figure depicts equilibrium payoffs with explicit contracts, with plausible payoffs in red and all other payoffs in yellow. The right-hand figure depicts equilibrium payoffs with flexible contracts, with plausible payoffs in red and all other payoffs in blue.

in which both firm types obtain a payoff of 5. More generally, payoff-plausibility eliminates equilibria when there is some type θ and a contract that, when the agent responds as if the type were θ, would give the type θ principal a strictly higher payoff than the equilibrium and all types below θ a strictly lower payoff.[6]

Payoff-plausibility selects precisely the red payoff pairs depicted in Figure 3. These are the payoffs that correspond to outcomes that can be obtained from the principal-optimal safe outcome by uniformly reducing the transfers paid by the firm types. With flexible contracts, there is a non-singleton line segment of such payoffs, as shown in Section OA.1.3, while there is only one such payoff with explicit contracts. As we will see in Section 6, in a broad class of environments nesting this example, payoff-plausibility selects the principal-optimal safe outcomes when only explicit contracts can be proposed, but frequently allows multiple equilibrium outcomes with flexible contracts.

Intuitively, payoff-plausibility eliminates any equilibrium whose payoffs are beneath the upper envelope with flexible contracts because, in any such equilibrium, the type

[6] With two firm types, the Intuitive Criterion (Cho and Kreps [1987]) is equivalent to payoff-plausibility. With more types, the Intuitive Criterion is usually much weaker, as illustrated by example in OA.4.1.

4 firm could propose a contract corresponding to a point on the upper envelope that is above and to the left of the equilibrium payoffs. The type 2 firm would do worse by such a proposal, while the type 4 firm would do better if the employee were to respond under the belief that $\theta = 4$. The requirement that plausible payoffs lie on the upper envelopes holds generally in a broad class of environments with two types. It is not clear that this always extends with more than two types. However, there are general thresholds that the payoffs in payoff-plausible equilibria must always meet. In particular, every principal type must always obtain a weakly higher payoff than they do in the principal-optimal safe outcomes. In this example, this amounts to the requirement that the type 4 firm always obtain a weakly higher payoff than 4, which is the reason for the horizontal lines at $U(4) = 4$ in Figure 3.

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

4 Framework

4.1 Primitives

The set of possible principal types is $\Theta = \{\theta_1, ..., \theta_N\}$, where the types are ordered so that $\theta_1 < ... < \theta_N$. The ex-ante probability of type θ is $\lambda(\theta) > 0$. If a relationship is formed, the principal's action set is the compact metric space X, with $x \in X$ denoting a typical principal action, while the agent's action set is the compact interval $Y = [\underline{y}, \overline{y}] \subset \mathbb{R}$, with $y \in Y$ denoting a typical agent action. Here, a principal action x could represent an investment, task assignment, incentive scheme, or monitoring system, and an agent action y could represent effort level, type of work, or social behavior. In addition to choosing an x, the principal gives a transfer $t \in \mathbb{R}$ to the agent. If a relationship is formed, $u(\theta, x, y) - t$ and $v(\theta, x, y) + t$ are the utilities of the principal and agent, respectively, when the principal's type is θ, the principal takes action x, gives transfer t, and the agent takes action y.[7] Both $u : \Theta \times X \times Y \to \mathbb{R}$ and $v : \Theta \times X \times Y \to \mathbb{R}$ are continuous.

If instead the principal and agent do not form a relationship, then both realize their outside options; the payoffs to all types of the principal and the agent from their outside options are normalized to 0.

Moreover, as in the informed firm and employee example, we assume that $v(\theta, x, y)$ is strictly concave in y for all x, so that $y^*(\tilde{\lambda}, x) \equiv \arg\max_{y \in Y} \sum_{\theta \in \Theta} \tilde{\lambda}(\theta) v(\theta, x, y)$ is singleton for all $\tilde{\lambda} \in \Delta(\Theta)$ and x. Additionally, we impose the following monotonicity assumptions

1. $u(\theta, x, y)$ is weakly increasing in y for all $\theta \in \Theta$ and $x \in X$.
2. $u(\theta, x, y)$ and $v(\theta, x, y)$ are weakly increasing in θ for all $x \in X$ and $y \in Y$.

and complementarity assumptions

3. $y^*(\theta, x)$ is weakly increasing in $\theta \in \Theta$ for all $x \in X$.
4. For all $\theta, \theta' \in \Theta$, $x \in X$, and $y, y' \in Y$ such that $\theta \geq \theta'$ and $y \geq y'$, $u(\theta, x, y) - u(\theta, x, y') \geq u(\theta', x, y) - u(\theta', x, y')$.

The monotonicity criteria state that (1) the principal always (weakly) prefers a higher agent action and (2) a higher principal type is good news in that both the agent and the principal gain (weakly) more by forming a relationship when the principal's type is higher. The first complementarity condition says that the agent's best response is weakly increasing in the principal's type. The second complementarity assumption requires that the principal's utility have increasing differences in their type and the agent's action.

[7] The assumption that the agent's utility is quasilinear in their transfer is made for simplicity. All results would hold if instead the agent's utility were of the form $v(\theta, x, y) + g(t)$ for some weakly concave function $g : \mathbb{R} \to \mathbb{R}$.

4.2 Contracts and the Principal-Agent Game

At the beginning of their interaction, the principal offers the agent a contract. A contract specifies a non-empty, finite menu of action-transfer pairs $C \subseteq X \times \mathbb{R}$ from which the principal must choose, and we will identify each contract with its corresponding C.[8] We will use $\mathcal{C}$ to denote the set of all possible contracts.

Formally, the ***principal-agent game*** proceeds as follows. The principal observes their type θ, and proposes a contract C to the agent. The agent observes the principal's choice of contract and then decides whether to accept the offer. If the agent rejects the offer, the game ends with the principal and agent each realizing their outside options. If instead the agent accepts the offer, the principal and agent form a relationship. Subsequently the principal chooses an action-transfer pair $(x, t) \in C$. The agent then observes the (x, t) and responds with an action y. After this the payoffs are realized.

4.3 Outcomes

We will focus will on *outcomes*, and in particular the outcomes that can arise under various notions of equilibria. To define outcomes, we introduce the object o, and we use the pair (θ, o) to denote the principal's type being θ and both parties receiving their outside option. Additionally, we will use $(\theta, \alpha, x, t, y) \in \Theta \times (0, 1] \times X \times \mathbb{R} \times Y$ to denote the principal's type being θ, a contract that is accepted with probability α being accepted, and (x, t, y) ultimately occurring.[9] An ***outcome*** is a probability distribution $p \in \Delta(\Theta \times ((0, 1] \times X \times \mathbb{R} \times Y \cup \{o\}))$.

