Discussion Paper Series – CRC TR 224

Discussion Paper No. 643
Project B 01

Neo-Optimum: A Unifying Solution to the Informed-Principal Problem

Tymofiy Mylovanov[1]
Thomas Tröger[2]

February 2025

[1]Department of Economics, University of Pittsburgh, mylovanov@gmail.com.
[2]Department of Economics, University of Mannheim, troeger@uni-mannheim.de

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Neo-Optimum: A Unifying Solution to the Informed-Principal Problem[*]

Tymofiy Mylovanov[†]and Thomas Tröger[‡]

January 21, 2025

Abstract

A mechanism proposal by a privately informed principal is a signal. The agents' belief updating endogenizes their incentives in the mechanism, implying that such design problems cannot be solved via optimizing subject to incentive constraints. We propose a solution, neo-optimum, that can be interpreted as principal-preferred perfect-Bayesian equilibrium. Its neologism-based definition allows an intuitive computation, as we demonstrate in several applications. Any Myerson neutral optimum is a neo-optimum, implying that a neo-optimum generally exists. In private-values environments, neo-optimum is equivalent to strong unconstrained Pareto optimum (Maskin-Tirole) and strong neologism-proofness (Mylovanov-Tröger). In information-design settings, any interim-optimum (Koessler-Skreta) is a neo-optimum. Our methods can be used to reconstruct the perfect-Bayesian equilibria in the informed-principal literature.

[*]Support by the German Research Foundation (DFG) through CRC TR 224 (Project B01) is gratefully acknowledged. Support by the National Science Foundation through grant 0922365, as well as the hospitality of the Study Center in Gerzensee, Switzerland, are also gratefully acknowledged.

[†]Department of Economics, University of Pittsburgh. Email: mylovanov@gmail.com.

[‡]Department of Economics, University of Mannheim. Email: troeger@uni-mannheim.de.

1 Introduction

Mechanism design—the theory of designing rules of interaction that provide incentives to reveal private information towards maximizing the principal's goal—is a cornerstone of economics, with many applications ranging from procurement and the regulation of firms (Laffont and Tirole, 1993) to the design of public institutions (Laffont, 2000), from macroeconomics (Kocherlakota, 2006) to testing and social distancing in pandemics (Tröger, 2025).

The standard approach to solving a mechanism-design problem relies on the revelation principle, which implies that any mechanism induces a direct revelation mechanism, thus transforming the design problem into an optimization subject to incentive constraints. As first recognized by Myerson (1983) and Maskin and Tirole (1990, 1992), this approach fails if the principal is privately informed about her own goals or about anything that concerns the incentives of the agents who participate in the mechanism. The proposal of a mechanism then is a signal in a signaling game that leads to an endogenous updated belief about the principal's private-information type. The incentives of the agents in the mechanism depend on the updated belief. Thus, the answer to the question *which* direct revelation mechanism is induced by a given mechanism proposal becomes endogenous.

An example is a seller who designs a profit-maximizing sales contract while being privately informed about the quality of her good. If, say, the contract includes a warranty then a buyer may be willing to pay a higher price not only because she gets the warranty, but also because she is triggered into believing in a higher quality of the good.

Modeling the principal as just a sender in a signaling game would, however, neglect the mechanism-design doctrine according to which the principal can select an equilibrium for the interaction. It is, for example, implausible that the principal offers a mechanism that, at the prior belief, yields a smaller payoff to all her private-information types than another mechanism. Yet, offering the low-payoff mechanism can be consistent with the logic of perfect-Bayesian equilibrium in signaling games because offering any alternative mechanism may trigger a "pessimistic" off-path belief that makes it unattractive.

The purpose of the paper is to provide a refinement of perfect-Bayesian equilibrium, neo-optimum, that is particularly suitable for informed-principal settings because it can be seen as "sender-preferred equilibrium". Neo-optimum exists broadly and connects the known solution approaches to informed-principal problems. Moreover, our methods can be used to reconstruct all (including non-refined) perfect-Bayesian equilibria in virtually all existing papers in the informed-principal literature.

Neo-optima are always at the weak Pareto frontier of the set of perfect-Bayesian equilibria. Thus the predictions of neo-optimum are generally different from refinements such as the intuitive criterion (Cho and Kreps, 1987) that are based on Kohlberg-Mertens stability. But neo-optimum is also generally different from the principal's ex-ante-optimal perfect-Bayesian equilibrium. The ex-ante criterion has the obvious limitation that rescaling the payoff function of some type of the principal—although being strategically irrelevant—can easily change the prediction. Neo-optimum is invariant with respect to payoff rescalings. Nevertheless, it turns out that in many settings the neo-optima are also ex-ante optimal for the principal.

To capture the broadest range of settings, we introduce a reduced-form description of informed-principal problems that focusses on the principal's payoff. A *payoff vector* refers to a payoff for each private-information type of the principal. For any (prior or updated) belief about the principal's type, a set of *feasible* payoff vectors is given. That's all.

In any particular application, the set of feasible payoff vectors at any belief will be determined by the details of the interaction. For example, a seller may be able to choose from a set of sales contracts, and each continuation equilibrium in the interaction following the proposal of a sales contract induces a particular feasible payoff vector. The reduced form specifies the feasibility sets for *all* beliefs because, a priori, the principal's proposal may trigger any updated belief.

While a reduced-form description is possible for any signaling game,[1] the feasibility structure of informed-principal problems has a very useful additional property, *composition-closedness*, that allows us build our entire analysis around the feasible belief-payoff-vector pairs. Composition-closedness can be seen as a reduced-form version of the inscrutability principle in Myerson (1983). To define, consider a finite family of belief-payoff-vector pairs. Imagine the principal has to choose one pair from this family. A *composition* is a belief-payoff-vector pair that arises from each principal type's payoff-maximizing choice from the family is Bayes consistent. Composition-closedness requires that any composition of feasible belief-payoff-vector pairs

[1] E.g., Mailath (1987).

is feasible.[2] The justification for composition-closedness is that for any finite family of mechanisms there exists a "grand" mechanism in which the principal gives herself the option to select a mechanism from the family.

Both established approaches to defining perfect-Bayesian equilibria for informed-principal signaling games, by Myerson (1983) and by Maskin and Tirole (1990, 1992), fit into our model. While the two approaches differ subtly with respect to the set of allowed mechanisms, in either approach, composition-closedness implies that focussing on fully pooling equilibria is without loss of generality, that is, all types of the principal propose the same mechanism on the equilibrium path.

Let us explain neo-optimum in more detail. A crucial ingredient is the concept of a neologism that is well-established in signaling games (Farrell, 1993). A neologism is a Bayes-consistent and feasible deviation relative to a given (not necessarily feasible) belief-payoff-vector pair. That is, the deviating payoff vector comes together with a deviating belief that puts probability 0 on types who would be harmed by the deviation, retains the relative likelihood across types that strictly gain, and can shift belief probability mass from indifferent types to strictly gaining types.

If no neologism exists for some payoff vector together with the prior belief, te payoff vector is *neologism-proof.* A payoff vector is a *neo-optimum* if (i) it is feasible at the prior belief and (ii) there exists a payoff vector *below* it that is a *limit* of neologism-proof payoff vectors.

Many established signaling-game refinements, including the intuitive criterion, rely on ideas related to neologisms. Farrell (1993) recognized that asking for a prior-belief feasible *and* neologism-proof payoff vector is generally too much; non-existence occurs in very simple signaling games. The subsequent literature responded by restricting the set of neologisms that are considered legitimate (e.g., Rabin (1990)), or by focussing on settings where existence is guaranteed. Neo-optimum is fundamentally different: it allows for arbitrary neologisms, but drops the requirement that the neologism-proof payoff vector is feasible. Rather, neo-optimum is content with being a limit of neologism-proof payoff vectors, or with being at least as good as such a limit for all types of the principal.

As a concrete example we consider a Spence (1973) job-market setting.

[2] As for a concrete example, suppose the principal can have two types. If the payoff vector $(1, 0)$ is feasible at the belief that puts probability 1 on the horizontal type and the payoff vector $(0, 2)$ is feasible at the belief that puts probability 1 on the vertical type, then composition-closedness requires that the payoff vector $(1, 2)$ is feasible at all beliefs.

It belongs to the class of informed-principal problems analyzed in Maskin and Tirole (1992). The principal is a worker who is privately informed about her productivity type, which can be high or low. To determine her task level and wage, she proposes a mechanism to an employer. One payoff vector that is feasible at all beliefs is the least-cost-separating one. It arises from both types getting a wage equal to their respective productivity types, the low-productivity type getting the lowest possible task level, and the high-productivity type getting the task level at which the low type is indifferent between the two types' outcomes. Another prominent payoff vector, feasible at the prior belief, arises from both types getting the lowest task level and the wage equal to average productivity; we call this the best-pooling payoff vector. It is then easy to see that the unique neo-optimum is the least-cost-separating payoff vector or the best-pooling payoff vector, whichever is preferred by the high-productivity type.[3]

As for the perfect-Bayesian equilibria in the Spence example, we recall that the least-cost-separating payoff vector is a neo-optimum if the prior belief puts a sufficiently high weight on the low-productivity type. From this it is immediate that the least-cost-separating payoff vector, being feasible at all beliefs, is a perfect-Bayesian equilibrium at any interior prior belief. Generalizing this logic to other settings recovers the perfect-Bayesian equilibria constructed in Maskin and Tirole (1992) and a number of subsequent papers (e.g., Koessler and Skreta (2016); Nishimura (2022); Balzer (2017); Dosis (2022); Zhao (2023))

We show that any Myerson (1983) neutral optimum is a neo-optimum. From Myerson's existence result, this implies that a neo-optimum exists in any Bayesian incentive problem as defined by Myerson. Neo-optimum is stronger than Myerson's other solution concepts, core and expectational equilibrium. Whenever it yields a unique prediction it identifies the unique neutral optimum. But neo-optimum is much more intuitive and easier to handle than neutral optimum. In particular, neo-optimum avoids any reference to Myerson's extension axiom that connects solutions across different settings.

Initiated by Maskin and Tirole (1990), a part of the informed-principal literature has considered settings with "private values" (e.g., Myerson (1985); Maskin and Tirole (1990); Tan (1996); Yilankaya (1999); Skreta (2009);

[3] Undefeated equilibrium (Mailath, Okuno-Fujiwara, and Postlewaite, 1993) yields an analogous prediction in the classical job-market signaling game where signals are not mechanisms but just education levels.

Mylovanov and Tröger (2014); Wagner, Mylovanov, and Tröger (2015)). Here, the principal is privately informed about her goals, that is, she "has private information that is not directly payoff relevant to the agents, but may influence her design" (Mylovanov and Tröger, 2012). An example would be a seller with private information about her opportunity cost of selling who designs a profit-maximizing sales procedure. The solution concepts proposed in this context, strong unconstrained Pareto optimum (SUPO) by Maskin and Tirole (1990) and its generalization, strongly neologism-proof allocations by Mylovanov and Tröger (2012, 2014) have so far remained disconnected from Myerson's (1983) approach.

We show that, in the generalized-private-values settings for which Mylovanov and Tröger (2012) show the existence of a strongly neologism-proof allocation, this solution concept is in fact *equivalent* to neo-optimum. In particular, in these settings any neutral optimum is strongly neologism-proof. This result resolves a question that has remained open essentially since the literature started. Another implication is that in quasilinear private-values environments (Mylovanov and Tröger, 2014), any neo-optimum, and thus any neutral optimum, is ex-ante optimal.

Koessler and Skreta (2019) consider an informed principal who can, partially or fully, "certify" her type. The model fits into our framework. Koessler and Skreta (2019) propose a solution concept, strong Pareto optimum (SPO), and show that any prior-feasible SPO is a perfect-Bayesian equilibrium and is ex-ante optimal. But a prior-feasible SPO exists only in settings with sufficiently a rich certifiability structure. Using neo-optimum instead of SPO as a solution concept in their setting, existence is generally guaranteed and the qualitative results of Koessler and Skreta (2019) remain largely intact.

Settings in which the principal is an information designer also fit into our framework. To analyze informed-principal information design, Koessler and Skreta (2023) introduce a new refinement of perfect-Bayesian equilibrium, interim optimality, and provide existence and characterization results. We show that any interim optimum is a neo-optimum. In the different variants of the introductory prosecutor-judge example in Koessler and Skreta (2023) the reverse is also true, that is, any neo-optimum is interim-optimal. Thus, the arguments given by Koessler and Skreta (2023) in favor of interim-optimum as a solution concept—predictive power and robustness—apply similarly to neo-optimum.

Balkenborg and Makris (2015) consider a common-value setup similar to Maskin and Tirole (1992), but in contrast to the latter focus on an equilibrium refinement called *assured allocation* that in two-type settings is the unique Myerson (1983) core allocation and hence is equivalent to neo-optimum and neutral optimum. However, in some settings the assured allocation is dominated by a stochastic allocation and is not a core allocation, implying that it is not a neo-optimum.

Section 2 introduces the model. Section 3 introduces the main concept, neo-optimum. In Section 4 we compare neo-optimum to the solution concepts in Myerson (1983). In Section 5 we compare neo-optimum to the established solution in private-values settings. In Section 6 we compare neo-optimum to the established solutions in settings with certifiability or information-design. Some proofs and examples are in the appendix.

2 Model

2.1 Informed-Principal settings

The *principal* is a privately informed entity. Let T denote the set of the principal's feasible private-information types. For technical simplicity, we assume that T is finite.[4] The set of probability distributions with support in T is denoted B. A *belief* about the principal's type is a $b \in B$, where $b(t)$ denotes the likelihood assigned to type t.

The principal's *payoff* is represented by a *vector* $U \in \mathbb{R}^T$, where $U(t)$ for all $t \in T$ denotes the principal's payoff when she has the type t.

An *informed-principal setting* is characterized by the *set of feasible belief-payoff-vector pairs*

$$K \subseteq B \times \mathbb{R}^T.$$

For any belief $b \in B$, we say that a payoff vector U is *feasible at* b or b*-feasible* if $(b, U) \in K$.[5] We assume that K is topologically closed and, for each $b \in B$, the set of b-feasible payoff vectors is non-empty.

When convenient, we will use the following language. We say that a payoff vector V is *above* a payoff vector U (or U is *below* V), written $V \geq U$, if $V(t) \geq U(t)$ for all $t \in T$. In other words, V is above U if and only if

[4] We see no obstacle against extending our central concept, neo-optimum, to continuous-type settings.

[5] Given any set $M \subseteq B \times \mathbb{R}^T$ and any $b \in B$, we will also use the notation $M(b) = \{U \mid (b, U) \in M\}$. For example, $K(b)$ is the set of b-feasible payoff vectors.

all sender types weakly prefer V to U. (We say that V *dominates* U if V is above U and $V \neq U$.)

Applications

As a first example, following Maskin and Tirole (1992) and inspired by Spence (1973), let the principal be a worker who proposes a mechanism for determining her wage w and task level e to a potential employer. The principal has one of two productivity types, $T = \{\theta_L, \theta_H\}$. Any belief b can be identified with the probability of the high type θ_H, that is, $b \in B = [0, 1]$. Let $w - e/\theta$ denote the principal's payoff if she has the type θ, works at level e, and gets the wage w. The employer then obtains the payoff $\theta - w$, whereas she gets 0 if she does not employ the principal. A mechanism is a game form in which the principal and the employer play, and each end node is a task-level-wage pair $(e, w) \in [0, \infty) \times [0, \theta_H]$. The mechanism is played if the employer accepts it. By the revelation principle, given any belief $b \in B$, the mechanism (or, more precisely, the action of proposing the mechanism) implements a task-level-wage pair[6] $(e_L(b), w_L(b))$ for type θ_L and a task-level-wage pair $(e_H(b), w_H(b))$ for type θ_H such that incentive compatibility is satisfied,

$$w_L - e_L/\theta_L \geq w_H - e_H/\theta_L \quad \text{and} \quad w_H - e_H/\theta_H \geq w_L - e_L/\theta_H, \tag{1}$$

and the employer's participation constraint is satisfied,

$$b(\theta_H - w_H) + (1 - b)(\theta_L - w_L) \geq 0. \tag{2}$$

Thus,

$$K = \{(b, (w_L - \frac{e_L}{\theta_L}, w_H - \frac{e_H}{\theta_H})) \mid (1), \quad (2)\}.$$

After standard manipulations, we obtain the following characterization,

$$\begin{aligned} K \quad = \quad \bigcup_{\overline{e} \geq 0} \bigcup_{b \in B} \{b\} \times \operatorname{conv} \Big\{ & (\theta_L, \theta_L + \frac{(\theta_H - \theta_L)^2}{\theta_H}), (\theta_L - \frac{\overline{e}}{\theta_L}, \theta_L - \frac{\overline{e}}{\theta_L}), \\ & (\theta_L - \frac{\overline{e}}{\theta_L}, \theta_L + \frac{(\theta_H - \theta_L)^2 - \overline{e}}{\theta_H}), \\ & (b\theta_H + (1 - b)\theta_L, b\theta_H + (1 - b)\theta_L) \Big\}, \end{aligned}$$

[6] Given the linearity of the payoff functions, probability distributions over task-level-wage pairs need not be considered.

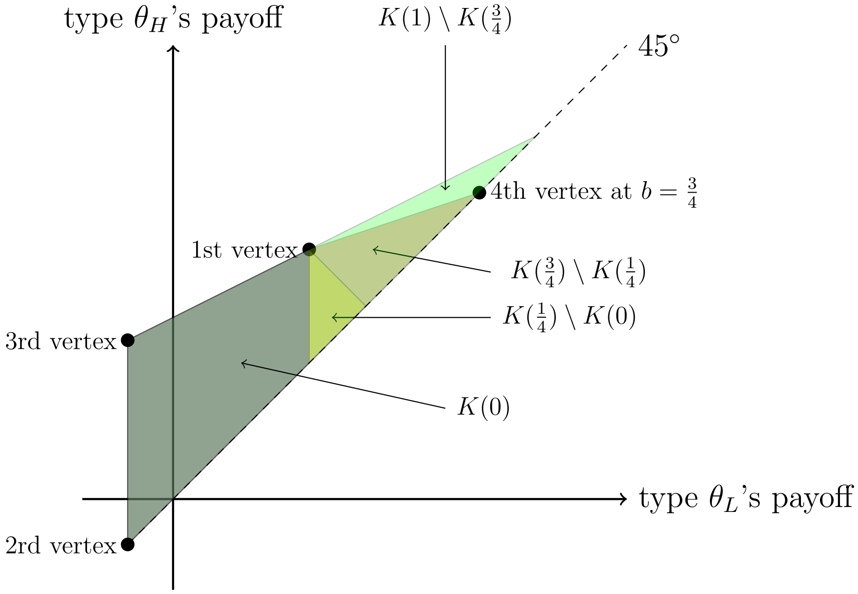
type θ_H's payoff
$K(1) \setminus K(\frac{3}{4})$
45°
4th vertex at $b = \frac{3}{4}$
1st vertex
$K(\frac{3}{4}) \setminus K(\frac{1}{4})$
$K(\frac{1}{4}) \setminus K(0)$
3rd vertex
$K(0)$
type θ_L's payoff
2rd vertex

Figure 1: The sets of feasible payoff vectors at different beliefs b in the Spence-job-market example for generic values of θ_L and θ_H. The vertices refer to the points that span the convex hull for a particular task-level bound $\overline{e}$ in the definition of K. The higher b, the larger the feasibility set so that $K(0) \subset K(1/4) \subset K(3/4) \subset K(1)$.

where conv$\{\dots\}$ is the convex hull of four "vertex" payoff vectors: the first is called least-cost separating, the second is the payoff vector where the low type, θ_L, chooses the task level $\overline{e}$ and gets the wage θ_L while the high type, θ_H, chooses the task level that, at wage θ_H, makes her indifferent to choosing the task level $\overline{e}$ and getting the wage θ_L, the third is the payoff vector where the low type chooses the task level $\overline{e}$ and gets the wage θ_L while the high type chooses the task level such that the low type is made indifferent to choosing the high type's task level and getting the wage θ_H, and the fourth arises from both types pooling at task level 0. The convex hull captures what is feasible at the belief b if there is a highest feasible task level $\overline{e}$.

The projections of K onto the space of payoff vectors for several beliefs b are represented graphically in Figure 1.

Secondly, going beyond a particular example, we would like to emphasize Myerson's (1983) "Bayesian incentive problems" as a broad framework that is covered by our model. The framework captures mechanism design problems with adverse selection and moral hazard, finitely many agents, arbitrary finite outcome spaces and arbitrary payoff functions.[7] A feasibility set K is naturally associated to any Bayesian incentive problem Γ, as follows. For any belief $b \in B$, the set $K(b)$ is determined via the incentive constraints in Γ when b is interpreted as a normalized likelihood vector in the sense of Myerson (1983, Section 5), and in the incentive constraints (1983, (3.1)) of all players except the principal (i.e., player 1) there is an additional factor $b(t_1)$, where t_1 is the first component of the summation variable t_{-i}. (An "incentive compatible mechanism" in the terminology of Myerson (1983) leads to a payoff vector that is b-feasible with b the uniform distribution on T.)

As a third application, our model covers the information-design framework of Bergemann and Morris (2019). Suppose the information designer is one of the players, say player 1 (i.e., $v = u_1$ in the terminology of Bergemann and Morris). As in the previous example, any belief b is interpreted as a normalized likelihood vector in the sense of Myerson (1983, Section 5). For any belief $b \in B$, the set $K(b)$ is determined via the obedience constraints (Definition 1 in Bergemann and Morris (2019)), augmented with an additional factor $b(t_1)$ for all $i \neq 1$.[8] A special case of this framework is Koessler and Skreta (2023), who assume that the principal is fully informed about the state of the world and the other players have no prior information.[9]

Compositions and inscrutiblity

Next we introduce our core structural property for feasibility sets, Assumption 1, which is a reduced-form generalization of Myerson's (1983) inscrutability principle. The core underlying concept, *composition*, is best understood via a thought experiment. Suppose that the principal had to choose a belief-payoff-vector pair from a given family. Depending on her type, she cares about the corresponding dimension in each payoff vector. A *composition* is a belief-payoff-vector pair that arises from a payoff-maximizing choice of each type, and the choice decisions of the various principal types are Bayes-

[7] The Spence example with the particular cost structure described above becomes a Bayesian incentive problem if we cut the outcome space at some highest feasible task level. Then any outcome can be represented as a probability distribution over four outcomes, combined of the highest/lowest task/wage levels.

[8] Correcting a typo in Bergemann and Morris (2019), the summation variables in their formula (2) should be a_{-i} and t_{-i}, where the latter has t_1 as its first coordinate.

[9] Koessler and Skreta (2023) sketch a more general model in Section 8 of their paper.

consistent.

A belief-payoff-vector pair $(b, U) \in B \times \mathbb{R}^T$ is the

$$\textit{composition} \text{ of } \ (b_k, U_k)_{k=1,\ldots,\overline{k}} \ \ (\overline{k} \geq 1, (b_k, U_k) \in B \times \mathbb{R}^T)$$

if, for all $t \in T$,

$$U(t) \ = \ \max_{k=1,\ldots,\overline{k}} U_k(t) \tag{3}$$

and if there exist (choice-probability) functions $c_k : T \to [0, 1]$ $(k = 1, \ldots, \overline{k})$ with $\sum_{k=1}^{\overline{k}} c_k(t) = 1$ $(t \in T)$ such that

$$\text{if } c_k(t) > 0, \text{ then } U_k(t) \geq U_l(t) \text{ for all } l \neq k, \tag{4}$$

and such that, for all k, Bayes' rule is satisfied,

$$c_k(t'')b(t'')b_k(t') = c_k(t')b(t')b_k(t'') \quad \text{for all } t', t'' \in T. \tag{5}$$

To digest the notation, suppose that each type t chooses among the payoff vectors $U_1, \ldots, U_{\overline{k}}$, where type t's probability of choosing any U_k is denoted $c_k(t)$. Then (3) expresses that U is the payoff vector that results from optimal choice, (4) expresses that each type chooses optimally, and (5) expresses that the belief b_k that is formed upon observing the choice k is consistent with the grand belief b and the choice probability distributions c_k. The fact that (5) captures Bayes rule is easiest to see if all numbers involved are strictly positive, implying

$$\frac{b_k(t')}{b_k(t'')} \ = \ \frac{c_k(t')b(t')}{c_k(t'')b(t'')}. \tag{6}$$

The right-hand side in this equation expresses the relative probability of entering the interaction and choosing k, across types t' and t'', taking the initial belief b into account; the equality with the left-hand side expresses that the belief that is formed upon observing k is consistent with the actual relative choice probabilities.

A special case covered by (5) is that some k ends up being never chosen. Formally, this case occurs if $c_k(t)b(t) = 0$ for all t. For such k, the conditions (5) are void because both sides are equal to 0. In all other k, we can find a type t'' such that $c_k(t'')b(t'') \neq 0$. Then the conditions (5) imply that

$b_k(t') = 0$ for any t' with $b(t')c_k(t') = 0$ and $b_k(t') > 0$ for any t' with $b(t')c_k(t') > 0$. Thus, in (6) either the numerators on both sides are strictly positive or the numerators on both sides are equal to zero, and the same is true for the denominators.

For any finite family of belief-payoff-vector pairs in K, let $\overline{K}$ denote the set of compositions of elements of K.

Assumption 1. *The feasibility set K is composition-closed, that is, $\overline{K} = K$.*

In all applications that we have described, the feasibility set K is composition-closed. Myerson (1983) calls this the inscrutability principle. Intuitively, the reason is as follows. For each k, by definition of the feasibility set, there exists a direct mechanism M_k that induces the payoff vector U_k at the belief b_k. Now consider the indirect mechanism $\hat{M}$ that gives the principal the option to select any of the direct mechanisms $M_1, \ldots, M_{\overline{k}}$ for play. The indirect mechanism then has an equilibrium in which any type t selects any M_k with probability $c_k(t)$, and in M_k the truth-telling and obedient equilibrium is played because the belief b_k prevails at the start of M_k.

2.2 Perfect-Bayesian equilibrium

In the previous section, we posited that the principal chooses among belief-payoff-vector pairs. That was a preparatory step towards conceptualizing the idea that the principal is a sender in a signaling game where the signals are mechanisms.

Which mechanisms should be allowed as signals for the principal in the mechanism-selection game? The standard literature on mechanism design where the principal has no private information evokes the revelation principle and so justifies the focus on direct revelation mechanisms. But now, for any belief b, any continuation equilibrium in any mechanism corresponds to a different direct mechanism, and the belief b is endogenous. The role of a mechanism as a signal generally depends on all its continuation equilibria for all possible beliefs. Here it can matter which equilibrium concept is used for continuation equilibria; this is particularly relevant if sequential mechanisms are allowed. Also, should one allow mechanisms such that a continuation equilibrium exists for some beliefs about the principal and not for others? It is not obvious how the set of possible mechanisms and continuation equilibria can be specified such that it does not appear restrictive and still a perfect

Bayesian equilibrium exists in the signaling game where the principal is the sender and signals are mechanisms.

In the literature, there are two different approaches to specifying the principal's set of mechanisms. First, Myerson (1983), given any Bayesian incentive problem, defines "generalized mechanisms" that allow for arbitrary finite message spaces for all players. The mechanism may reveal information about these messages privately to each player, to influence her private action, and some public outcome is implemented. The equilibrium concept for continuation equilibria in generalized mechanisms is Nash equilibrium.

The second approach, which can be traced back to Maskin and Tirole (1990, 1992), is to avoid the explicit specification of the set of feasible mechanisms, and instead restrict the continuation-equilibrium payoff properties of mechanisms. Given any (in whatever way specified) mechanism μ, we define a set of belief-payoff-vector pairs M^μ, as follows: $(b, U) \in M^\mu$ if and only if the payoff vector U can be induced by (in whatever way specified) continuation-equilibrium play of μ at the belief b. A set of belief-payoff-vectors $M \subseteq B \times \mathbb{R}^T$ is called a *Kakutani set*[10] if, for all $b \in B$, the set of payoff vectors $M(b) = \{U \mid (b, U) \in M\}$ is non-empty and convex, and the set M is compact (hence, the correspondence $b \mapsto M(b)$ is upper hemi-continuous). Rather than explicitly describing which μ's are feasible, Maskin and Tirole (1990, 1992) assume that only such μ are feasible where M^μ is a Kakutani set (whether *all* such μ are feasible will be irrelevant for our purposes).

Myerson's (1983) approach allows some mechanisms that are not allowed by Maskin and Tirole (1990) because, for some generalized mechanisms M and beliefs b, the set $M(b)$ is a non-singleton set of isolated points, hence non-convex. Maskin and Tirole's (1990, 1992) approach, on the other hand, is not restricted to Myerson's Bayesian incentive problems.

According to both approaches, in a perfect-Bayesian equilibrium (or, "expectational equilibrium", in Myerson's framework) of the mechanism-selection game, we can assume without loss of generality from the point of view of the principal's equilibrium payoff vector that all types of the principal pool at the same mechanism.

The possibility of pooling follows from composition-closedness. To see this, let b^* denote the (interior) prior belief about the principal.[11] Let

[10] The terminology is adapted from Pęski (2022).

[11] Myerson does not need to specify a prior belief as it is implicitly built into the definition

$\mu_1, \ldots, \mu_{\overline{k}}$ denote the list of mechanisms that are chosen with positive probability by at least one type of principal. In a perfect-Bayesian equilibrium, each type of principal chooses an optimal mechanism from the list, and the beliefs b_k are consistent with these choices in the sense of Bayes' rule. For any k, let U_k denote the payoff vector that is induced by the continuation-equilibrium play of μ_k. Let $c_k(t)$ denote the probability that any type t chooses the mechanism μ_k. By composition-closedness (with $b = b^*$), the payoff vector $U = \max_k U_k$ is b^*-feasible. Thus, without loss of generality all types pool at a mechanism that induces U at the belief b^*.

According to Myerson (1983), given any Bayesian incentive problem, in an *expectational equilibrium* all types of the principal pool at an incentive-compatible direct revelation mechanism such that for any generalized mechanism μ' there exists a belief b' and a continuation equilibrium in μ' at belief b' such that no type of principal gains from deviating to μ'.

The following alternative concept follows the spirit of Maskin and Tirole (1990, 1992). Given any informed-principal setting K and any interior prior belief b^*, a payoff vector U is a *Kakutani perfect-Bayesian equilibrium* if U is b^*-feasible, and for any Kakutani set $M' \subseteq K$ there exists $(b', U') \in M'$ such that U' is below U.

The following observation is often useful for fencing the set of Kakutani perfect-Bayesian equilibria; the lemma's conclusion is immediate from the assumption because $M' = B \times \{\underline{U}\}$ is a Kakutani set.