[8] We could allow for contracts that were compact subsets of $X \times \mathbb{R}$ without changing any results.

[9] It is convenient to ignore the actual contract that is proposed since, beyond the (α, x, t, y) it induces, it is payoff-irrelevant.

4.4 Solution Concepts

4.4.1 Perfect Bayesian Equilibrium

Our baseline solution concept is perfect Bayesian equilibrium.[10] A PBE consists of (1) a strategy for each principal type, which amounts to a contract proposal distribution and a rule that takes each contract into a distribution over the (x, t) allowed by that contract, (2) an agent strategy, which amounts to a rule mapping contracts into acceptance probabilities and a rule mapping contracts and principal action-transfer pairs into agent actions, and (3) a belief update rule that gives the agent's interim beliefs upon the proposals of arbitrary contracts as well as after an arbitrary contract has been proposed, accepted, and the agent has observed an arbitrary (x, t) allowed by the contract. For simplicity, we require that each principal type's strategy induces a distribution over contract proposals that has finite support.[11]

Such a collection of strategies for each player and agent belief update rule is a PBE if and only if the following conditions hold. (1) Each principal type plays optimally: Their expected payoff must be no less than the payoff they could get by playing an arbitrary contract and subsequent (x, t) given the play of the agent. (2) The agent plays optimally: For each contract, their acceptance decision and their subsequent choices of actions conditional on the various (x, t) maximize their expected utility given their posterior beliefs about the principal's type. (3) The agent's posterior beliefs at the contract proposal stage must be consistent with their prior, the contract proposal rules of the principal types, and Bayes' rule whenever possible. Likewise, after accepting a given contract, the agent's posterior belief upon observing a given (x, t) must be consistent with their interim belief about the principal's type when the corresponding contract is proposed, the action-transfer pair selection rules of the principal types, and Bayes' rule whenever possible.

[10] In Clark [2022], which studies general mechanism design by informed principals, PBE could not be applied, because the games considered there do not have perfectly observed actions.

[11] All results extend to the case where the principal can use an arbitrary distribution over contract proposals. Details are available from the author upon request.

4.4.2 Payoff-Plausibility

Perfect Bayesian equilibrium is often excessively permissive in the principal-agent game, so in our analysis, we will frequently apply the criterion of *payoff-plausibility.*

Definition 1. *The profile of principal type expected utilities* $(U(\theta_1), ..., U(\theta_N))$ *is* **plausible** *if* $U(\theta) \geq 0$ *for all* $\theta \in \Theta$ *and, for all* $n \in \{1, ..., N\}$,

$$\begin{aligned} U(\theta_n) \geq \max_{(x,t)\in X\times\mathbb{R}} &\ u(\theta_n, x, y^*(\theta_n, x)) - t \\ s.t.\ &\ v(\theta_n, x, y^*(\theta_n, x)) + t \geq 0, \\ &\ u(\theta_{n'}, x, y^*(\theta_n, x)) - t \leq U(\theta_{n'})\ \forall n' < n. \end{aligned} \tag{1}$$

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

Payoff-plausibility requires that each principal type θ get a non-negative payoff that is at least as high as that from proposing any (x, t) that satisfies agent IR and principal IC constraints when the agent responds under the belief that the type is θ. In particular, the agent IR constraint guarantees that the agent obtains a weakly positive expected utility from (x, t) under type θ when they best-respond with $y^*(\theta, x)$. The principal IC constraint says that every principal type smaller than θ must obtain a weakly lower payoff from proposing (x, t) and having the agent respond under the belief that the type is θ than their payoff in the profile.

Clark [2022] discusses the relationship of payoff-plausibility to various adaptations of signaling game refinements to the principal-agent game in an environment where the principal can propose arbitrary mechanisms. In particular, it shows that payoff-plausibility characterizes both the set of robust neologism proof (Clark [2021]) equilibria and the set of strongly justified communication equilibria (Clark and Fudenberg [2021]) in MCS environments, which is essentially the class of environments considered in this paper modulo a few additional technical restrictions. Similar arguments show that this relationship between payoff-plausibility, robust neologism proofness, and strongly justified communication equilibrium hold in the principal-agent game of this paper in all MCS environments, and are available from the author upon request.

5 Equilibrium Outcomes

In this section, we consider the outcomes that can arise in PBE. We obtain properties that all such outcomes obey, and we identify a special class of outcomes that are always consistent with PBE. A ***PBE outcome*** is an outcome $p \in \Delta(\Theta \times ((0,1] \times X \times \mathbb{R} \times Y \cup \{o\}))$ that is induced by the strategies used in a PBE. Because of the restriction to equilibria in which the contract proposal distributions used by the principal types have finite support, all PBE outcomes have finite-support outcomes. For a finite-support outcome p, we let $Z_p \subseteq (0,1] \times X \times \mathbb{R} \times Y$ be the (finite) set of (α, x, t, y) that occur with positive probability under p. Here and throughout the paper, we use $U(\theta, p)$ to denote the expected utility of a type θ principal under outcome p.[12]

A PBE outcome must satisfy various conditions. For example, standard arguments show that it must be incentive compatible and individually rational for the principal.

Definition 2. *Outcome p is satisfies* **principal incentive compatibility and individual rationality** *if*

$$U(\theta, p) \geq \max\left\{ \max_{(\alpha,x,t,y)\in Z_p} \alpha(U(\theta,x,y) - t), 0 \right\} \text{ for all } \theta \in \Theta.$$

Additionally, it must satisfy agent incentive compatibility and individual rationality constraints.

Definition 3. *Outcome p is satisfies* **agent incentive compatibility and individual rationality** *if, for every* $(\alpha, x, t, y) \in Z_p$,

1. $y = y^*(\tilde{\lambda}, x)$ *where* $\tilde{\lambda}(\theta) = p(\theta, \alpha, x, t, y) / \sum_{\theta' \in \Theta} p(\theta', \alpha, x, t, y)$ *for all* $\theta \in \Theta$*,*

[12] Formally, this is given by $U(\theta, p) = \sum_{(\theta,\alpha,x,t,y)\in \text{supp}(p)} \alpha(u(\theta, x, y) - t)$.