Lemma 1. *If there exists a payoff vector $\underline{U}$ that is feasible at all beliefs, then $\underline{U}$ is below all Kakutani Perfect-Bayesian equilibria.*

To illustrate the concept of Kakutani perfect-Bayesian equilibrium, consider again our Spence-job-market example. The least-cost separating payoff vector, $U^{lcs} = (\theta_L, \theta_L + (\theta_H - \theta_L)^2/\theta_H)$, illustrated as the 1st vertex in Figure 1, is b^*-feasible for any b^*. Any Kakutani set M leaves a trace $\cup_{b \in B} M(b)$ in the space of payoff vectors. By definition, the trace includes a 0-feasible payoff vector U', and from our earlier characterization of the 0-feasibility set in the Spence example it follows that $U'(t) \leq U^{lcs}(t)$ for all $t \in T$. Thus, using the "pessimistic belief" $b' = 0$ we see that U^{lcs} is a Kakutani perfect-Bayesian equilibrium. On the other hand, because U^{lcs} is feasible at all beliefs, Lemma 1 implies that any Kakutani perfect-Bayesian equilibrium U is above U^{lcs}. It

of a Bayesian incentive problem. Translated to our terminology, Myerson's prior is the uniform distribution on T.

follows that the set of Kakutani perfect-Bayesian equilibria is

$$\{U \mid U \text{ is } b^*\text{-feasible, } \forall t \in T: \; U(t) \geq U^{lcs}(t)\}.$$

For later use, we note the following.

Lemma 2. *In any informed-principal setting with any interior prior, the set of Kakutani Perfect-Bayesian equilibria is closed.*

Proof. Denote the feasibility set by K and the prior by b^*. Consider a sequence of equilibria $(U_n) \to U^*$. Then U_n is b^*-feasible for all n. Because the set of b^*-feasible payoff vectors, $K(b^*)$, is closed, it follows that U^* is b^*-feasible.

Consider any Kakutani set $M' \subseteq K$. There exists a sequence $(b'_n, U'_n) \in M'$ such that $U'_n(t) \leq U_n(t)$ for all n and t. Because M' is compact, there exists $(b', U') \in M'$ such that $(b'_n, U'_n) \to (b', U')$ along some subsequence. Thus, $U'(t) \leq U^*(t)$ for all t, proving that U^* is a Kakutani Perfect-Bayesian equilibrium. □

3 Neo-optimum

In many informed-principal settings (such as the Spence-job-market example above with b^* being sufficiently close to 1), multiple Kakutani perfect-Bayesian equilibria exist. Similarly, Myerson (1983) observes that multiple expectational equilibria exist in many Bayesian incentive problems.

The main goal of our paper is to show how to select perfect-Bayesian equilibria that are "sender preferred" in an intuitive sense. Importantly, our refinement will be invariant to scaling the utility of each sender type. Thus, the refinement is a-priori unrelated to ex-ante optimality for the sender. Our refinement provides a unified perspective of the informed-principal literature, and opens the door to solving new problems.

We start from a concept inspired by Farrell (1993). Consider a feasibility set K, an interior belief b, and a payoff-vector $U \in \mathbb{R}^{|T|}$. A belief-payoff-vector pair $(\hat{b}, \hat{U}) \in K$ is a *neologism* for (b, U)[12] if $\hat{U}(\check{t}) > U(\check{t})$ for some $\check{t} \in T$, and the following conditions hold for all $t \in T$:

$$\text{if } \hat{U}(t) > U(t) \text{ then } \hat{b}(t)b(t') \geq \hat{b}(t')b(t) \text{ for all } t' \in T, \tag{7}$$

$$\text{if } \hat{U}(t) < U(t) \text{ then } \hat{b}(t) = 0. \tag{8}$$

[12] Sometimes we say instead that $(\hat{b}, \hat{U})$ is a neologism for U at b.

Intuitively, a neologism that can be seen as a Bayes-consistent and feasible deviation relative to a given (not necessarily feasible) belief-payoff-vector pair. The deviating payoff vector $\hat{U}$ comes together with a belief that puts probability 0 on types who would be harmed by the deviation (see (8)), retains the relative likelihood across types that strictly gain (see (7) with switched roles of t and t', yielding $b_1(t)b(t') = b_1(t')b(t)$), and can shift belief probability mass from indifferent types to strictly gaining types (see (7) with t' such that $\hat{U}(t') = U(t')$).

A payoff vector U is *b-neologism-proof* if no neologism exists for (b, U). As implicitly suggested by the principal's "speeches" proposed in Myerson (1983), an ideal solution for the principal would be a payoff vector that is feasible at the prior belief b^*, and is b^*-neologism-proof. In many settings, however, such a payoff vector does not exist. Below we will review this well-known fact in our Spence example; see Farrell (1993) for a different example in an elementary signaling game.

In the spirit of the literature following Farrell (1993), one may respond to the non-existence problem by restricting the set of neologisms that are considered legitimate. We follow an alternative approach: we select the b^*-feasible payoff vectors that are *above limits of* b^*-neologism-proof payoff vectors. Note that the selected payoff vector itself may not be b^*-neologism-proof. Here is the definition.

Given any interior prior belief b^*, a payoff vector U is a *b^*-neo-optimum* if U is b^*-feasible and there exists a payoff vector $V \leq U$ such that V is a limit of b^*-neologism-proof payoff vectors.

Our first remark is that all neo-optima lie on the principal's weak Pareto frontier. This is immediate from the definition of neo-optimum.

Remark 1. *Let b^* denote an interior belief. Consider a payoff vector U such that $U(t) < \hat{U}(t)$ for all $t \in T$ for some b^*-feasible $\hat{U}$. Then U is not a b^*-neo-optimum.*

Another important remark is that the notion of neo-optimum is independent of each principal type's utility scale: if a positive affine transformation is applied to some type's utility, then the set of neo-optima remains unchanged. This reveals a fundamental difference to the notion of the principal's *ex-ante optimum*, which by definition is any payoff vector that maximizes $\sum_t b^*(t)U(t)$ among all b^*-feasible payoff vectors U.

In the rest of the paper, we will show that neo-optima are refinements of perfect-Bayesian equilibria, exist broadly, often lead to a unique prediction, and provide a unified perspective to the various solution concepts that have been proposed in the informed-principal literature.

In a given application, given any b^*, one can find the neo-optima via the following intuitive steps. First, characterize the b^*-neologism-proof payoff vectors (independently of their feasibility) by checking any possible neologism. Then consider the topological closure of this set and include all vectors above elements of this set. The intersection with the set of b^*-feasible payoff vectors is the set of neo-optima.

Examples

Consider the Spence setting with a prior belief $b^* > (\theta_H - \theta_L)/\theta_H$. This inequality guarantees that the high type strictly prefers the best pooling equilibrium,

$$U^{pool*} = (b^*\theta_H + (1-b^*)\theta_L, b^*\theta_H + (1-b^*)\theta_L),$$

over the least-cost separating payoff vector, U^{lcs}. Note that in this case there is a multiplicity of Kakutani perfect-Bayesian equilibria. From Remark 1 it is immediate that U^{pool*} is the unique neo-optimum. It is instructive to explicitly compute the set $P(b^*) \subseteq \mathbb{R}^T$ of b^*-neologism-proof payoff vectors. First, we show that only payoff vectors weakly above the least-cost separating one can be neologism-proof:

$$P(b^*) \subseteq \{U \mid U \geq U^{lcs}\}.$$

To see this, note that $(\hat{b}, U^{lcs}) \in K$ for all $\hat{b} \in B$. Thus, for any U such that $U(\theta_L) < U^{lcs}(\theta_L)$ and $U(\theta_H) \geq U^{lcs}(\theta_H)$, the belief-payoff vector pair $(0, U^{lcs})$ is a neologism for (b^*, U). Vice versa, for any U such that $U(\theta_L) \geq U^{lcs}(\theta_L)$ and $U(\theta_H) < U^{lcs}(\theta_H)$, the belief-payoff vector pair $(1, U^{lcs})$ is a neologism for (b^*, U). Lastly, for any U such that $U(\theta_L) < U^{lcs}(\theta_L)$ and $U(\theta_H) < U^{lcs}(\theta_H)$, the belief-payoff vector pair (b^*, U^{lcs}) is a neologism for (b^*, U).

Second, no payoff vector that is dominated by the best pooling one can be neologism-proof.

$$P(b^*) \cap \{U \mid U \leq U^{pool*},\ U \neq U^{pool*}\} = \emptyset.$$

This follows from using (b^*, U^{pool*}) as a neologism.

Third, consider the segment between the least-cost separating payoff vector and (θ_H, θ_H), the best pooling payoff vector at the belief 1. We define

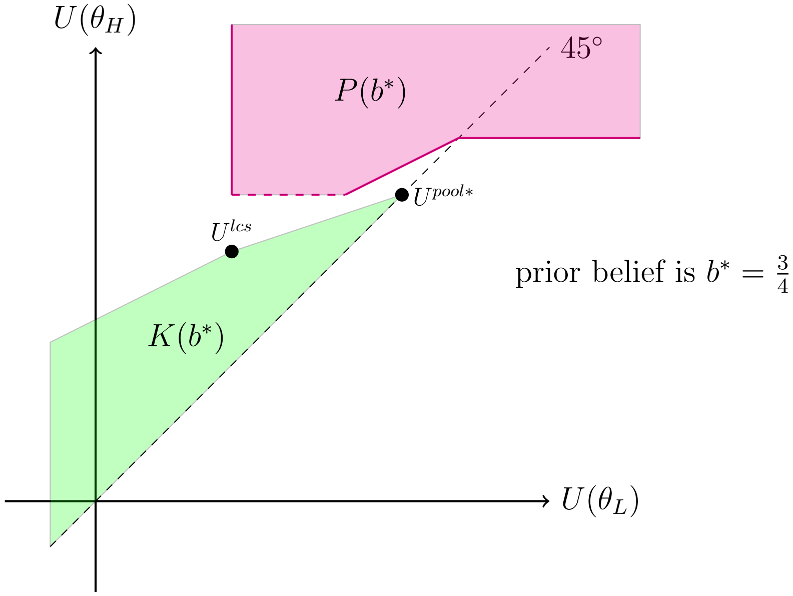

Figure 2: In the Spence-job-market example, at the prior belief $b^* = 3/4$, the set of feasible payoff vectors $K(b^*)$ has an empty intersection with $P(b^*)$, the set of payoff vectors U such that (b^*, U) is neologism-proof (the set extends infinitely to the upper right). The unique neo-optimum is U^{pool*}.

the "lower right" of this segment as the set of payoff vectors U such that $U(\theta_L) \geq \hat{U}(\theta_L)$ and $U(\theta_H) < \hat{U}(\theta_H)$ for some $\hat{U}$ in the segment. Given any such U, the pair $(1, \hat{U})$ is a neologism for (b^*, U). Thus,

$$P(b^*) \cap \big(\text{lower right of segment between } U^{lcs} \text{ and } (\theta_H, \theta_H)\big) = \emptyset.$$

The three restrictions that we have described characterize $P(b^*)$; the set is sketched in Figure 2.

It is apparent that no b^*-feasible and b^*-neologism payoff vector exists. It is also apparent that U^{pool*} is the unique payoff vector that is above a point in the topological closure of $P(b^*)$. Thus, U^{pool*} is the unique neo-optimum.

In the case $b^* = (\theta_H - \theta_L)/\theta_H$, the set of neo-optima equals the segment spanned by U^{lcs} and U^{pool*}.

In cases with a prior belief $b^* < (\theta_H - \theta_L)/\theta_H$, the payoff vector U^{lcs} is the unique Kakutani perfect-Bayesian equilibrium and thus is the unique neo-optimum; this is implied by Proposition 1 below.

Note that neo-optimum in general differs from ex-ante optimum. If the prior belief b^* is slightly below $(\theta_H - \theta_L)/\theta_H$, then the payoff vector U^{lcs} is the unique neo-optimum and the payoff vector U^{pool*} is the unique ex-ante optimum.[13]

A class of settings in which neo-optimum is trivially unique and identical to ex-ante optimum is characterized by the property that $U(t) = U(t')$ for all $(b, U) \in K$ and all $t, t' \in T$. Such settings are considered in Koessler and Skreta (2016):[14] different types of the principal have identical goals and represent different information about the agents' payoffs.

In Appendix B we present several cases of mechanism-design by a privately informed seller with interdependent values. These examples reveal some aspects that are not present in our Spence example. In particular, one example features b-feasibility sets that are not nested across different beliefs b, and the other example has multiple Kakutani perfect-Bayesian equilibria such that different types of the principal prefer different equilibria, yet there is a unique neo-optimum.

Our next result shows that neo-optimum is a refinement of perfect-Bayesian equilibrium.

Proposition 1. *Let b^* denote an interior belief. Any b^*-neo-optimum is a Kakutani Perfect-Bayesian equilibrium.*

Towards proving this, the main technical hurdle is Lemma 3 below. (This is also the essential step towards many of the equilibrium constructions in the informed-principal literature.) It says that from any given finite list of Kakutani sets belief-payoff-vector pairs can be selected to form a composition that entails any given belief.

Lemma 3. *Let $\mathcal{P}$ be a finite set of Kakutani sets. Let $b \in B$.*

Then there exists a composition $(b, U^*) \in B \times I\!R^T$ of some $(b^*_P, U^*_P)_{P \in \mathcal{P}}$, where $(b^*_P, U^*_P) \in P$ for all P.

The proof works as follows. For each $n = 1, 2, \ldots$, we define a function Ψ_n that continuously maps each list of payoff vectors that can occur in the signals in $\mathcal{P}$ to a vector of smoothed optimal choice probabilities, that is,

[13] Alternatively, we may maximize the principal's ex-ante expected payoff among all Kakutani perfect-Bayesian equilibria. This "ex-ante-optimal equilibrium" is identical to neo-optimum in the Spence example for all $b^* \neq (\theta_H - \theta_L)/\theta_H$. For an example where ex-ante-optimal equilibrium and neo-optimum are different for an open set of prior beliefs, consider the discussion at the end of the informed-seller example in the appendix.

[14] See also Izmalkov and Balestrieri (2012).

every type chooses every $P \in \mathcal{P}$ with a positive probability, and as n tends to infinity, almost all weight is put on utility-maximizing Ps. Requiring strictly positive choice probabilities is a form of trembling which guarantees that by Bayesian updating from b a unique belief about the sender's type is assigned to each signal.

We have another correspondence given by the signals themselves. Suppose a belief is given for each signal in $\mathcal{P}$. We consider the correspondence that assigns to any such list of beliefs a set of lists of payoff vectors by applying the signals to the beliefs. Combining this correspondence with Ψ_n, we obtain a correspondence that has a fixed point by Kakutani's Theorem.[15]

A fixed point consists of, for each signal, a belief and a payoff vector that belongs to the signal at this belief such that the beliefs are consistent with the sender's smoothed optimal choice.

By taking n to infinity we consider a sequence of fixed points with trembling probabilities tending to 0. By choosing an appropriate subsequence, we can guarantee that the sequence converges. In the limit, there is no trembling restriction so that the beliefs are fully consistent with the sender's optimal choice among the signals. Because the signals are compact, each limit payoff vector belongs to the respective signal at the limit belief. By construction, the maximum of the limit payoff vectors together with belief b is the composition of the list of limit payoff vectors together with the limit beliefs. The details of the proof are in the Appendix.

Proof of Proposition 1. Consider a b^*-feasible and b^*-neologism-proof payoff vector U. It is sufficient to show that U is a Kakutani perfect-Bayesian equilibrium. By Lemma 2, the conclusion then extends to limit points. By definition of equilibrium, it then extends to points above.

Consider any Kakutani set M in K. By Lemma 3, there exists a composition $(b^*, \cdot)$ of some $(b_1, V) \in M$ and some element of $B \times \{U\}$.

In particular, $(b_1, V) \in K$. Let $c_1(t)$ denote the choice probabilities of M in the composition.

Suppose that $V \not\leq U$. For any $t \in T$ with $V(t) > U(t)$, we have $c_1(t) = 1$ by definition of a composition. Applying (5) with $t'' = t$, (7) follows.

By assumption, $V(\hat{t}) > U(\hat{t})$ for some $\hat{t} \in T = \operatorname{supp}(b^*)$, implying $c_1(\hat{t}) = 1$.

[15] The detour through introducing trembles is needed to guarantee that Ψ_n is single-valued, and thus the combined correspondence is convex-valued, a prerequisite for Kakutani's Theorem.

Now consider any $t \in T$ with $V(t) < U(t)$. Then $c_1(t) = 0$. Because b is interior, $b(\hat{t}) > 0$. Applying (5) with $t' = t$ and $t'' = \hat{t}$ now yields (8).

Thus, (b_1, V) is a neologism for (b^*, U). But this contradicts the assumption on U. Hence, $V \leq U$, as was to be shown. □

As an immediate corollary to Proposition 1, we have a way of establishing other (and, in many settings, all) Kakutani perfect-Bayesian equilibria. Virtually all equilibria in the informed-principal literature can be reconstructed in this way.

Corollary 1. *Consider any interior prior belief b^* and another interior belief b'. Any payoff vector that is b^*-feasible and is a b'-neo-optimum is a Kakutani perfect-Bayesian equilibrium at the prior belief b^*.*

Note that the result applies to the equilibria in Maskin and Tirole (1992) (see also the correcting formulation in Dosis (2022)) and the literature building on it: if the "Rothschild-Stiglitz-Wilson allocation" (which is feasible at all beliefs) is undominated for some interior belief b', then its induced payoff vector is b'-neologism-proof and hence is a b'-neo-optimum, implying by Corollary 1 that it is a Kakutani perfect-Bayesian equilibrium at any interior prior belief b^*.

4 Neo-optimum versus Myerson's solution concepts in Bayesian incentive problems

In this section, we use the simpler term neo-optimum instead of b^*-neo-optimum because we follow Myerson's formulation of Bayesian incentive problems in which the prior belief b^* is the uniform distribution on T. We show that a neo-optimum exists in any Bayesian incentive problem and explain its connection to Myerson's solution concepts neutral optimum, expectational equilibrium, and core.

Given any Bayesian incentive problem, Myerson (1983) identifies a set of payoff vectors that he calls *neutral optima*. He argues that neutral optima represent a "fair" compromise across all types of the principal. By verifying Myerson's axioms in the proof below, we obtain the following reult.

Proposition 2. *Consider any Bayesian incentive problem. Any neutral optimum is a neo-optimum.*

An immediate conclusion is that in the Spence example above with a highest feasible task level, the neutral optimum is generically unique and identical to neo-optimum for all prior beliefs $b^* \neq (\theta_H - \theta_L)/\theta_H$.

Myerson shows that a neutral optimum always exists. Thus, we have the following.

Corollary 2. *In any Bayesian incentive problem, a neo-optimum exists.*

This is a very broad existence result because Myerson allows for arbitrary outcome spaces, any number of agents, arbitrary payoff functions, and moral hazard; the essential restriction is that type and outcome spaces are assumed to be finite.

We emphasize that neo-optimum is not only much easier to handle than neutral optimum in applications, but neo-optimum is also conceptually simpler than neutral optimum because it avoids any reference to Myerson's (1983) Extension axiom, which relates properties of solutions across different Bayesian incentive problems.

While we will not dwell on Myerson's axioms, we still find it worthwhile to mention that the axioms are very useful for understanding why a prior-belief-feasible and neologism-proof payoff vector often fails to exist. The problem with neologism-proofness as a solution concept is that it violates two axioms, Openness and Domination. Indeed, as the Spence example above shows, a payoff vector can be neologism-proof while payoff vectors arbitrarily close to it may fail to be, and some payoff vectors above it may also fail to be. The concept of a neo-optimum relaxes the ideal of a prior-belief-feasible and neologism-proof payoff vector just enough so that all of Myerson's axioms are satisfied, thus restoring existence.

Proof of Proposition 2. Let b^* denote the uniform distribution on T. We say that a payoff vector $U \in \mathbb{R}^T$ is *neo-blocked* if for some $\epsilon > 0$, a neologism exists for all (b^*, V) such that $V \leq U + \epsilon$. To complete the proof, it is sufficient to show that the concept of neo-blocking satisfies Myerson's four axioms.

The axioms Extension, Domination, and Openness are clear by construction. Consider the axiom Strong Solution. Let U be a strong solution. We have to show that U is not neo-blocked. For this it is sufficient that no neologism exists for (b^*, U). Suppose that $(\hat{b}, \hat{U})$ is a neologism. The main step is to show that (*) there exists a belief b' such that $(b^*, \max\{\hat{U}, U\})$ is a composition of $(\hat{b}, \hat{U})$ and (b', U).

By definition of a strong solution, (b', U) is feasible. Given (*), composition-closedness then implies that $(b^*, \max\{\hat{U}, U\})$ is feasible. Because, by definition of a strong solution, U is undominated, $\max\{\hat{U}, U\} = U$, implying that $\hat{U} \leq U$, contradicting the fact that $(\hat{b}, \hat{U})$ is a neologism.

To show (*), take any $\check{t}$ such that $\hat{U}(\check{t}) > U(\check{t})$ and define $\hat{d} = b^*(\check{t})/\hat{b}(\check{t})$. For all $t \in T$, define

$$c_1(t) = \frac{\hat{d}\;\hat{b}(t)}{b^*(t)}, \quad c_2(t) = 1 - c_1(t), \quad b'(t) = \frac{b^*(t)c_2(t)}{1 - \hat{d}}.$$

This together with the definition of a neologism implies $c_1(t) = 1$ for all t with $\hat{U}(t) > U(t)$, $0 \leq c_1(t) \leq 1$ for all t with $\hat{U}(t) = U(t)$, and $c_1(t) = 0$ for all t with $\hat{U}(t) < U(t)$. Next,

$$\sum_t b^*(t)c_2(t) = 1 - \sum_t b^*(t)c_1(t) = 1 - \sum_t \hat{d}\;\hat{b}(t) = 1 - \hat{d},$$

implying that b' is a probability distribution. Moreover, using the definitions above it is straightforward to verify that

$$c_1(t'')b^*(t'')\hat{b}(t') = c_1(t')b^*(t')\hat{b}(t'')$$

and

$$c_2(t'')b^*(t'')b'(t') = c_2(t')b^*(t')b'(t'')$$

for all $t', t'' \in T$. This completes the proof of (*). □

The following result establishes that neo-optimum is an equilibrium refinement.

Proposition 3. *In any Bayesian incentive problem, any neo-optimum is an expectational equilibrium.*

Proof. Consider a payoff vector U such that no neologism exists for (b^*, U), where b^* is the uniform distribution. Consider any generalized mechanism μ' as defined in Myerson (1983). It is sufficient to show that there exists a belief b' and a continuation equilibrium in μ' at belief b' such that no type of principal gains from deviating to μ'.

Consider a fictitious game in which the principal (with type distributed according to b^*) first chooses between getting the payoff vector U and the game ends, or deciding that μ' will be played. Because the fictitious game is finite, there exists a sequential equilibrium. Let b' denote a belief at the start of the continuation game μ' that is consistent with the sequential equilibrium. For all t, denote by $c_1(t)$ the probability that type t decides to play μ', and denote by $V(t)$ her expected payoff in the continuation game μ'.

By definition of a sequential equilibrium, V is b'-feasible.

Suppose that $V \nleq U$. By the sequential-equilibrium conditions, (b', V) is a neologism for (b^*, U). But this contradicts the assumption on U. Hence, $V \leq U$, so that at belief b' no type of principal gains from choosing μ'. □

To relate neo-optimum to Myerson's (1983) other solution concept, core, we need additional notation. We formulate the relation generally for all informed-principal settings, not restricted to Bayesian incentive problems. Given any interior belief b^* and a non-empty set $S \subseteq T$, let b^S denote the belief derived from the information that the type belongs to S, that is,

$$\begin{aligned} & b^S(t) = 0 \quad \text{for all } t \notin S, \\ \text{and} \quad & b^S(t')b^*(t'') = b^S(t'')b^*(t') \quad \text{for all } t', t'' \in S. \end{aligned}$$

A payoff vector U is called a b^*-*core* payoff vector if U is b^*-feasible and for any payoff vector V that dominates U there exists $S \supseteq \{t \in T \mid V(t) > U(t)\}$ such that V is not b^S-feasible. Myerson motivates the concept with ideas involving neologisms. While a core payoff vector always exists if b^* is the uniform distribution, it is not always an equilibrium. Next we show that neo-optimum is stronger than the core.

Proposition 4. *Let b^* denote an interior belief. Then any b^*-neo-optimum is a b^*-core payoff vector.*

Proof. Denote $D = \{t \in T \mid V(t) > U(t)\}$.

Suppose that U is b^*-feasible, but is not a core payoff vector, that is, there exists a a payoff vector V that dominates U and V is b^S-feasible for all $S \supseteq D$.

Fix $\epsilon > 0$ such that $V(t) > U(t) + \epsilon$ for all $t \in D$.

Consider any $W \leq U + \epsilon$.

Note that $D \subseteq D^W := \{t \in T | V(t) > W(t)\}$. Thus, V is b^{D^W}-feasible, implying that (b^{D^W}, V) is a neologism for (b^*, W).

We conclude that U is not a b^*-neo-optimum. □

5 Neo-optimum in private-values environments

Mylovanov and Tröger (2012) establish a solution for a principal who "has private information that is not directly payoff relevant to the agents, but may influence her design"—the private-values case. The concept, strongly neologism-proof allocations, is a generalization of strong unconstrained Pareto optimum (SUPO) defined by Maskin and Tirole (1990). Here we show it is equivalent to neo-optimum.

Proposition 5. *Consider a separable generalized-private-values environment in the sense of Mylovanov and Tröger (2012). Consider any interior prior b^*. Then a payoff vector U is strongly neologism-proof if and only if U is a b^*-neo-optimum.*

Together with Proposition 2 this resolves a long-standing open question concerning the relation between the private-values solution concepts and neutral optimum: in any separable generalized-private-values environment that is also a Bayesian incentive problem, any neutral optimum is strongly neologism-proof. Since strong neologism-proofness often yields sharp properties related to competitive equilibria (Maskin and Tirole, 1990)—such as ex-ante optimality in quasi-linear settings (Mylovanov and Tröger, 2014)—the same properties apply to any neutral optimum.

From the existence result in Mylovanov and Tröger (2012) together with the only-if part of Proposition 5, we can also conclude that a neo-optimum broadly exists in private-value settings, including settings that do not satisfy the finiteness properties of Bayesian incentive problems as defined by Myerson (1983).

As an intermediate step towards proving Proposition 5, we employ yet another solution concept. In a sense, this is the missing piece that allows to connect private-values settings to neutral optimum. The concept was invented by Koessler and Skreta (2023) in a non-private-values context of information design. Given any belief b^*, a payoff vector U is b^*-*interim-optimal* if (i) U is b^*-feasible and (ii) there does not exist a belief b together with a b-feasible payoff vector V such that $\text{supp}(b) \subseteq \{t \in T | V(t) > U(t)\}$. Interim-optimality is easily seen to be at least as strong as neo-optimum (and the result has nothing to do with private values):

Remark 2. *Consider any informed-principal setting. Given any interior belief b^*, any b^*-interim-optimal payoff vector is a b^*-neo-optimum.*

Proof. Consider any b^*-interim optimal payoff vector U. Then, for all $\epsilon > 0$, no neologism exists for $(b^*, U + \epsilon)$. Thus, U is a limit of b^*-neologism-proof payoff vectors, showing that it is a neo-optimum. □

Interim-optimality—in contrast to neo-optimum—is not a generally applicable solution concept because existence may fail, as can be seen in our Spence example with $b^* > (\theta_H - \theta_L)/\theta_H$.

To prove Proposition 5, we first show that in private-values settings, interim-optimality and neo-optimality are in fact equivalent. Then we use the separability assumption in Mylovanov and Tröger (2012) to show that interim-optimality and strong neologism-proofness are equivalent.

Translated into our current, abstract framework, a *private-values setting* is a set of belief-payoff-vector pairs K with the following property. There exist closed and convex sets of P and $Q \supseteq P$ in some linear space, and a linear mapping Π from Q into $\mathbb{R}^T$. Each element of Q is a possible "allocation for a principal type".

Given any allocation family $(\rho_t)_{t \in T}$ with $\rho_t \in Q$, we say that the condition "Principal's Incentive Compatibility" (PIC) is satisfied if $\Pi(\rho_t)(t) \geq \Pi(\rho_{t'})(t)$ for all $t, t' \in T$. Given any allocation family $(\rho_t)_{t \in T}$ in Q together with a belief $b \in B$, we say that the condition "Agents' Feasibility" (AF) is satisfied if $\sum_{t \in T} b(t)\rho_t \in P$.

The feasible set K is given as follows: for any belief $b \in B$, we have $U \in K(b)$ if and only if there exists an allocation family $(\rho_t)_{t \in T}$ in Q such that $U(t) = \Pi(\rho_t)(t)$ and PIC holds, and AF holds for $(\rho_t)_{t \in T}$ together with the belief b.

This captures as a special case the "generalized private-values environments" of Mylovanov and Tröger (2012), where Q is the set of all maps from the profile of agent types (excluding the principal's type) into the space of outcomes, P is the subset of Q in which the agents' incentive and participation constraints are satisfied, and $\Pi(x)$ is the principal's expected-payoff vector from any $x \in Q$. What we call an "allocation family" here is called an "allocation" in Mylovanov and Tröger (2012).

Proposition 6. *Consider any private-values setting. Given any interior belief b^*, a payoff vector is b^*-interim-optimal if and only if it is a b^*-neo-optimum.*

The only-if part was shown in Remark 2. To show the if-part (for details see the Appendix), we start with a b^*-feasible payoff vector U that is not interim optimal and show that it is not a neo-optimum.

By assumption, there exists a feasible "deviation" (b'', U'') such that in U'' all types in the support of b'' are strictly better off than in U. In general, (b'', U'') is not a neologism because the relative probabilities of different types in the support of b'' are unrestricted, and types outside the support may also be better off in U'' than in U. The idea behind our proof is to apply a sequence a "surgeries" in which we change the deviation multiple times such that eventually a neologism $(\hat{b}, \hat{U})$ for (b^*, U) is obtained.

Because in U'' all types in the support of b'' are also strictly better off than in the payoff vectors in a neighborhood of U and below, the surgery constructions extend to the existence of neologisms for all such payoff vectors, implying that U is not a neo-optimum.

Starting with b'' and an allocation family (ρ''_t) for U'', the basic idea behind our surgeries is that we build a new belief b' together with a new allocation family (ρ'_t) such that

$$\sum_{t \in T} b'(t)\rho'_t = \sum_{t \in T} b''(t)\rho''_t,$$

where each type's ρ'-allocation will be a convex combination of various types' ρ''-allocations. By construction, the new allocations belong to Q, and AF remains true for (ρ'_t) (resp., for its resulting payoff vector) together with b'.

If some type t's new allocation ρ'_t arises from a convex combination involving some type $\check{t}$'s old allocation $\rho''_{\check{t}}$, then we say that type t obtains a chunk of type $\check{t}$'s allocation. Note that in this process a corresponding piece of probability mass from $b''(\check{t})$ must be moved into $b'(t)$ so that AF remains true.

The possibility of surgeries yields considerable freedom to construct new deviations (b', U'), but care is needed to guarantee that PIC remains true so that (b', U') is feasible. Several observations are helpful towards verifying PIC: first, if a type does not gain from choosing some other types' allocations, then she also cannot gain from any convex combination of these allocations; second, if a type does not gain from choosing another type's allocation, then any convex combination of her own and that type's allocation is still at least as good for her as that type's allocation; third, if a type strictly loses from choosing another type's allocation, then this remains true for any perturbation of her original allocation.

As a first surgery, the belief is kept fixed and each type outside the support of b'' gets restricted to choose her most preferred allocation among the allocations of types in the support of b''. This will keep PIC in place and can only lead to a reduction of utility for the types outside the support.