2. $\sum_{(\theta',\alpha,x',t',y')\in supp(p)} p(\theta',\alpha,x',t',y')(v(\theta',x',y')+t) \geq 0$,

3. $\alpha = 1$ *if* $\sum_{(\theta',\alpha,x',t',y')\in supp(p)} p(\theta',\alpha,x',t',y')(v(\theta',x',y')+t) > 0$.

5.1 Safe Outcomes

We now focus on a special class of *safe* outcomes that, in addition to the constraints above, are consistent with regret-free play by the agent.

Definition 4. *Outcome p is* **safe** *if it satisfies principal incentive compatibility and individual rationality as well as the following: For all $(\theta,\alpha,x,t,y) \in supp(p)$,*

1. $y = y^*(\theta,x)$,

2. $v(\theta,x,y)+t \geq 0$, *and*

3. $\alpha = 1$ *if* $v(\theta,x,y)+t > 0$.

For every safe outcome, there is a strategy profile that induces the outcome and is such that the agent's play, both in terms of contract acceptance/rejection and choice of action, is always optimal for every principal type, contract they propose with positive probability, and subsequent continuation play. This strategy profile can be obtained by identifying each (θ,α,x,t,y) with a contract with one option, (x,t). Condition 1 guarantees that playing y is optimal against x when the type is θ, and Conditions 2 and 3 guarantee that the agent's acceptance/rejection this contract is optimal when type proposes it. Moreover, we can identify every (θ,o) with some contract that has a transfer so low that the agent would be guaranteed a strictly negative payoff from accepting the contract.

5.2 Principal-Optimal Safe Outcomes

Within the class of safe outcomes, those which are uniformly optimal for the principal types play a key role in our analysis. In this subsection, we define our notion of *principal-optimal* safe outcomes and show that they are always PBE outcomes.

Definition 5. *Safe outcome p is a* **principal-optimal** *safe outcome if it gives every type of the principal a weakly higher payoff than every other safe outcome p': $U(\theta, p) \geq U(\theta, p')$ for all $\theta \in \Theta$ and safe p'.*

Since the principal prefers higher agent actions and the agent's optimal action increases with the principal's type, higher principal types would like to separate from lower principal types. Complementarity between the principal's type and the agent's action allows the higher principal types to credibly do so by paying higher transfers to the agent. The following proposition shows that principal-optimal safe outcomes exist and characterizes the corresponding payoffs to the principal types.

Proposition 1. *Principal-optimal safe outcomes exist, and the payoffs $\{U^*(\theta)\}_{\theta \in \Theta}$ of the principal types from the principal-optimal safe outcomes are characterized iteratively as follows. For $n \in \{1, ..., N\}$, $U^*(\theta_n) = \max\{U^\dagger(\theta_n), 0\}$, where*

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

As with (1) in the definition of payoff-plausibility, the first constraint in (2) is simply the agent's individual rationality condition given (θ_n, x, t) when the agent responds with $y^*(\theta_n, x)$; the second constraint is a principal incentive compatibility condition guaranteeing that lower types than θ_n weakly prefer their principal-optimal safe outcome to $(x, t, y^*(\theta, x))$. The proof of Proposition 1 is in Section OA.2. It shows that the $U^*(\theta)$ is an upper bound on the payoff of the type θ principal in every safe outcome. This is clearly the case for safe outcomes with no acceptance probabilities strictly between 0 and 1 since, for every (θ_n, x, t) that occurs with positive probability, the (x, t) must satisfy the constraints in (2). Moreover, because of the monotonicity and complementarity assumptions on the payoff functions, it can be shown that, for every safe outcome, there is a safe outcome that gives each principal type a weakly higher payoff and never has acceptance probabilities strictly between 0 and 1. The proof then obtains a safe outcome in which each θ gets payoff $U^*(\theta)$. The outcome is obtained by assigning each type θ for whom $U^\dagger(\theta) = 0$ to the outside option o and each type θ for whom $U^\dagger(\theta) > 0$ to some (x, t) that solves their optimization problem in (2). The constraints in (2) are such that all upward principal incentive compatibility constraints hold, and the proof uses the monotonicity and complementarity assumptions on the payoff functions to show that the downward principal incentive compatibiltiy constraints must hold as well.

We now establish that these principal-optimal safe outcomes are always consistent with equilibrium play.

Theorem 1. *Any principal-optimal safe outcome is a PBE outcome.*

The proof is given in Appendix A. The difficult part is showing that, for each contract, there is continuation play consistent with PBE that would deter every principal type from proposing the contract when they obtain their principal-optimal safe payoff. (The rest of the proof identifies a strategy profile that induces a given principal-optimal safe outcome.) This part of the proof constructs a sequence of modified principal-agent games such that (1) in the limit, each principal type gets their principal-optimal safe payoff, and (2) the limit strategy profiles can be used to find a continuation PBE for each contract that gives each principal type weakly less than their equilibrium payoff. In the modified games, for an arbitrary $(x, t) \in X \times \mathbb{R}$, each principal type θ can forego proposing a contract and take an outside option which automatically gives them their payoff from $(x, t, y^*(\theta, x))$ provided that (1) the agent gets a weakly positive payoff from $(x, t, y^*(\theta, x))$, and (2) every lower type $\theta' < \theta$ would obtain a lower payoff from $(x, t, y^*(\theta, x))$ than they do in equilibrium. This guarantees that each principal type obtains at least their principal-optimal safe payoff. To prevent them from obtaining higher payoffs, we impose costs to using this outside option when either of these conditions are violated.

6 Consequences of Payoff-Plausibility

We now apply payoff-plausibility to refine the set of PBE. We show that the principal-optimal safe outcome provides a payoff benchmark that every payoff-plausible PBE must meet. We also discuss how, with flexible contracts, payoff-plausibility can permit outcomes with higher principal payoffs than the principal-optimal safe outcome, while payoff-plausibility typically selects the least-cost separating equilibria when only explicit contracts can be proposed. However, there is a class of environments in which payoff-plausibility selects precisely the principal-optimal safe outcomes, though this selection would not hold if the principal had unlimited commitment power and could propose arbitrary mechanisms.

6.1 The Principal-Optimal Safe Benchmark

Section 5.2 showed that principal-optimal safe outcomes are always PBE outcomes. They are additionally always payoff-plausible, and they provide payoff benchmarks that every payoff-plausible equilibrium must meet.