If after this operation there exists a type t outside the support who still obtains more than her U utility, then, as a second surgery, we move some probability mass to her from her most preferred type in the initial support. AF and PIC are still in place, but now we have included t into the support. In this way, we obtain a deviation (b', U') such that in U' all types in the support of b' are strictly better off than in U, and all types outside the support are weakly better off in U than in U'.

If the support of b' contains a single type, we have obtained a neologism and are done. If it contains two types, say t and $\check{t}$, the remaining surgeries are still comparatively easy. It is useful to introduce auxiliary variables that capture probabilities relative to the prior; we call the numbers $b'(t)/b^*(t)$ and $b'(\check{t})/b^*(\check{t})$ the r-values of the types t and, resp., $\check{t}$ at the belief b'. If both types have the same r-value, then (b', U') is a neologism for (b^*, U) and we are done.

Otherwise one type, say t, has a smaller r value than the other type, $\check{t}$. Now imagine that we change the deviation continuously, by moving an ever larger chunk of the allocation of type $\check{t}$, and a corresponding piece of belief probability mass, to type t. Along the way, any other type (i.e., the types outside the support of b') always chooses her most preferred allocation among the current allocations of the types t and $\check{t}$. In this process, the r-value of type t increases while the r-value of type $\check{t}$ decreases, and the utility of type t_1 can drop. AF und PIC remain intact.

This process is continued until one of two things happens. Either both types' r values are equalized, or the utility of type t drops to her U utility. In both cases we have arrived at a neologism and are done.

The general argument, where the type space (and thus the support of b') can have any cardinality, is very much more complicated. The main reasons for the complications are that the number of incentive constraints in PIC increases fast (quadratically) with the cardinality of the type space, and that we have to find a deviation that equalizes the r-values across a potentially large number of types. These complications may have contributed to the fact that the underlying puzzle—the relation between neutral optimum and private-values solution concepts—has remained open essentially since the start of the informed-principal literature in the 1980s. In the following we provide a roadmap through the general argument.

The key to the general argument is the introduction of a special class of deviations. A feasible pair (b', U') is a *deviation if at least one type has utility $> U$, all types not in $\text{supp}(b')$ have utility $\leq U$ and each of them obtains the same allocation as one of the types in $\text{supp}(b')$, all types in $\text{supp}(b')$ have utility $\geq U$, and the U-utility types $t \in \text{supp}(b')$ have r-values $\leq r^*$, where r^* is defined as the "target value" of r that would be reached if all r-values of types with $> U$ utility were equal, that is

$$\sum_{U'(t) > U(t)} \left(r^* - r_{b'}(t)\right) b^*(t) = 0, \tag{9}$$

where $r_{b'}(t) = b'(t)/b^*(t)$ denotes the r-value of any type t at the belief b'. (Note that r^* is defined separately for each *deviation.)

Not all *deviations are allowed deviations in the definition of interim-optimality because some types in the support of b' can have utility equal to U. However, a *deviation, with no U-utility type in the support of b', exists by the first and second surgery arguments above.

If a *deviation is such that the r-values of all $> U$-utility types are equalized then, by construction, the *deviation is a neologism for (b^*, U) and we are done.

Rather than explicitly describing the sequence of surgeries to be applied to the initial *deviation, we cut through to the end by considering a *deviation with the "right" properties.

Consider the *deviations that have a minimal cardinality of the support of b' among all *deviations. Among these, consider the *deviations that have a maximum number of U-utility types in the support. Among these, we consider a *deviation that has a maximum number of types with r-value equal to r^*.

We claim that any such *deviation (b', U') has the desired neologism properties. Suppose otherwise.

Let r_0^* denote the value of r^* for (b', U'). Then there exists a type t_1 in the support of b' with $> U$ utility and an r-value below r_0^*. Let $T^{\leq}$ denote the set of $> U$-utility types with r-values $\leq r_0^*$. Let $T^{>}$ denote the $> U$-utility types with r-values $> r_0^*$.

Starting with (b', U'), we now consider the problem of maximizing the r-value of type t_1 via surgery subject to constraints. We consider surgery that concerns the types in $T^{\leq} \cup T^{>}$, while the other types in the support of b' keep their allocations, and each type outside the support of b' chooses her best available allocation among those of the types in the support of b'.

Using the numeration from the proof for reference, the constraints are that (18) all types in $T^{>}$ keep their allocations, (19) each type in $T^{\leq}$ obtains a convex combination of the allocations of the types in $T^{\leq} \cup T^{>}$, (20) the r-value of type t_1 remains $\leq r_0^*$, (21) the r-values of the types in $T^{\leq} \setminus \{t_1\}$ remain the same as at the belief b', (22) the r-values of the types $T^{>}$ remain $\geq r_0^*$, (23) each type in $T^{\leq}$ weakly prefers her new allocation to the (old and new) allocation of each of the types in $\text{supp}(b') \setminus (T^{\leq} \cup T^{>})$, (24) each type in $T^{\leq}$ weakly prefers her new allocation to the new allocations of the types in $T^{\leq}$, and to the allocations of the types in $T^{>}$, and (25) the utility each type in $T^{\leq}$ does not fall below her U-utility.

We will show that at a solution to the maximization problem, denoted $(\hat{b}, \hat{u})$, the constraints (20), (22), (23), and (25) are not binding. This will allow us to increase the solution value via a perturbation that satisfies all constraints, and thus obtain a contradiction.

By construction, the solution $(\hat{b}, \hat{u})$ is a *deviation, where $\hat{b}$ has the same (minimum cardinality) support as b'. By the assumed maximality of the number of U-utility types in the support, at the optimum $(\hat{b}, \hat{u})$, the utility of no type in $T^{\leq}$ has dropped to her U-utility, that is, the constraints (25) are not binding. Note also that the r^*-value for $(\hat{b}, \hat{u})$ is still equal to r_0^*.

At $\hat{b}$, the r-value of type t_1 must still be strictly below r_0^*, and the r-values of the types in $T^{>}$ must still be strictly above r_0^* because the number of types with r-values equal to r^* was assumed to be already maximal at (b', U'), and by constraint (21) any type who before the optimization had an r value equal to r^* keeps it. Thus, the constraints (20) and (22) are not binding.

Suppose a constraint (23) is binding, that is, some type $t_i \in T^{\leq}$ is indifferent to a type $\mathring{t}$ that belongs to the support of $\hat{b}$ and who obtains her U utility. Then we can do a surgery where all the probability mass and allocation of $\mathring{t}$ is moved to type t_i, yielding a new *deviation where the type $\mathring{t}$ does not belong to the belief support anymore, but this contradicts the minimality of the support of $\hat{b}$ among all *deviations.

Now we describe the perturbation of $(\hat{b}, \hat{U})$. Only the allocations of the types in a subset of $T^{\leq}$ are changed. In the subset we include all types with allocations that type t_1 likes as well as her own allocation, and then include all types that any type included in the first round is indifferent to, and so on, until all indifferences in $T^{\leq}$ are exhausted. We denote this subset (which can be the singleton $\{t_1\}$) by $T^{\leq}_{=}$.

As a perturbing surgery, the allocation of each type in $T^{\leq}_{=}$ is now changed such that a small fraction of her new allocation comes from her respective most preferred type in $T^{>}$. The fraction will be the same for all types in $T^{\leq}_{=}$, implying that incentive compatibility relative to each other and to the types in $T^{>}$ remains intact. By construction, there are no indifferences from types in $T^{\leq}_{=}$ to types in $T^{\leq} \setminus T^{\leq}_{=}$ if the perturbation is small.

Due to the new allocation chunks and corresponding probability masses, the types in $T^{\leq}_{=}$ will now have increased r-values. For all types except t_1, the r values must be brought back to their previous levels to satisfy constraint (21).

To this end, we consider a directed graph with nodes $T^{\leq}_{=}$ where each edge corresponds to an indifference. We select a tree with root t_1 in $T^{\leq}_{=}$. All the direct predecessor types of the tree's end nodes get chunks of the end node's allocations and corresponding probability masses such that the end nodes are back to their correct r values. These corrections are iterated backwards through the tree. Due to the indifferences along the way, the involved types keep their utility levels. Eventually only type t_1 gains probability mass, yielding the desired contradiction.

To prove Proposition 5 (for details see the appendix), it is—in light of Proposition 6—sufficient to show that a payoff vector is strongly neologism-proof if and only if it is interim optimal.

The direction "only if" is immediate from the definitions. To show "if", consider an interim-optimal payoff vector U and suppose it is not strongly neologism-proof. Then there exists a feasible deviation pair (q_0, U') such that, for all $t \in \operatorname{supp}(q_0)$, we have that (i) $U'(t) \geq U(t)$ with strict inequality for at least one type $t = t'_0$, and (ii) $U(t)$ is below the "maximum feasible payoff" as defined in Mylovanov and Tröger (2012). We now do surgery in order to find a deviation as required in the definition of interim-optimality. By the "separability" assumption, there exists an allocation such that all agents' constraints are satisfied strictly; we perturb the allocation family underlying U' by having type t'_0 offer this separating allocation with a small probability and simultaneously slightly increasing the probability mass for type t'_0. She will still be strictly better off than in U. Given the new allocation family, the agents' constraints are satisfied strictly. Thus, we can again perturb it without violating the agents' constraints; we do this by giving all types in $\operatorname{supp}(q_0) \setminus \{t'_0\}$ their "maximum feasible payoff" with a small probability. Now all types in the belief support are strictly better off than in U, but PIC may not hold anymore. It can be restored by further surgery, using the method from Mylovanov and Tröger (2012). If one type in the belief support is attracted to the allocation of another type in the belief support, then we let the first type offer the average allocation of what both types used to offer, and move all the probability mass from the second type to the first type. This procedure continues until incentive compatibility is satisfied for the types in the (remaining) support. Let the types outside the support choose their optimum among the allocations of the types in the support. Then we have a deviation as considered in the definition interim-optimality.

6 Neo-optimum in settings with certification or information-design

Koessler and Skreta (2019) consider a seller-principal who proposes a mechanism to guide her interaction with a single buyer. Values are interdependent: each trader has private information concerning both traders' valuations. The seller can provide partial or complete evidence about (i.e., "certify") her type. The model fits into our framework, with Lemma 1 in Koessler and Skreta (2019) describing the sets of feasible payoff vectors for all beliefs.

Koessler and Skreta (2019) propose a solution concept, strong Pareto optimum (SPO). They show that any prior-feasible SPO is an expectational equilibrium and is ex-ante optimal, but it exists only in settings with sufficient (e.g., full) certifiability. Using neo-optimum instead of SPO as a solution concept in their setting, existence is generally guaranteed and their qualitative results remain largely intact. In the following we sketch how.

Remark 3. *Let b^* denote an interior prior belief. Any b^*-feasible SPO as defined in Koessler and Skreta (2019) is a b^*-neo-optimum.*

Proof. Consider a b^*-feasible SPO (payoff/profit vector) V^*. We show that for any $\epsilon > 0$, the payoff vector $V^*+\epsilon$ is neologism-proof. Suppose there exists a neologism (b, U) for $(b^*, V^* + \epsilon)$. Define a payoff vector V via $V(t) = U(t)$ for all $t \in \text{supp}(b)$, and $V(t) = V^*(t)$ for all other t. Then (using the terminology of Koessler and Skreta (2019)) V is "buyer-feasible" at the belief $\pi = b$. Moreover, $V(t) \geq V^*(t)$ for all $t \in T$, with strict inequality for all $t \in \text{supp}(b)$, contradicting the definition of an SPO payoff vector. We conclude that V^* is a limit of b^*-neologism-proof payoff vectors, and hence is a b^*-neo-optimum. □

Thus, all the properties that we have established for neo-optima also hold for any prior-feasible SPO. Moreover, in all cases where there is a unique neo-optimum, it is identical to prior-feasible SPO if the latter exists.

Note that Koessler and Skreta (2019) includes two extreme cases, full certifiability and no certifiability ("soft information"). Importantly, SPO is a generally useful solution concept *only* under sufficient certifiability, while neo-optimum can always be used.

Remark 4. *Let b^* denote an interior prior belief. A b^*-feasible SPO as defined in Koessler and Skreta (2019) may not exist with soft information, but a b^*-neo-optimum always exists.*

Koessler and Skreta (2019) themselves remark on the non-existence problem. Interestingly, non-existence can also be seen from our Spence-job-market example, restricted via a highest feasible task level $\overline{e}$. This setting is included as a soft-information case in Koessler and Skreta (2019). The worker-principal is the seller. Selling her labor with a certain probability p means working at the task level $e = p\overline{e}$. Thus, our Spence example with the prior b^* slightly below $(\theta_H - \theta_L)/\theta_H$ yields instances where no b^*-feasible SPO exists because it would be ex-ante optimal by the results in Koessler and Skreta (2019) and would be a b^*-neo-optimum by our Remark 3.

Neo-optimum, however, always exists: any setting considered in Koessler and Skreta (2019) is a Bayesian incentive problem, except that feasibility is defined without truthtelling constraints for the principal; as observed in Koessler and Skreta (2023), Myerson's 1983 proof that a neutral optimum exists still applies and, by the same logic as in our Proposition 2, any neutral optimum is a neo-optimum.

In the other extreme case, full certifiability, the main qualitative insight of Koessler and Skreta (2019) is that a prior-feasible SPO exists and is ex-ante optimal. But this is also true for neo-optimum.

Remark 5. *With full certifiability, any neo-optimum is ex-ante optimal.*

Indeed, in the proof of Proposition 3 in Koessler and Skreta (2019), the pair $(\pi^0, \tilde{V})$ is a neologism for $(\pi^0, \hat{V})$ because with full certifiability there are no incentive constraints for the seller.

Koessler and Skreta (2023) analyze information design by a privately informed designer. Their setting fits into our framework.[16] Koessler and

[16] Note a special feature of information-design settings: different types of the principal

Skreta (2023) introduce a new solution concept, interim optimality, which we have generalized to arbitrary informed-principal settings in Section 5 above. The main results in Koessler and Skreta (2023) concern existence of interim-optimum in their setting, the proof that interim-optima are perfect-Bayesian equilibria, and characterization results in special settings.

From Remark 2 in Section 5 we know that any interim optimum is a neo-optimum. In the different variants of the introductory prosecutor-judge example in Koessler and Skreta (2023) the reverse is also true, that is, any neo-optimum is interim-optimal. The question to what extent this reverse implication holds in general information-design settings is left for future research.

Consider the three-actions variation of the prosecutor-judge setting in Koessler and Skreta (2023) with a belief that the defendant is guilty with a probability $b < 1/3$. A payoff vector $(U(t_G), U(t_I))$ is b-feasible if and only if

$$U(t) = 2\mu(a_2|t) + 3\mu(a_3|t) \quad \text{for } t \in \{t_G, t_I\},$$

where μ is a mechanism that satisfies the relevant obedience constraints, that is, μ is b-incentive compatible as defined in Koessler and Skreta (2023).

To prepare, we extend the arguments in Koessler and Skreta (2023) to show the following.

Lemma 4. *Consider a belief* $0 < b < 1/3$ *in the three-actions variation of the prosecutor-judge setting in Koessler and Skreta (2023). The set of b-feasible payoff vectors is the convex hull of the points* $(0, 0)$*,* $(3, 0)$*,* $(3, \frac{3}{2}\frac{b}{1-b})$*, and* $(2, 4\frac{b}{1-b})$*.*

See Figure 5 for an illustration of the b-feasibility set and the b'-feasibility set for some $0 < b < b' < 1/3$.

Koessler and Skreta (2023) show that, in the three-action and four-action prosecutor-judge examples with a prior belief $b^* < 1/3$, a payoff vector U^* is interim-optimal if and only if

(*) U^* is b^*-feasible and $U^*(t_G) = 3$.