Theorem 2.

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

Theorem 2 follows from combining the characterizations of the principal-optimal safe payoffs in Proposition 1 and the requirements of payoff-plausibility. In particular, the proof of Theorem 2.1 shows that, for any equilibrium that does not principal-payoff-dominate the principal-optimal safe outcome, there must be a lowest principal type θ whose expected utility violates payoff plausibility. Theorem 2.2 is an immediate consequence of the observation that, in the principal-optimal safe outcomes, each principal type's payoff precisely equals their plausibility threshold.

6.2 Flexible Versus Explicit Contracts

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

For an example where payoff-plausibility allows pooling under flexible contracts, consider again the firm and employee of Section 3, except now suppose that $\Theta = \{1, 2, 4\}$ and $\lambda(1) = \lambda(2) = \lambda(4) = 1/3$. Here there is an additional low type $\theta = 1$, and all three types are equally likely. One payoff-plausible outcome is for the low type and medium type to pool and give all profit residuals to the employee ($s(1) = s(2) = 1$) along with the same base transfer of $t(1) = t(2) = -2.05$. The corresponding level of effort exerted by the employee is $e = 3/2$. The high type separates by giving half of the profit to the employee ($s(4) = 1/2$) along with a base transfer of $t(4) = -.05$; the corresponding level of effort exerted by the employee is $e = 2$. This outcome, which gives each principal type a strictly higher payoff than the principal-optimal safe outcome, is payoff-plausible because both the low and medium types get at least their first-best payoff, while the high type's payoff precisely equals their plausibility threshold.

In contrast, if only explicit contracts can be proposed, payoff-plausibility selects the principal-optimal safe outcome both in this example and in a broad set of environments.

Definition 6. *An environment is* **quasi-strict at** $x \in X$ *if*

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

$$u(\theta, x, y') > u(\theta', x, y) - u(\theta', x, y').$$

An environment is **quasi-strict** *if it is quasi-strict at every $x \in X$.*

Quasi-strictness strengthens some of the maintained assumptions to hold strictly.

Proposition 2. *In quasi-strict environments, payoff-plausibility selects the principal-optimal safe outcomes when contracts must be explicit.*

When only explicit contracts can be proposed, payoff-plausibility precludes pooling in quasi-strict environments, because the highest type $\overline{\theta}$ involved in pooling would gain strictly more than the lower types from being recognized as $\overline{\theta}$, and the agent's expected utility conditional on the highest pooling type must be weakly positive.[13] Separating outcomes obtained with explicit contracts must be safe, so the result then follows since payoff-plausibility requires that every principal type obtain at least their principal-optimal safe payoff.[14]

6.3 Doubly Complementary Environments

Even with flexible contracts, payoff-plausibility does select the principal-optimal safe outcomes in environments where there are complementarities between the principal's action and the principal's type and agent's action. In these environments, $X = X_1 \times X_2 \times ... \times X_K$ is a Cartesian product of various component spaces, and $X_1 = [\underline{x}_1, \overline{x}_1] \subset \mathbb{R}$ so that one of the action component spaces is an interval of real numbers. To avoid boundary issues, we assume that $\max_{y \in Y} u(\theta_n, \overline{x}_1, x_{-1}, y) + v(\theta_n, \overline{x}_1, x_{-1}, y) < 0$ for all $x_{-1} \in X_2 \times ... \times X_K$, which ensures that the highest value of x_1 is prohibitively costly.

Definition 7. *An environment is* **doubly complementary** *if it satisfies:*

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

[14] The firm-employee example is not quasi-strict, because the strict complementarity conditions fail at $s = 0$, and the strict monotonicity condition and second strict complementrity condition fail at $s = 1$. Section OA.3 states and proves a more general version of Proposition 2 that does cover the example. Intuitively, neither the issues at $s = 0$ nor $s = 1$ prevent the conclusion of Proposition 2, because quasi-strictness holds at arbitrarily close values of s.

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

The first condition says the agent's best response is weakly increasing in the x_1 component of the principal's action. The second condition requires that the difference in principal utility from a higher x_1, holding fixed the remaining components of the principal's action as well as the agent's action, is higher for a higher principal type, and strictly so at points where when the higher principal type gets a strictly higher utility than the lower type.

These requirements are satisfied in many economic applications, such as the following modified version of the informed firm and employee example.

Example 2. As before, the firm has private information $\theta \in \{2, 4\}$ about the profitability or quality of a task for which they seek to hire an employee, a hired employee will choose an effort level $e \in \mathbb{R}_+$ that affects the probability of the task being successful, and the firm will pay a profit-share s and transfer t to the agent. However, unlike before, the firm makes a costly investment $i \in \mathbb{R}_+$ that increases the productivity of the employee's effort. The utility functions of the firm and employee are $U(\theta, i, s, t, e) = \theta(1-s)\ln(1+i)e/2 - i^2/2 - t$ and $V(\theta, i, s, t, e) = \theta s \ln(1+i)e/2 - e^2/2 + t$, respectively. The conditions of Definition 7 can be readily verified when taking i to be the first component of the firm's action. ■

Theorem 3. *In an environment that is doubly complementary and quasi-strict, the payoff-plausible PBE outcomes are the principal-optimal safe outcomes.*[15]

The proof, which is in Section C of the Online Appendix, shows that every payoff-plausible PBE outcome must be such that a (θ, x, t) from which the agent would get

[15] Because the environment in the firm and employee example is not quasi-strict, the assumptions of Theorem 3 are not met. In Appendix C, we state and prove a stronger version of the theorem that covers the doubly complementary firm and employee example.

strictly positive utility while playing $y^*(\theta, x)$ occurs with 0 probability. Intuitively, for any PBE outcome in which such a (θ, x, t) occurs with positive probability, there is some $\alpha \in (0, 1]$ and $y \in Y$ such that θ gets their equilibrium payoff from (α, x, t, y) while all other types would get a weakly lower payoff. Without loss of generality, we can take θ to be the highest type for which (α, x, t, y) occurs with positive probability, which means that $y \leq y^*(\theta, x)$. Then θ could propose an action x' with a slightly increased first component relative to x, and adjust their transfer so that if the agent accepts and plays $y^*(\theta, x')$, the agent obtains a strictly higher payoff than 0, while θ is strictly better off, and every lower type is strictly worse off than in equilibrium. But this violates payoff-plausibility. The proof then shows that, since the agent's expected utility in PBE must be non-negative, the probability of a (θ, x, t, y) from which the agent gets a strictly negative utility is 0 in all payoff-plausible PBE outcomes. It then follows that the probability of a (θ, x, t, y) for which $y \neq y^*(\theta, x)$ must also be 0, so every payoff-plausible outcome must be safe. Since, by Theorem 2.1, every payoff-plausible outcome principal-payoff-dominates the principal-optimal safe outcome, it follows that every payoff-plausible outcome must be a principal-optimal safe outcome.