We claim that condition (*) is also necessary for U^* to be a b^*-neo-optimum.

cannot imitate each other in a given direct mechanism. Thus, even in settings with state-independent preferences the feasibility of a payoff vector does not exclude the possibility that different types obtain different payoffs.

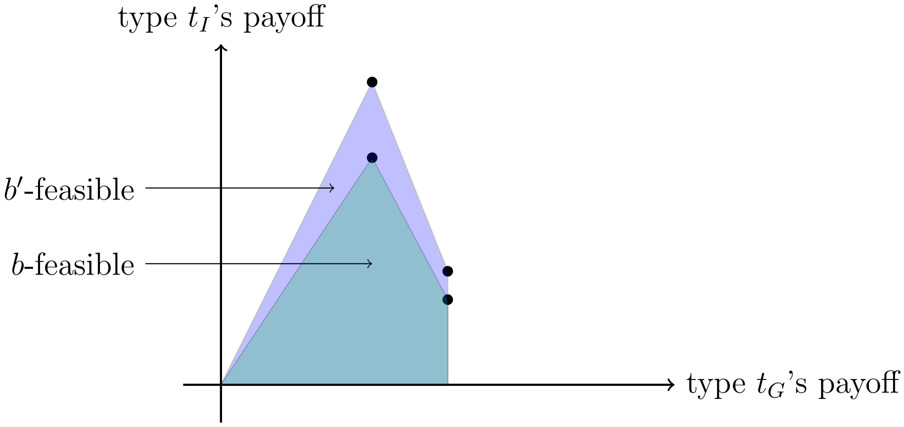

Figure 3: The sets of feasible payoff vectors at different beliefs $0 < b < b' < 1/3$ in the three-action prosecutor-judge example in Koessler and Skreta (2023). The thick vertices refer to the ex-ante optimal point $(2, 4\frac{b}{1-b})$ (or, resp., $(2, 4\frac{b'}{1-b'})$) and the point $(3, \frac{3}{2}\frac{b}{1-b})$ (or, resp., $(3, \frac{3}{2}\frac{b'}{1-b'})$) where the information designer splits the believed probability of the guilty type t_G into 0 and 2/3.

Consider the three-action version and consider a b^*-neo-optimum U^*. By feasibility, $U^*(t_G) \leq 3$. Suppose that $U^*(t_G) < 3$.

Defining $V = (3, \frac{3}{2}\frac{b^*}{1-b^*})$, note first that (b^*, V) is a neologism for all $U \leq V$, $U \neq V$.

By Lemma 4, for any b^*-feasible U with $U(t_G) < 3$, there exists a b'-feasible V with $b' > b^*$, $V(t_G) > U(t_G)$, and $V(t_I) = U(t_I)$. Thus, (b', V) is a neologism for U.

These arguments contradict the fact that U^* is a b^*-neo-optimum. This completes the proof that U^* satisfies (*) in the three-action version.

The arguments extend easily to the four-action version. Because the designer's payoff from action $\overline{a}_0$ is smaller than from action a_3, the b-feasibility sets in the four-action version are more restricted than in the three-action version, for all b. But the points $(0, 0)$, $(3, \frac{3}{2}\frac{b}{1-b})$, and $(2, 4\frac{b}{1-b})$ are still feasible for all $b < 1/3$. Thus the above arguments showing the necessity of (*) extend to the four-action version.

In the two-action version of the prosecutor-judge example in Koessler and Skreta (2023), the equivalence between neo-optimum and interim-optimum follows with the help of Remark 1.

Appendix A: omitted proofs

Proof of Lemma 3. Fix any $\epsilon > 0$. Denote by $\overline{\mathcal{U}}$ the (compact) convex hull of $\cup_{P\in\mathcal{P},b'\in B}P(b')$. Define, for any $n \in \mathbb{N}$, any list of payoff vectors $(U_P)_{P\in\mathcal{P}}$ with $U_P \in \overline{\mathcal{U}}$, and any $t \in T$, the choice probability

$$\Psi_n((U_P)_{P\in\mathcal{P}})_P(t) \;=\; \frac{n^{U_P(t)}}{\sum_{P'\in\mathcal{P}} n^{U_{P'}(t)}}.$$

Note that, for any n, the map

$$\Psi_n :\; \overline{\mathcal{U}}^{|\mathcal{P}|} \longrightarrow \mathbb{R}^{|\mathcal{P}|\cdot|T|},\; (U_P)_{P\in\mathcal{P}} \mapsto \Psi_n((U_P)_{P\in\mathcal{P}}) \quad \text{is continuous.}$$

Denote by $\mathcal{C}_n$ the (compact) convex hull of $\Psi_n(\overline{\mathcal{U}}^{|\mathcal{P}|})$. Note that $\sum_{P\in\mathcal{P}} c_P(t) = 1$ for all $t \in T$ and all $(c_P)_{P\in\mathcal{P}} \in \mathcal{C}_n$.

For any $(c_P)_{P\in\mathcal{P}} \in \mathcal{C}_n$, define $\hat{b}(c_P) \in B$ by

$$\hat{b}(c_P)(t) \;=\; \frac{b(t)c_P(t)}{\sum_{t'\in T} b(t')c_P(t')} \quad \text{for all } t \in T.$$

Define a correspondence

$$C_n :\; \overline{\mathcal{U}}^{|\mathcal{P}|} \times \mathcal{C}_n \longrightarrow \overline{\mathcal{U}}^{|\mathcal{P}|} \times \mathcal{C}_n$$

by

$$C_n((U_P)_{P\in\mathcal{P}}, (c_P)_{P\in\mathcal{P}}) \;=\; (P(\hat{b}(c_P)))_{P\in\mathcal{P}} \times \Psi_n((U_P)_{P\in\mathcal{P}}).$$

Because $\hat{b}$ is continuous, the correspondence C_n is upper-hemicontinuous and convex-valued. Moreover, $\overline{\mathcal{U}}^{|\mathcal{P}|} \times \mathcal{C}_n$ is convex and compact. Hence, by Kakutani's Theorem, C_n has a fixed point

$$((U_{P,n})_{P\in\mathcal{P}}, (c_{P,n})_{P\in\mathcal{P}})\,.$$

Now choose a subsequence $n_l \to \infty$ such that, for all P,

$$b_{P,n_l} \overset{\text{def}}{=} \hat{b}(c_{P,n_l}) \to b^*_P \text{ for some } b^*_P \in B,$$

and

$$U_{P,n_l} \to U^*_P \text{ for some } U^*_P \in \overline{\mathcal{U}}, \tag{10}$$

and, for all $t \in T$,

$$c_{P,n_l}(t) \to c^*_P(t) \text{ for some } c^*_P(t) \geq 0.$$

Hence,

$$s_{P,n_l} \overset{\text{def}}{=} \sum_{t' \in T} b_{P,n_l}(t') c_{P,n_l}(t') \to s^*_P \overset{\text{def}}{=} \sum_{t' \in T} b^*_P(t') c^*_P(t').$$

Observe that, by the definition of the fixed point,

$$U_{P,n_l} \in P(b_{P,n_l}) \text{ for all } l.$$

Taking the limit $l \to \infty$ and using the upper-hemicontinuity of P, we obtain the conclusion

$$U^*_P \in P(b^*_P). \tag{11}$$

By definition of $\hat{b}$, for all P and t,

$$b(t) c_{P,n_l}(t) \quad = \quad b_{P,n_l}(t) s_{P,n_l}.$$

Taking the limit $l \to \infty$, we obtain

$$b(t) c^*_P(t) \quad = \quad b^*_P(t) s^*_P \quad \text{for all } P \text{ and } t. \tag{12}$$

Observe that, for all $P \in \mathcal{P}$ and $t \in T$,

$$\text{if } c^*_P(t) > 0, \text{ then } U^*_P(t) \geq U^*_{P'}(t) \text{ for all } P' \in \mathcal{P}. \tag{13}$$

Indeed, if we had $U^*_P(t) < U^*_{P'}(t)$ for some P', then

$$U_{P',n_l}(t) - U_{P,n_l}(t) > \epsilon \overset{\text{def}}{=} \frac{U^*_{P'}(t) - U^*_P(t)}{2} > 0$$

for all large l, implying—by definition of Ψ_{n_l}—that

$$\frac{c_{P',n_l}(t)}{c_{P,n_l}(t)} = (n_l)^{U_{P',n_l}(t) - U_{P,n_l}(t)} > (n_l)^\epsilon \to \infty,$$

contradicting the fact that $0 < c^*_P(t) = \lim_l c_{P,n_l}(t)$.

Next, $\sum_{P \in \mathcal{P}} c^*_P(t) = 1$ for all t by definition of $\hat{b}$. Using (12) with $t = t'$ and $t = t''$,

$$b^*_P(t') c^*_P(t'') b(t'') = c^*_P(t') b(t') b^*_P(t'') \quad \text{for all } t', t'' \in T.$$

The proof is completed by defining $U^*(t) = \max_{P \in \mathcal{P}} U^*_P(t)$ for all t.

Proof of the if-part of Proposition 6. Consider a b^*-neo-optimum U.

Suppose that U is not interim-optimal, that is, a feasible pair (b', U') exists such that $U'(t) > U(t)$ for all $t \in \text{supp}(b')$. The pair (b', U') applies to all payoff vectors in a neighborhood of U and to everything below. Thus, to obtain a contradiction it is sufficient to show that a neologism exists for (b^*, U).

Given any belief b', define $r_{b'}(t) = b'(t)/b^*(t)$ for all $t \in T$.

Call a feasible pair (b', U') a *deviation if there exists an allocation family $(\rho'_t)_{t\in T}$ such that $U'(t) = \Pi(\rho'_t)(t)$ for all t, $U'(t) \le U(t)$ for all $t \in T\backslash\text{supp}(b')$,

$$\{\rho'_t \mid t \in T\} = \{\rho'_t \mid t \in \text{supp}(b')\},$$

$U'(t) \ge U(t)$ for all $t \in \text{supp}(b')$ with a strict inequality for at least one type, and any type $t \in \text{supp}(b')$ with $U'(t) = U(t)$ satisfies $r_{b'}(t) \le r^*$, where we define r^* via the equation

$$r^* \cdot \sum_{U'(t)>U(t)} b^*(t) + \sum_{U'(t)=U(t)} b'(t) = 1. \tag{14}$$

Note that we can also write this equation in the form (9).

As a first step, we show that a *deviation exists. Because U is not interim-optimal, a feasible pair (b'', U'') exists such that $U''(t) > U(t)$ for all $t \in \text{supp}(b'')$. Let $(\rho''_t)_{t\in T}$ denote a corresponding family of allocations. That is,

$$U''(t) = \Pi(\rho''_t)(t) \ge \Pi(\rho''_{\check{t}})(t) \tag{15}$$

for all $t, \check{t} \in T$, and

$$\sum_{t'\in T} b''(t')\rho''_{t'} \in P. \tag{16}$$

For all $t \in T \setminus \text{supp}(b'')$, select a

$$v(t) \in \arg\max_{\check{t}\in\text{supp}(b'')} \Pi(\rho''_{\check{t}})(t), \tag{17}$$

and let $v(t) = t$ for all $t \in \text{supp}(b'')$. Then define $\rho'_t = \rho''_{v(t)}$ and $U'(t) = \Pi(\rho'_t)(t)$ for all $t \in T$.

In other words, every type outside $\text{supp}(b'')$ gets restricted to their respective best allocation of a type in $\text{supp}(b'')$, while the types in $\text{supp}(b'')$ keep their allocations.

The construction has a second part in which we move from the belief b'' to a new belief b'. If a type t outside $\mathrm{supp}(b'')$ has utility $U'(t) > U(t)$, then we move a bit of probability to her from the type $v(t)$. This includes t into the support $\mathrm{supp}(b')$, without affecting AF because the types t and $v(t)$ get the same allocation.

We will now describe the second part of the construction more formally. For any $t' \in \mathrm{supp}(b'')$, let $w(t')$ denote the number of types outside $\mathrm{supp}(b'')$ which choose the allocation of type t' and still get more than their U utility. That is,

$$w(t') = |\{t \in T \setminus \mathrm{supp}(b'') \mid v(t) = t',\ U'(t) > U(t)\}|.$$

Given any $\epsilon > 0$, define for all $t \in T$,

$$b'(t) = \begin{cases} 0 & \text{if } t \notin \mathrm{supp}(b'') \ \text{ and } \ U'(t) \leq U(t), \\ \epsilon & \text{if } t \notin \mathrm{supp}(b'') \ \text{ and } \ U'(t) > U(t), \\ b''(t) - \epsilon w(t) & \text{if } t \in \mathrm{supp}(b''), \end{cases}$$

where ϵ is chosen so small that $b'(t) > 0$ for all $t \in \mathrm{supp}(b'')$.

PIC holds for the allocation family $(\rho'(t))_{t\in T}$. To see this, note that, for all $t, t' \in T$,

$$\Pi(\rho'_t)(t) = \Pi(\rho''_{v(t)})(t) \geq \Pi(\rho''_{v(t')})(t) = \Pi(\rho'_{t'})(t),$$

where in the cases with $t \in \mathrm{supp}(b')$ the above inequality follows from (15) with $\check{t} = v(t')$, and in the other cases the inequality follows from (17).

To show that AF holds for the allocation family $(\rho'_t)_{t\in T}$ together with the belief b', recall (16) and note that

$$\sum_{t'\in T} b'(t')\rho'_{t'} = \sum_{t'\in T} b''(t')\rho''_{t'}.$$

We conclude that (b', U') is a feasible pair. Moreover, by construction,

$$\mathrm{supp}(b') = \{t \in T \mid U'(t) > U(t)\}.$$

Thus, (b', U') is a *deviation.

Let D_2 denote the set of *deviations (b', U') such that $|\mathrm{supp}(b')|$ is minimal among all *deviations. Let D_1 denote the set of *deviations (b', U') in D_2 such that $|\{t \in \mathrm{supp}(b') \mid U'(t) = U(t)\}|$ is maximal among all *deviations in D_2. Let D_0 denote the set of *deviations (b', U') in D_1 such that $|\{t \in \operatorname{supp}(b') \mid U'(t) > U(t), r_{b'}(t) = r^*\}|$ is maximal among all *deviations in D_1.

In the following, we consider a *deviation $(b', U') \in D_0$. Let $(\rho'(t))_{t\in T}$ denote an allocation family for this *deviation. Let r_0^* denote the r^*-value for this *deviation.

It remains to show that (b', U') is a neologism for (b^*, U). To prove this, we have to show that all types t with $U'(t) > U(t)$ satisfy $r_{b'}(t) = r_0^*$.

Suppose otherwise. Then there exists a type t_1 with $U'(t_1) > U(t_1)$ and $r_{b'}(t_1) < r_0^*$ as well as a type t with $U'(t) > U(t)$ and $r_{b'}(t) > r_0^*$. Let $t_1, \ldots, t_n$ denote the types with $U'(t_i) > U(t_i)$ and $r_{b'}(t_i) \leq r_0^*$. Let $t_{n+1}, \ldots, t_{n+m}$ denote the types with $U'(t_i) > U(t_i)$ and $r_{b'}(t_i) > r_0^*$.