This result would not hold if the principal could commit to arbitrary randomizations over action-transfer pairs. Intuitively, this is because a high-type principal and low-type principal could pool in such a way that the agent's expected utility conditional on the high type is negative. So while the the agent's actions would increase if the high-type principal were to separate from the low-type principal, doing so would be prohibitively costly for the high-type principal.[16] Thus, Theorem 3 illustrates a significant qualitative difference in results and tightness of conclusions between this framework and those in which the principal has unlimited commitment power.

[16] OA.4.2 presents an example exhibiting this phenomena.

7 Conclusion

This paper takes a novel approach to the study of informed principal environments. In particular, we avoided making either the extreme assumption that contracts precisely determine the principals' future actions or the extreme assumption that the principal can commit to an arbitrary mechanism requiring a high degree of complexity and commitment power. Instead, we assumed that the principal can propose a contract that gives them flexibility over their future actions, but does not allow them to commit to non-degenerate randomizations.

Using this approach, we focused on a natural class of environments with complementarities. The principal-optimal safe outcomes are important equilibrium objects. They are always payoff-plausible PBE outcomes, and they give a threshold for the principal-type payoffs that must be met in every payoff-plausible PBE. Unlike when only explicit contracts can be proposed, payoff-plausibility does not generally select the principal-optimal safe outcomes. However, payoff-plausibility with flexible contracts does select precisely the principal-optimal safe outcomes in the subclass of doubly complementary environments, which is not the case when the principal can commit to an arbitrary mechanism. This is one illustration of the greater tightness and prediction power that obtains when assuming flexible contracts with limited rather than unlimited commitment. Hopefully this approach and similar frameworks will be useful due to the greater realism of the underlying assumptions and the appealing nature of the results generated.

References

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

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

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

N. C. Bedard. The strategically ignorant principal. *Games and Economic Behavior*, 102:548–561, 2017b.

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

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

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

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

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

D. Clark. The informed principal with agent moral hazard. Working Paper, 2022.

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

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

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

A. Dosis. On the informed principal model with common values. *Rand Journal of Economics* (forthcoming), 2022.

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

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

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

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

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

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

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

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

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

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

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

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

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

Y. Sun. Contracts that reward innovation: Delegated experimentation with an informed principal. Working Paper, 2021.

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

A Proof of Theorem 1

Lemma 1. *For every contract $C \in \mathcal{C}$, there is a continuation PBE following the proposal of C that results in each principal type receiving a weakly lower payoff than their principal-optimal safe payoff.*

As discussed before, the principal-agent game is modified so that, for an arbitrary $(x, t) \in X \times \mathbb{R}$, each principal type θ can forego proposing a contract and automatically obtain their payoff from $(x, t, y^*(\theta, x))$ under certain conditions. These conditions are that (1) the agent would get a weakly positive payoff from $(x, t, y^*(\theta, x))$, and (2) every lower type $\theta' < \theta$ would obtain a lower payoff from $(x, t, y^*(\theta, x))$ than they do in equilibrium. When these conditions are violated, the principal type θ experiences costs from taking this option, which become prohibitively high in the limit. Additionally, we give each principal type a small additional utility $\eta > 0$ from taking the automatic option, so that regular contracts are proposed and accepted with probability 0 in the limit. This is because the agent cannot get strictly negative utility in the limit, and, for a highest type that, with positive probability, proposes contracts and plays subsequent action-transfer pairs that give the agent a weakly positive utility, there would be a profitable deviation to one of these modified outside options. So, for each type θ, with probability 1, either their outside option is realized, or they are taking a modified outside option corresponding to some $(x, t, y^*(\theta, x))$. But as discussed above, in the limit, no type plays such an outside option when either the agent's resulting utility would be negative or there is a lower type that would get a higher-than-equilibrium payoff by mimicking them. These facts, along with the characterization of the principal-optimal safe payoffs in Proposition 1, guarantee that, in the $\eta \to 0$ limit, no principal type obtains a higher payoff than their principal-optimal safe payoff. Otherwise, there are essentially no changes from the true principal-agent game, which is why the play following a given contract corresponds to a valid continuation PBE in the true game.

Proof of Lemma 1. Let $\{X_j\}_{j\in\mathbb{N}}$, $\{T_j\}_{j\in\mathbb{N}}$, and $\{Y_j\}_{j\in\mathbb{N}}$ be sequences of finite action and transfer sets such that $\lim_{j\to\infty} X_j = X$, $\lim_{j\to\infty} T_j = \mathbb{R}$, and $\lim_{j\to\infty} Y_j = Y$. For a given $j \in \mathbb{N}$, let $\mathcal{C}_j = P(X_j \times T_j) \setminus \{\emptyset\}$ be the set of non-empty subsets of $X_j \times T_j$. Additionally, fix some $B_j > \max_{(\theta,x,t,y)\in\Theta\times X_j\times T_j\times Y_j} \max\{|u(\theta, x, y) - t|, |v(\theta, x, y) + t|\}$.