Now consider the following problem, where y stands for the probability mass assigned to type t_1, and $x_{k,i}$ stands for the fraction of the allocation of type t_k that is reassigned to type t_i.

$$\max_{\substack{y,\ (x_{k,i})_{k=1,\ldots,n+m,}\\ i=1,\ldots,n+m}} \quad y,$$

$$\begin{aligned}
\text{s.t.}\quad & x_{k,i} = \mathbf{1}_{k=i} \text{ for all } k \text{ and all } i \geq n+1, && (18)\\
& x_{k,i} \geq 0 \text{ for all } k \text{ and all } i \leq n, && \\
& \sum_{k=1}^{n+m} x_{k,i} = 1 \text{ for all } i \leq n, && (19)\\
& y \leq b^*(t_1) r_0^*, && (20)\\
& b'(t_k) = x_{k,1} y + \sum_{i=2}^{n} x_{k,i} b'(t_i) \quad \text{for all } k \leq n, && (21)\\
& b'(t_k) - x_{k,1} y - \sum_{i=2}^{n} x_{k,i} b'(t_i) \geq b^*(t_k) r_0^* \quad \text{for all } k > n, && (22)\\
& \sum_{k=1}^{n+m} x_{k,i} \Pi(\rho'_k)(t_i) \geq \Pi(\rho'_{\check{t}})(t_i) \text{ for all } \check{t} \in \operatorname{supp}(b') \setminus \{t_1, \ldots, t_{n+m}\}, && (23)\\
& \sum_{k=1}^{n+m} (x_{k,i} - x_{k,j}) \Pi(\rho'_k)(t_i) \geq 0 \text{ for all } i \leq n \text{ and all } j, && (24)\\
& \sum_{k=1}^{n+m} x_{k,i} \Pi(\rho'_k)(t_i) \geq U(t_i) \text{ for all } i \leq n. && (25)
\end{aligned}$$

Note that all constraints are satisfied at the point $y = b'(t_1)$ and $x_{k,i} = \mathbf{1}_{k=i}$ for all k and i. Thus, the feasibility set is non-empty and the solution value—which exists by the extreme-value theorem of Weierstrass—is $\geq b'(t_1)$.

Given any solution $\hat{y}$, $(\hat{x}_{k,i})$, define an allocation family $(\hat{\rho}_t)$ as follows:

$$\hat{\rho}_{t_i} = \sum_{k=1}^{n+m} \hat{x}_{k,i} \rho'_{t_k} \quad \text{for all } i = 1, \ldots, n+m;$$

$\hat{\rho}_t = \rho'_t$ for all $t \in \operatorname{supp}(b') \setminus \{t_1, \ldots, t_{n+m}\}$, and $\hat{\rho}_t = \hat{\rho}_{\hat{v}(t)}$ for all $t \in T \setminus \operatorname{supp}(b')$, where we choose any

$$\hat{v}(t) \in \arg \max_{\check{t} \in \operatorname{supp}(b')} \Pi(\hat{\rho}_{\check{t}})(t).$$

Define a utility vector $\hat{U}$ via $\hat{U}(t) = \Pi(\hat{\rho}_t)(t)$ for all t. Define a belief $\hat{b}$ as follows: $\hat{b}(t_1) = \hat{y}$; $\hat{b}(t_k) = b'(t_k)$ for $k = 2, \ldots, n$;

$$\hat{b}(t_k) = b'(t_k) - \hat{x}_{k,1}\hat{y} - \sum_{i=2}^{n} \hat{x}_{k,i} b'(t_i) \quad \text{for all } k = n+1, \ldots, n+m;$$

$\hat{b}(t) = b'(t)$ for all $t \in T \setminus \{t_1, \ldots, t_{n+m}\}$. Note that $\hat{b} \in B$ because

$$\begin{aligned}
\sum_{t \in T} \hat{b}(t) \quad &= \quad \sum_{t \in \operatorname{supp}(b') \setminus \{t_1, \ldots, t_{n+m}\}} b'(t) + \hat{y} + \sum_{k=2}^{n} b'(t_k) \\
&\qquad + \sum_{k=n+1}^{n+m} \left(b'(t_k) - \hat{x}_{k,1}\hat{y} - \sum_{i=2}^{n} \hat{x}_{k,i} b'(t_i) \right) \\
&= \quad \sum_{t \in \operatorname{supp}(b') \setminus \{t_1\}} b'(t) + \left(1 - \sum_{k=n+1}^{n+m} \hat{x}_{k,1} \right) \hat{y} - \sum_{i=2}^{n} \sum_{k=n+1}^{n+m} \hat{x}_{k,i} b'(t_i)
\end{aligned}$$

and, recalling $b' \in B$ and using constraint (19), the above chain continues as

$$\begin{aligned}
&= \quad 1 - b'(t_1) + \sum_{k=1}^{n} \hat{x}_{k,1}\hat{y} - \sum_{i=2}^{n} \left(1 - \sum_{k=1}^{n} \hat{x}_{k,i} \right) b'(t_i) \\
&= \quad 1 - \sum_{i=1}^{n} b'(t_i) + \sum_{k=1}^{n} \hat{x}_{k,1}\hat{y} + \sum_{i=2}^{n} \sum_{k=1}^{n} \hat{x}_{k,i} b'(t_i) \quad = \quad 1,
\end{aligned}$$

where the last equation follows from the formula that is obtained by summing the constraints (21) across all $k \leq n$. Next we show that $(\hat{b}, \hat{U})$ is feasible. To verify condition AF for $(\hat{b}, \hat{U})$, note that

$$\sum_{t\in T\setminus\{t_1,\dots,t_{n+m}\}} \hat{b}(t)\hat{\rho}_t \quad = \quad \sum_{t\in T\setminus\{t_1,\dots,t_{n+m}\}} b'(t)\rho'_t$$

and

$$\begin{aligned}
\sum_{i=1}^{n+m} \hat{b}(t_i)\hat{\rho}_{t_i} \quad &= \quad \hat{y}\hat{\rho}_{t_1} + \sum_{i=2}^{n} b'(t_i)\hat{\rho}_{t_i} + \sum_{k=n+1}^{n+m} \hat{b}(t_k)\rho'_{t_k} \\
&= \quad \hat{y}\sum_{k=1}^{n+m} \hat{x}_{k,1}\rho'_{t_k} + \sum_{i=2}^{n} b'(t_i)\sum_{k=1}^{n+m} \hat{x}_{k,i}\rho'_{t_k} \\
&\qquad + \sum_{k=n+1}^{n+m} \left(b'(t_k) - \hat{x}_{k,1}\hat{y} - \sum_{i=2}^{n} \hat{x}_{k,i} b'(t_i) \right) \rho'_{t_k} \\
&= \quad \sum_{k=1}^{n} \left(\hat{y}\hat{x}_{k,1} + \sum_{i=2}^{n} b'(t_i)\hat{x}_{k,i} \right) \rho'_{t_k} + \sum_{k=n+1}^{n+m} b'(t_k)\rho'_{t_k} \\
&\overset{(21)}{=} \textstyle\sum_{k=1}^{n+m} \quad b'(t_k)\rho'_{t_k}.
\end{aligned}$$

We also have to verify PIC. For any $t \in T \setminus \text{supp}(b')$ and $t' \in \text{supp}(b')$,

$$\Pi(\hat{\rho}_t)(t) = \Pi(\hat{\rho}_{\hat{v}(t)})(t) \geq \Pi(\hat{\rho}_{t'})(t),$$

by definition of $\hat{v}(t)$. Similarly, for any $t, t' \in T \setminus \text{supp}(b')$,

$$\Pi(\hat{\rho}_t)(t) = \Pi(\hat{\rho}_{\hat{v}(t)})(t) \geq \Pi(\hat{\rho}_{v(t')})(t) = \Pi(\hat{\rho}_{t'})(t).$$

For all $t \in \text{supp}(b') \setminus \{t_1, \dots, t_n\}$ and all $t' \in T$,

$$\Pi(\hat{\rho}_t)(t) = \Pi(\rho'_t)(t) = \max_{\check{t}\in T} \Pi(\rho'_{\check{t}})(t) \geq \Pi(\hat{\rho}_{t'})(t).$$

For all $t \in \{t_1, \dots, t_n\}$ and all $t' \in \{t_1, \dots, t_{n+m}\}$, constraint (24) directly implies $\Pi(\hat{\rho}_t)(t) \geq \Pi(\hat{\rho}_{t'})(t)$.

This then also implies that for all $t \in \{t_1, \dots, t_n\}$ and all $t' \in T \setminus \text{supp}(b')$, $\Pi(\hat{\rho}_t)(t) \geq \Pi(\hat{\rho}_{\hat{v}(t')})(t) = \Pi(\hat{\rho}_{t'})(t)$.

For all $t \in \{t_1, \dots, t_n\}$ and all $t' \in \text{supp}(b') \setminus \{t_1, \dots, t_{n+m}\}$, constraint (23) directly implies $\Pi(\hat{\rho}_t)(t) \geq \Pi(\hat{\rho}_{t'})(t)$. This completes the proof of PIC.

Next we show that $(\hat{b}, \hat{U})$ is a *-deviation.

For all $i = 1, \ldots, n$, we have $\hat{U}(t_i) \geq U(t_i)$ by constraint (25). For all $t \in \mathrm{supp}(b') \setminus \{t_1, \ldots, t_{n+m}\}$, we have $\hat{U}(t) = U'(t) = U(t)$ by construction. For all $i = n+1, \ldots, n+m$, we have $\hat{U}(t_i) = U'(t_i) > U(t_i)$ by construction. For all $t \in T \setminus \mathrm{supp}(b')$,

$$\hat{U}(t) = \Pi(\hat{\rho}_{\hat{v}(t)}) \leq \max_{\check{t} \in \mathrm{supp}(b')} \Pi(\rho'_{\check{t}})(t) \leq \Pi(\rho'_t)(t) = U'(t) \leq U(t).$$

In particular, for any type $t \in T$, if $\hat{U}(t) > U(t)$ then $U'(t) > U(t)$, and all types $t \in T$ with $\hat{U}(t) = U(t)$ satisfy $r_{\hat{b}}(t) \leq r_0^*$. Thus, to complete the proof that $(\hat{b}, \hat{U})$ is a *-deviation, it remains to show that the r^* value for $(\hat{b}, \hat{U})$ satisfies $r^* \geq r_0^*$.

Using the definition (14),

$$\begin{aligned} 1 &= r_0^* \cdot \sum_{U'(t) > U(t)} b^*(t) + \sum_{U'(t) = U(t)} b'(t) \\ &\geq r_0^* \cdot \sum_{\hat{U}(t) > U(t)} b^*(t) + \sum_{\substack{U'(t) > U(t), \\ \hat{U}(t) = U(t)}} \hat{b}(t) + \sum_{U'(t) = U(t)} b'(t) \\ &= r_0^* \cdot \sum_{\hat{U}(t) > U(t)} b^*(t) + \sum_{\hat{U}(t) = U(t)} \hat{b}(t), \end{aligned}$$

implying that $r^* \geq r_0^*$.

Next we show that the constraints (20), (22), (23), and (25) are not binding at the solution $\hat{y}$, $(\hat{x}_{k,i})$.

Note that $(\hat{b}, \hat{U}) \in D_2$ because $\mathrm{supp}(\hat{b}) = \mathrm{supp}(b')$. Moreover, because any type $t \in \mathrm{supp}(\hat{b})$ with $U'(t) = U(t)$ also satisfies $\hat{U}(t) = U(t)$, we even have $(\hat{b}, \hat{U}) \in D_1$. Thus, $\hat{U}(t_i) > U(t_i)$ for all $i = 1, \ldots, n$, implying that the constraints (25) are not binding.

As a consequence, $r^* = r_0^*$.

By construction, any type $t \in T$ with $r_{b'}(t_1) = r_0^*$ also satisfies $r_{\hat{b}}(t_1) = r_0^*$. Thus, we even have $(\hat{b}, \hat{U}) \in D_0$, implying that the constraints (20) and (22) are not binding.

To show that the constraints (23) are not binding, we suppose that

$$\hat{U}(t_i) = \Pi(\hat{\rho}_{\check{t}})(t_i)$$

for some $i \leq n$ and some $\mathring{t} \in \operatorname{supp}(b') \setminus \{t_1, \dots, t_{n+m}\}$ and derive a contradiction. Define an allocation family $(\mathring{\rho}_t)_{t\in T}$ as follows. Let

$$\mathring{\rho}_{t_i} = \frac{\hat{b}(t_i)}{\hat{b}(t_i) + \hat{b}(\mathring{t})}\hat{\rho}_{t_i} + \frac{\hat{b}(\mathring{t})}{\hat{b}(t_i) + \hat{b}(\mathring{t})}\hat{\rho}_{\mathring{t}}; \tag{26}$$

let $\mathring{\rho}_t = \hat{\rho}_t$ for all $t \in \operatorname{supp}(b') \setminus \{t_i, \mathring{t}\}$; for all $t \in \{\mathring{t}\} \cup T \setminus \operatorname{supp}(b')$, let $\mathring{\rho}_t = \hat{\rho}_{\mathring{v}(t)}$, where we choose any

$$\mathring{v}(t) \in \arg \max_{\check{t} \in \operatorname{supp}(b') \setminus \{\mathring{t}\}} \Pi(\mathring{\rho}_{\check{t}})(t).$$

Define a belief $\mathring{b}$ as follows. Let

$$\mathring{b}(t_i) = \hat{b}(t_i) + \hat{b}(\mathring{t}); \tag{27}$$

let $\mathring{b}(t_i) = 0$; let $\mathring{b}(t) = \hat{b}(t)$ for all $\in T \setminus \{t_i, \mathring{t}\}$.

Define a payoff vector $\mathring{U}$ via $\mathring{U}(t) = \Pi(\mathring{\rho}_t)(t)$ for all $t \in T$.

By the supposed indifference, $\mathring{U}(t_i) = \hat{U}(t_i)$. Moverover, the set of alternative allocations to choose from has shrunk:

$$\{\mathring{\rho}_{t'} | t' \in T \setminus \{t_i\}\} \subseteq \{\hat{\rho}_{t'} | t' \in T \setminus \{t_i\}\}.$$

Thus, because $(\hat{\rho}_t)_{t\in T}$ satisfies PIC, the allocation family $(\mathring{\rho}_t)_{t\in T}$ also satisfies PIC for $t = t_i$ and all $t' \in T$. The same holds for $t \in \operatorname{supp}(b') \setminus \{t_i, \mathring{t}\}$ and all $t' \neq t_i$ because $\mathring{\rho}_t = \hat{\rho}_i$.

The allocation family $(\mathring{\rho}_t)_{t\in T}$ satisfies PIC for all $t \in \operatorname{supp}(b') \setminus \{t_i, \mathring{t}\}$ and for $t' = t_i$ because

$$\Pi(\mathring{\rho}_{t_i})(t) \overset{(26)}{\leq} \max\{\Pi(\hat{\rho}_{t_i})(t), \Pi(\hat{\rho}_{\mathring{t}})(t)\} \leq \Pi(\hat{\rho}_t)(t) = \Pi(\mathring{\rho}_t)(t).$$

Finally, the allocation family $(\mathring{\rho}_t)_{t\in T}$ satisfies PIC for all $t \in \{\mathring{t}\} \cup T \setminus \operatorname{supp}(b')$ and all t' due to the definition of $\mathring{v}(t)$. This completes the verification of PIC.

To verify AF for $(\mathring{\rho}_t)_{t\in T}$ together with $\mathring{b}$, note that

$$\sum_{t \in T \setminus \{t_i, \mathring{t}\}} \mathring{b}(t)\mathring{\rho}_t \quad = \quad \sum_{t \in T \setminus \{t_i, \mathring{t}\}} \hat{b}(t)\hat{\rho}_t,$$

and

$$\mathring{b}(t_i)\mathring{\rho}_{t_i} + \mathring{b}(\mathring{t})\mathring{\rho}_{\mathring{t}} \quad = \quad \hat{b}(t_i)\hat{\rho}_{t_i} + \hat{b}(\mathring{t})\hat{\rho}_{\mathring{t}}$$

by (26) and (27). Thus, $(\mathring{b}, \mathring{U})$ is feasible.

Also note that $\mathrm{supp}(\mathring{b}) = \mathrm{supp}(b') \setminus \{\mathring{t}\}$.

To obtain a contradiction it remains to verify that $(\mathring{b}, \mathring{U})$ is a *deviation because then $(b', U') \notin D_2$.