We now describe the strategy space of the type θ principal in the j-th game. Part of this player's choice is over which contracts to propose. We force θ to propose all contracts in $\mathcal{C}_j$ with positive probability. Additionally, we allow θ to propose special contracts of the form (θ, x, t), and we use $\mathcal{C}_{\theta,j} = \{(\theta, x, t) : x \in X_j, t \in T_j\}$ to denote the set of such contracts. Moreover, we prevent θ from proposing any contract of the form (θ', x, t) with $\theta' \neq \theta$. Thus the distribution over contract proposals used by θ must belong to

$$\Delta_{j,\theta}(\mathcal{C}_j \cup \mathcal{C}_{\theta,j}) = \Big\{ \mathscr{C} \in \Delta(\mathcal{C}_j \cup \mathcal{C}_{\theta,j}) : \;\; \mathscr{C}(C) \geq \frac{1}{j|\mathcal{C}_j|(1 + B_j)} \; \forall C \in \mathcal{C}_j \setminus \mathcal{M}^0_{j,\theta} \Big\}.$$

Moreover, when a given contract is accepted, we force θ to tremble and play every option in the contract with positive probability. Formally, the distribution over action-transfer pairs used by θ when contract C is accepted must belong to

$$\Pi_j(C) = \Big\{ \chi \in \Delta(C) : \chi(x, s, t) \geq \frac{1}{j|\mathcal{C}|(1 + B_j)} \; \forall (x, s, t) \in C \Big\}.$$

A valid strategy for θ in the j-th game is any pair $(\mathscr{m}_\theta, \boldsymbol{\chi}_\theta)$ consisting of a $\mathscr{m}_\theta \in \Delta_{j,\theta}(\mathcal{C}_j \cup \mathcal{C}_{\theta,j})$ and a rule $\boldsymbol{\chi}_\theta : \mathcal{C}_j \to \Delta(X_j \times T_j)$ for how to play when an arbitrary contract in $\mathcal{C}_j$ is accepted that satisfies $\boldsymbol{\chi}_\theta(C) \in \Pi_j(C)$ for all $C \in \mathcal{C}_j$.

The strategy space of the agent is unaltered from the principal-agent game, aside from the addition of trembles. For every contract (μ, M_P), we require the probability α that the agent accepts its proposal to be no less than $1/(j(1+B_j))$. A valid strategy for the agent in the j-th game is any pair $(\boldsymbol{\alpha}, \mathbf{y})$ consisting of (1) a rule governing the probability of mechanism acceptance $\boldsymbol{\alpha} : \mathcal{C}_j \to [1/(j(1 + B_j)), 1]$ and (2) a rule governing the agent's choice of actions $\mathbf{y} : \mathcal{C}_j \times X_j \times T_j \to \Delta(Y_j)$.

We now develop the payoffs of the various players for an arbitrary strategy profile ζ. Fix an $\eta > 0$. For any $\theta \in \Theta$, let $\widetilde{U}_j(\theta, C, \boldsymbol{\alpha}, \boldsymbol{\chi}_\theta, \mathbf{y})$ and $\widetilde{V}_j(\theta, C, \boldsymbol{\alpha}, \boldsymbol{\chi}_\theta, \mathbf{y})$ be the unmodified expected payoffs to the principal and agent, respectively, when the principal's type is θ, the contract $C \in \mathcal{C}_j$ is proposed, the agent uses the acceptance probability rule $\boldsymbol{\alpha}$, and subsequent play is governed by the rules $\boldsymbol{\chi}_\theta$ and $\mathbf{y}$.

The agent's payoff is

$$V_j(\zeta) = \sum_{\theta \in \Theta} \lambda(\theta) \left[\sum_{C \in \mathcal{C}_j} \mathscr{C}_\theta(C) \widetilde{V}_j(\theta, C, \boldsymbol{\alpha}, \boldsymbol{\chi}_\theta, \mathbf{y}) \right].$$

This is precisely the agent's total expected utility from play over contracts in $\mathcal{C}_j$.

We require more notation to specify the payoffs of the principal types. We define the type θ virtual payoff from strategy profile ζ

$$\begin{aligned}\widehat{U}_j(\theta,\zeta) = &\sum_{C\in\mathcal{C}_j} \mathscr{C}_\theta(C)\widetilde{U}_j(\theta,C,\boldsymbol{\alpha},\boldsymbol{\chi}_\theta,\mathbf{y}) \\ &+ \sum_{(x,t)\in X_j\times T_j} \mathscr{M}_\theta(\theta,x,s,t)\left(u(\theta,x,y^*(\theta,x)) - \left(1-\frac{\theta}{j}\right)s - t + \eta\right)\end{aligned}$$

to be the total expected utility of the principal if the principal's type were θ, the principal followed the contract proposal rule $\mathscr{C}_\theta$, the play that followed proposal of $C \in \mathcal{C}_j$ proceeded according to the rules $\boldsymbol{\alpha}$, $\boldsymbol{\chi}_\theta$, and $\mathbf{y}$, the agent were to accept the proposal of an arbitrary $(\theta,x,t)\in\mathcal{C}_{\theta,j}$ with probability 1 and subsequently take action $y^*(\theta',x)$, and the principal received an additional η in utility from proposing a contract of the form (θ,x,t). We will impose modifications to the payoffs of the principal types so that it is costly for θ to propose any $(\theta,x,t)\in\mathcal{C}_{\theta,j}$ whenever either som lowere principal type $\theta' < \theta$ would prefer to propose (θ,x,t) (and have the agent respond according to $y^*(\theta,x)$) to their outcome or the agent gets a low utility from $(\theta,x,t,y^*(\theta,x))$. Let $A > \max_{(\theta,x,t,y)} u(\theta,x,y) + v(\theta,x,y)$, and let $f_j : \mathbb{R}\to\mathbb{R}_+$ be the continuous function given by $f_j(z) = \max\{0, A\min\{jz,1\}\}$. Note that $f_j(z) = 0$ for all $z\le 0$ and j, and $\lim_{j\to\infty} f_j(z) = A$ for all $z>0$. Let $c_{j,\theta,\zeta} : \mathcal{C}_{\theta,j}\to\mathbb{R}_+$ be the "cost" function given by

$$c_{j,\theta,\zeta}(\theta,x,t) = \sum_{\theta'<\theta} f_j\left(u(\theta',x,y^*(\theta,x)) - t - \widehat{U}_j(\theta',\zeta)\right) + f_j\left(v(\theta',x,y^*(\theta,x)) + t\right).$$

Note that $c_{j,\theta,\zeta}(\theta,x,t) \ge A$ if some principal type $\theta'\ne\theta$ would get a payoff from $(x,t,y^*(\theta,x))$ that exceeds their virtual payoff by $1/j$ or the agent would get a utility lower than $-1/j$ from $(\theta,x,t,y^*(\theta,x))$. On the other hand, $c_{j,\theta,\zeta}(\theta,x,t) = 0$ if every principal type $\theta'\notin\widetilde{\Theta}$ gets a weakly higher virtual payoff than they would from $(x,t,y^*(\theta,x))$ and the agent gets a weakly positive utility from $(\theta,x,t,y^*(\theta,x))$. We set the payoff of θ from the strategy profile ζ in the j-th game to be

$$U_j(\theta,\zeta) = \widehat{U}_j(\theta,\zeta) - \sum_{(x,t)\in X_j\times T_j} \mathcal{C}_\theta[(\theta,x,t)]c_{j,\theta,\zeta}(\theta,x,t).$$

The important feature of the cost terms is that θ would never want to propose a $(\theta,x,t) \in \mathcal{C}_{\theta,j}$ if either $u(\theta',x,y^*(\theta,x)) - t \geq \widehat{U}_j(\theta',\zeta) + 1/j$ for some $\theta' \neq \theta$ or $v(\theta,x,y^*(\theta,x)) + t \leq -1/j$. 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 $v(\theta,x,y^*(\theta,x)) + t \geq 0$, then the artificial cost from proposing (θ,x,t) is 0 for θ.

Standard arguments show that the j-th game has a Nash equilibrium. Let $p_j \in \Delta(\Theta \times ((0,1] \times X_j \times T_j \times Y_j \cup \{o\}))$ be the outcome induced by the corresponding contract proposal strategies used by the principal types and the following continuation play for each contract: For any $C \in \mathcal{C}_j$, the principal types and agent play as they do in the Nash equilibrium, and, for any $(\theta,x,t) \in \mathcal{C}_{\theta,j}$, the agent accepts with probability 1 and then plays $y^*(\theta,x)$. Suppose (by restricting attention to a subsequence if needed) that $\lim_{j\to\infty} U_j(\theta,p_j)$ exists for all $\theta \in \Theta$ and that there is some $p^* \in \Delta(\Theta \times ((0,1] \times X \times \mathbb{R} \times Y \cup \{o\}))$ such that $\lim_{j\to\infty} p_j = p^*$. Standard upper hemicontinuity arguments show that, for every $C \in \mathcal{C}$, there is some continuation PBE that gives every principal type $\theta \in \Theta$ a weakly lower payoff than $U(\theta,p^*) + \eta$. We will show that $U(\theta_n,p^*) \leq U^*(\theta_n) + \eta(n-1)$ for all $n \in \{1,...,N\}$, where $U^*(\theta)$ is the type θ principal-optimal safe payoff. Since $\eta > 0$ can be arbitrarily chosen, combining this with further standard upper hemicontinuity arguments then implies that, for every $C \in \mathcal{C}$, there is some continuation PBE that gives every principal type a weakly lower payoff than their principal-optimal safe payoff.

We establish that, as $j \to \infty$, the probability of a contract in $\mathcal{C}_j$ being proposed and accepted in the j-th Nash equilibrium converges to 0. Suppose towards a contradiction that this probability does not converge to 0. Then there must be some $(\theta,\alpha,x,t,y) \in \Theta \times (0,1] \times X \times \mathbb{R} \times Y$ such that (1) $\alpha(u(\theta,x,y) - t) \geq \lim_{j\to\infty} U_j(\theta,p_j)$, (2) $y \leq y^*(\theta,x)$, (3) $\alpha(u(\theta',x,y) - t) \leq \lim_{j\to\infty} U_j(\theta',p_j)$ for all $\theta' < \theta$, and (4) $v(\theta,x,y) + t \geq 0$. We first establish that $p^*(v(\theta,x,y)+t \geq 0|\alpha > 0) = 1$. This means that, for all sufficiently high j, there is an $(x_j,t_j) \in X_j \times T_j$ such that (1) $u(\theta,x_j,y^*(\theta,x_j)) - t_j + \eta/4 > U_j(\theta,p_j)$, (2) $u(\theta',x_j,y^*(\theta,x_j)) - t_j < \widehat{U}_j(\theta',p_j)$, and (3) $v(\theta,x,y^*(\theta,x_j)) + t_j > 0$. However, this means that, for all sufficiently high j, the type θ principal could deviate from the equilibrium by proposing (θ,x_j,t_j) with the highest probability possible and thereby secure a payoff strictly higher by $\eta/2 > 0$, contradicting Nash equilibrium.

It thus follows that, for every $\theta \in \Theta$, either $U(\theta,p^*) = 0$ or there is some $(x,t) \in X \times T$ such that $U(\theta,p^*) \leq u(\theta,x,y^*(\theta,x)) - t$, $U(\theta',p^*) + \eta \geq u(\theta',x,y^*(\theta,x)) - t$ for all $\theta' < \theta$, and $v(\theta,x,y^*(\theta,x)) + t \geq 0$. We establish by induction that $U(\theta_n,p^*) \leq U^*(\theta_n) + \eta(n-1)$ for all $n \in \{1,...,N\}$. Consider first $n = 1$. If $U(\theta_1,p^*) = 0$, then this holds trivially. If instead there is some $(x,t) \in X \times T$ such that $U(\theta_1,p^*) \leq u(\theta_1,x,y^*(\theta_1,x)) - t$ and $v(\theta_1,x,y^*(\theta_1,x)) + t \geq 0$, then (x,t) solves the type θ_1 optimization problem given by (2), so $U(\theta_1,p^*) \leq U^*(\theta_1)$. Assume that $U(\theta_{n'},p^*) \leq U^*(\theta_{n'}) + \eta(n'-1)$ for all $n' < n$. If $U(\theta_n,p^*) = 0$, then $U(\theta_n,p^*) \leq U^*(\theta_n) + \eta(n-1)$ holds trivially. If instead there is some $(x,t) \in X \times T$ such that $U(\theta_n,p^*) \leq u(\theta_n,x,y^*(\theta_n,x)) - t$, $U(\theta_{n'},p^*) + \eta \geq u(\theta_{n'},x,y^*(\theta,x)) - t$ for all $n' < n$, and $v(\theta_n,x,y^*(\theta_n,x)) + t \geq 0$, then $(x,t+\eta(n-1))$ solves the type θ_n optimization problem given by (2), so $U(\theta_n,p^*) - \eta(n-1) \leq u(\theta_n,x,y^*(\theta_n,x)) - t - \eta(n-1) \leq U^*(\theta_n)$, which gives $U(\theta_n,p^*) \leq U^*(\theta_n) + \eta(n-1)$. ■

Proof of Theorem 1. Consider a principal-optimal safe outcome p. Fix an arbitrary $x' \in X$ and $\bar{t}$ such that $\bar{t} > \max_{(\theta,x,y)} \max\{u(\theta,x,y), v(\theta,x,y)\}$. Further, for every $\theta \in \Theta$, let $A_\theta = \{\alpha \in (0,1] : p(\theta,\alpha) > 0\}$ and, for every $\alpha \in A_\theta$, consider the contract $C_{\theta,\alpha} = \{(x,t) \in X \times \mathbb{R} : p(\theta,\alpha,x,t) > 0\} \cup \{(x',\bar{t}+\theta)\}$. This contract consists of all the principal-action transfer pairs that are chosen by type θ following the acceptance of contracts that are accepted with probability α under p, as well $(x',\bar{t}+\theta)$, an option which type θ would never play but will be used to ensure that $C_{\theta,\alpha}$ is played only by type θ. We will use the contract $C_o = \{(x',-\bar{t}-\theta)\}$ to induce the outside option since this would never be accepted by the agent.

We identify a candidate PBE that induces p. For each principal type, the contract proposal distribution $\mathscr{C}_\theta \in \Delta(\mathcal{C})$ is such that, for every $\alpha \in (0,1]$ satisfying $p(\theta,\alpha) > 0$, $\mathscr{C}_\theta(C_{\theta,\alpha}) = p(\theta,\alpha)/\alpha$, and $\mathscr{C}_\theta(C_{\theta,o}) = 1 - \sum_{\alpha \in A_\theta} p(\theta,\alpha)/\alpha$. Following the proposal of a contract of the form C_θ, the agent's belief puts probability 1 on the principal's type being θ and the agent rejects with probability 1. Following the proposal of a contract of the form $C_{\theta,\alpha}$ for some $\alpha \in A_\theta$, the agent's belief puts probability 1 on the principal's type being θ, and the agent accepts with probability α and then plays $y^*(\theta,x)$ should $(x,t) \in C_{\theta,\alpha}$ be observed. Should a contract of the form $C_{\theta,\alpha}$ be proposed and accepted, the type θ principal chooses among the allowed options according to the probability distribution χ_θ given by $\chi_\theta(x,t) = p(\theta,\alpha,x,t)/p(\theta,\alpha)$, while all other principal types choose according to some probability distribution that puts full weight on their optimal options in $C_{\theta,\alpha}$ given that the agent will respond according to $y^*(\theta,x)$. Additionally, let the strategies of the principal types, the agent's strategy, and the agent's belief update rule be such that, for all other contracts, the induced play following the contract's proposal matches that of a continuation PBE which deters proposal of the contract. By construction, this strategy profile and belief update rule constitute a PBE and induce outcome p. ■

B Proof of Theorem 2.1

Proof. Let p be a payoff-plausible outcome. Suppose towards a contradiction that there is some $n \in \{1,...,N\}$ for which θ_n obtains a lower expected utility than their principal-optimal safe payoff, and let n be the lowest such value. Since $U(\theta_n,p) \geq 0$, there must be some $(x,t,y^*(\theta_n,x))$ that gives type θ_n their principal-optimal safe payoff, every lower type $\theta_{n'}$ for $n' < n$ a weakly lower payoff than their principal-optimal safe payoff, and the agent a weakly positive utility conditional on type θ_n. Then (x,t) satisfies the constraints in the type-θ_n problem given by (1). Thus, we have $U(\theta_n,p) \geq u(\theta_n,x,y^*(\theta_n,x)) - t$, which contradicts θ_n obtaining a lower expected utility than their principal-optimal safe payoff. ■

C Proof of Theorem 3

The following generalization of Theorem 3 implies that payoff-plausibility selects the principal-optimal safe outcomes in Example 2.

Theorem 3′. *Consider a doubly complementary environment in which, for every $x \in X$, either quasi-strictness holds at x or there exists a sequence $\{x_i\}$ converging to x such that quasi-strictness holds at each x_i. Then payoff-plausibility selects the principal-optimal safe outcomes.*

Proof. Consider a payoff-plausible PBE outcome p. We first show that $p(v(\theta, x, y^*(\theta, x))+ t > 0) = 0$. Suppose towards a contradiction that there is some θ such that $p(v(\theta, x, y^*(\theta, x))+ t > 0|\theta) > 0$, and suppose that $\overline{\theta}$ is the highest type for which this is true. Then there are $\alpha \in (0, 1]$, $x \in X$, $t \in \mathbb{R}$, and $\tilde{\lambda} \in \Delta(\Theta)$ such that (1) $\alpha(u(\overline{\theta}, x, y^*(\tilde{\lambda}, x)) - t) = U(\overline{\theta}, p)$, (2) $\alpha(u(\theta, x, y^*(\tilde{\lambda}, x)) - t) \leq U(\theta, p)$ for all $\theta \neq \overline{\theta}$, (3) $\tilde{\lambda}(\theta \leq \overline{\theta}) = 1$, and (4) $v(\overline{\theta}, x, y^*(\overline{\theta}, x)) + t > 0$. Consider (x', t') such that $t' = \alpha t + u(\overline{\theta}, x', y^*(\overline{\theta}, x')) - \alpha u(\overline{\theta}, x, y^*(\tilde{\lambda}, x))$. By construction, this (x', t') is such that, when the agent responds with $y^*(\overline{\theta}, x')$, the type $\overline{\theta}$ principal obtains the same payoff as in p. Moreover, we can choose an x' with $x'_k = x_k$ for all $k \neq 1$ and $x'_1 > x_1$ close enough to x_1 so that all lower type principals would achieve a strictly lower payoff from $(x', t', y^*(\overline{\theta}, x'))$ than p and the agent gets a strictly higher utility from $\overline{\theta}$ playing (x', t') than their outside option. Thus, for sufficiently small $\varepsilon > 0$, $(x', t' - \varepsilon)$ would satisfy the constraints of the type $\overline{\theta}$ optimization problem in (1) and give $\overline{\theta}$ a strictly higher payoff than in p, which contradicts payoff-plausibility.

Since the agent's total expected utility must be weakly positive, it thus follows that $p(v(\theta, x, y)+t < 0) = 0$. Thus, $p(y \neq y^*(\theta, x)) = 0$, and $p(\alpha < 1, v(\theta, x, y)+t > 0) = 0$, so p must be safe. As p is an arbitrary payoff-plausible PBE outcome, we conclude that every payoff-plausible PBE outcome must be safe. As every payoff-plausible outcome must principal-payoff-dominate the principal-optimal safe outcome, it follows that only the principal-optimal safe outcomes can be payoff-plausible. ■