By construction, $\mathring{U}(t) = \hat{U}(t)$ for all $t \in \mathrm{supp}(\mathring{b})$ and $\mathring{U}(t) \leq \hat{U}(t)$ for all $t \in T \setminus \mathrm{supp}(\mathring{b})$. Thus, using the definition (14),

$$\begin{aligned} 1 &= r_0^* \cdot \sum_{\hat{U}(t) > U(t)} b^*(t) + \sum_{\hat{U}(t) = U(t)} \hat{b}(t) \\ &= r_0^* \cdot \sum_{\mathring{U}(t) > U(t)} b^*(t) + \sum_{\mathring{U}(t) = U(t)} \mathring{b}(t) \ + \hat{b}(\mathring{t}), \end{aligned}$$

implying that the r^* value for $(\mathring{b}, \mathring{U})$ satisfies $r^* > r_0^*$. Thus, $(\mathring{b}, \mathring{U})$ is a *deviation.

To obtain the final contradiction, we will now define a perturbation of the presumed max-solution that satisfies all constraints and increases the solution value.

Denote $T^{\leq} = \{t_1, \ldots, t_n\}$. Given the allocation family $(\hat{\rho}_t)_{t \in T}$, we say that $(v_1, \ldots, v_l)$ (where $l \geq 1$) is a *chain-indifference path* in $T^{\leq}$ if $v_1, \ldots, v_l \in T^{\leq}$ and $\Pi(\hat{\rho}_{v_{i+1}})(v_i) = \Pi(\hat{\rho}_{v_i})(v_i)$ for all $i < l$.

Let $T^{\leq}_{=}$ denote the types $t \in T^{\leq}$ such that a chain-indifference path $(v_1, \ldots, v_l)$ exists with $v_1 = t_1$ and $v_l = t$.

An *indifference graph* $(T^{\leq}_{=}, g)$ is defined as a directed graph such that (i) the set of nodes equals $T^{\leq}_{=}$ and (ii) $\Pi(\hat{\rho}_{t'})(t) = \Pi(\hat{\rho}_t)(t)$ for each edge $(t, t') \in g$.

By definition of $T^{\leq}_{=}$, there exists an indifference graph such that, for all $t' \in T^{\leq}_{=}$, there exists a (chain-indifference) path from t_1 to t'. Requiring this property, let $(T^{\leq}_{=}, g)$ denote an indifference graph with a minimal number of edges.

Then $(T^{\leq}_{=}, g)$ is a tree with root t_1; that is, no edge points to t_1, and there exists a unique path from t_1 to each node in $T^{\leq}_{=}$. (To see the uniqueness statement, suppose that paths p_1 and p_2 lead to the same node, and $(t_1', t'') \in p_1$, $(t_2', t'') \in p_2$ with $t_1' \neq t_2'$ are edges where the two paths join. Then $(T^{\leq}_{=}, g \setminus \{(t_1', t'')\})$ is an indifference graph will a smaller number of edges in which still there exists a path from t_1 to any other node—contradiction.)

For each $t \in T^{\leq}_{=}$, let the index of a "most preferred type" among those with $r > r^*$ (recall that the constraints (22) are not binding) be denoted

$$\iota(t) \in \arg \max_{j \in \{n+1,\ldots,n+m\}} \Pi(\hat{\rho}_{t_j})(t).$$

For all i with $t_i \in T^{\leq}_{=}$, define

$$\sigma(i) = \{j \in \{1, \ldots, n\} \mid t_j \text{ is a direct successor of } t_i \text{ in } (T^{\leq}_{=}, g)\}.$$

Note that $\sigma(i) = \emptyset$ means that t_i is an end node in $(T^{\leq}_{=}, g)$. For each $j \neq 1$ with $t_j \in T^{\leq}_{=}$, let $\sigma^{-1}(j)$ denote the index of the direct predecessor of t_j in $(T^{\leq}_{=}, g)$.

Fix any $0 < \epsilon < 1$. The following definition works recursively from the end nodes backwards through the tree. Define

$$\omega_j \quad = \quad \frac{b'(t_j)}{b'(t_{\sigma^{-1}(j)})} \frac{\epsilon + (1-\epsilon)\sum_{k \in \sigma(j)} \omega_k}{(1-\epsilon)} \quad \text{for all } j \notin \sigma(1) \text{ with } t_j \in T^{\leq}_{=}$$

and

$$z_j \quad = \quad b'(t_j) \frac{\epsilon + (1-\epsilon)\sum_{k \in \sigma(j)} \omega_k}{(1-\epsilon)} \quad \text{for all } j \in \sigma(1).$$

Define

$$\mathring{y} \quad = \quad \frac{1}{1-\epsilon}\hat{y} + \sum_{j \in \sigma(1)} z_j \tag{28}$$

and

$$\omega_j \quad = \quad \frac{z_j}{\mathring{y}} \quad \text{for all } j \in \sigma(1).$$

Thus, replacing $z_j = \omega_j \mathring{y}$ in (28) and solving for $\mathring{y}$, we find that

$$\mathring{y} \quad = \quad \frac{\hat{y}}{(1 - \sum_{j \in \sigma(1)} \omega_j)(1-\epsilon)}. \tag{29}$$

For all i with $t_i \in T^{\leq}_{=}$ and all $k = 1, \ldots, n+m$, define

$$\mathring{x}_{k,i} \quad = \quad \left(\hat{x}_{k,i} \left(1 - \sum_{j \in \sigma(i)} \omega_j \right) + \sum_{j \in \sigma(i)} \hat{x}_{k,j}\omega_j \right) (1-\epsilon) + \mathbf{1}_{k=\iota(t_i)}\epsilon. \tag{30}$$

For all $i \leq n$ with $t_i \notin T^{\leq}_{=}$, and all $i = n{+}1, \ldots, n{+}m$ and all $k = 1, \ldots, n{+}m$, define $\mathring{x}_{k,i} = \mathbf{1}_{k=i}$.

First note that

$$0 < \omega_j \to_{\epsilon \to 0} 0 \quad \text{for all } j \neq 1 \text{ with } t_j \in T^{\leq}_{=}.$$

(This is seen recursively, arguing backwards from the end nodes in $(T^{\leq}_{=}, g)$.)

Thus, through choosing ϵ sufficiently close to 0, we can guarantee that $\sum_{j \in \sigma(i)} \omega_j$ is close to 0 for all $i = 1, \ldots, n$, implying

$$\mathring{x}_{k,i} \geq 0$$

and, using (29),

$$\mathring{y} > \hat{y}.$$

In particular, once we show that $\mathring{y}, (\mathring{x}_{k,i})$ satisfies all remaining constraints of our max-problem, then we have a contradiction to the assumption that $\hat{y}, (\hat{x}_{k,i})$ is a solution.

First of all, recall that the constraints (20), (22), (23), and (25) are not binding at the solution $\hat{y}$, $(\hat{x}_{k,i})$.

Thus, because $\mathring{x}_{k,i} \to \hat{x}_{k,i}$ and $\mathring{y} \to \hat{y}$ as $\epsilon \to 0$, the constraints (20), (22), (23), and (25) are also strictly satisfied at $\mathring{y}, (\mathring{x}_{k,i})$, assuming ϵ is sufficiently close to 0.

Define the auxiliary variables $\hat{b}(t_1) = \hat{y}$ and $\hat{b}(t_i) = b'(t_i)$ for all $i = 2, \ldots, n$. Defining the column vectors $b'^n = (b'(t_1), \ldots, b'(t_n))^T$ and $\hat{b}^n = (\hat{b}(t_1), \ldots, \hat{b}(t_n))^T$ and the square matrix $\hat{X} = (\hat{x}_{k,i})_{k \leq n,\ i \leq n}$, constraint (21) reads

$$b'^n = \hat{X} \hat{b}^n. \tag{31}$$

Defining the square matrices $\mathring{X} = (\mathring{x}_{k,i})_{k \leq n,\ i \leq n}$ and $H = (h_{j,i})_{j \leq n,\ i \leq n}$ via $h_{j,i} = \mathbf{1}_{j=i}$ if $t_i \notin T^{\leq}_{=}$, and

$$h_{i,i} = \left(1 - \sum_{j \in \sigma(i)} \omega_j\right)(1{-}\epsilon), \quad h_{j,i} = \omega_j (1{-}\epsilon) \text{ for all } j \in \sigma(i), \quad h_{j,i} = 0 \text{ otherwise,}$$

if $t_i \in T^{\leq}_{=}$, definition (30) implies the matrix-product equation

$$\mathring{X} = \hat{X} H. \tag{32}$$

Now define the auxiliary variables $\mathring{b}(t_1) = \mathring{y}$ and $\mathring{b}(t_i) = b'(t_i)$ for all $i = 2, \dots, n$. Defining the column vector $\mathring{b}^n = (\mathring{b}(t_1), \dots, \mathring{b}(t_n))^T$, the definition of the ω_j variables implies that

$$\mathring{b}(t_{\sigma^{-1}(j)})\omega_j(1-\epsilon)+\mathring{b}(t_j)(1-\sum_{k\in\sigma(j)}\omega_k)(1-\epsilon) = \hat{b}(t_j) \quad \text{for all } j \neq 1 \text{ with } t_j \in T^{\leq}_{=},$$

and (29) implies

$$\mathring{b}(t_1)(1 - \sum_{k\in\sigma(1)}\omega_k)(1 - \epsilon) = \hat{b}(t_1).$$

In matrix notation,

$$H\mathring{b}^n = \hat{b}^n.$$

Together with (31) and (32) this implies

$$b'^n = \mathring{X}\mathring{b}^n.$$

That is, constraint (21) holds for $\mathring{y}, (\mathring{x}_{k,i})$.

That constraint (19) holds for $(\mathring{x}_{k,i})$ is seen by summing (30) across all $k = 1, \dots, n+m$ and noting that (19) holds for $(\hat{x}_{k,i})$.

It remains to verify (24) for $(\mathring{x}_{k,i})$, that is, for all $i \leq n$ and all j,

$$\sum_{k=1}^{n+m}(\mathring{x}_{k,i} - \mathring{x}_{k,j})\Pi(\rho'_k)(t_i) \geq 0. \tag{33}$$

Consider any i with $t_i \notin T^{\leq}_{=}$ and any $j > n$, or $j \leq n$ with $t_j \notin T^{\leq}_{=}$. Then (33) is immediate because $\mathring{x}_{k,i} = \hat{x}_{k,i}$ and $\mathring{x}_{k,j} = \hat{x}_{k,j}$ and (24) holds for $(\hat{x}_{k,i})$.

Consider any i with $t_i \notin T^{\leq}_{=}$ and any j with $t_j \in T^{\leq}_{=}$. Then (33) follows from (30) because

$$\begin{aligned}
&\sum_{k=1}^{n+m}(\mathring{x}_{k,i}-\mathring{x}_{k,j})\Pi(\rho'_k)(t_i)\\
=&\sum_{k=1}^{n+m}(\hat{x}_{k,i}-\mathring{x}_{k,j})\Pi(\rho'_k)(t_i)\\
=&(1-\sum_{l\in\sigma(j)}\omega_l)(1-\epsilon)\sum_{k=1}^{n+m}(\hat{x}_{k,i}-\hat{x}_{k,j})\Pi(\rho'_k)(t_i)\\
&+\sum_{l\in\sigma(j)}\omega_l(1-\epsilon)\sum_{k=1}^{n+m}(\hat{x}_{k,i}-\hat{x}_{k,l})\Pi(\rho'_k)(t_i)+\epsilon\sum_{k=1}^{n+m}(\hat{x}_{k,i}-\hat{x}_{k,\iota(t_j)})\Pi(\rho'_k)(t_i)\\
\geq&0,
\end{aligned}$$

where the inequality follows because (24) holds for $(\hat{x}_{k,i})$.

Consider any i with $t_i \in T^{\leq}_{=}$ and $j \leq n$ with $t_j \notin T^{\leq}_{=}$. By definition of the indifference tree, (24) holds as a strict inequality for $(\hat{x}_{k,i})$. Thus, assuming that ϵ is sufficiently close to 0, (33) holds.

Consider any i with $t_i \in T^{\leq}_{=}$. By definition of the indifference tree,

$$\sum_{k=1}^{n+m}\mathring{x}_{k,i}\Pi(\rho'_k)(t_i)=(1-\epsilon)\sum_{k=1}^{n+m}\hat{x}_{k,i}\Pi(\rho'_k)(t_i)+\epsilon\;\Pi(\rho'_{\iota(t_i)})(t_i). \tag{34}$$

Because (24) holds for $(\hat{x}_{k,i})$ with $j=\iota(t_i)$, we conclude that

$$\sum_{k=1}^{n+m}\mathring{x}_{k,i}\Pi(\rho'_k)(t_i)\geq\Pi(\rho'_{\iota(t_i)})(t_i).$$

Thus, for any $j > n$, using the definition of $\iota(t_i)$,

$$\sum_{k=1}^{n+m}\mathring{x}_{k,i}\Pi(\rho'_k)(t_i)\geq\Pi(\rho'_{t_j})(t_i),$$

implying (33).

Finally, consider any i with $t_i \in T^{\leq}_{=}$ and any j with $t_j \in T^{\leq}_{=}$. Applying

(34), and applying it again with i replaced by j, we find

$$\begin{aligned}
&\sum_{k=1}^{n+m} (\mathring{x}_{k,i} - \mathring{x}_{k,j})\,\Pi(\rho'_k)(t_i)\\
=&(1-\epsilon)\sum_{k=1}^{n+m} (\hat{x}_{k,i} - \hat{x}_{k,j})\,\Pi(\rho'_k)(t_i) + \epsilon\left(\Pi(\rho'_{\iota(t_i)})(t_i) - \Pi(\rho'_{\iota(t_j)})(t_i)\right)\\
\geq&0,
\end{aligned}$$

where the inequality follows because both terms are ≥ 0—the left term because (24) holds for $(\hat{x}_{k,i})$ and the right term by definition of $\iota(t_i)$.

In summary, we have shown that $\mathring{y}, (\mathring{x}_{k,i})$ satisfies all constraints of the max problem and $\mathring{y} > \hat{y}$, contradicting the fact that $\hat{y}, (\hat{x}_{k,i})$ is a solution. □

Proof of Proposition 5. In this proof, we use the notation from Mylovanov and Tröger (2012). The payoff vector that corresponds to an allocation ρ is U_0^ρ. We now say that the payoff vector is strongly neologism-proof if the underlying allocation is strongly neologism-proof.

In light of Proposition 6, it is sufficient to show that a payoff vector is strongly neologism-proof if and only if it is interim optimal.

The direction "only if" is immediate from the definitions. To show "if", let ρ denote an allocation such that U_0^ρ is interim optimal and suppose that ρ is not strongly neologism-proof. Then there exists a belief q_0 and a q_0-feasible allocation ρ' such that q_0 puts zero probability on all types that are strictly better off in ρ than in ρ' or that already obtain in ρ the maximum feasible payoff. Moreover, there exists a type $t'_0 \in \text{supp}(q_0)$ such that $U_0^{\rho'}(t'_0) > U_0^\rho(t'_0)$.

Let $f : \mathbf{T_{-0}} \to \mathcal{Z}$ denote the expected agent-allocation at the belief q_0, that is,

$$f(\cdot) = \sum_{t_0 \in T_0} q_0(t_0)\rho'(t_0, \cdot).$$

By the q_0-feasibility of ρ', the agents' constraints are satisfied at f.

By separability, there exists an agent-allocation $e : \mathbf{T_{-0}} \to \mathcal{Z}$ such that all agents' incentive and participation constraints are satisfied strictly.

For any $\delta < 1$, define a belief $\hat{q}_0$ and an allocation $\hat{\rho}$ such that, for all $t_0 \neq t'_0$, $\hat{q}_0(t_0) = \delta q_0(t_0)$ and $\hat{\rho}(t_0, \cdot) = \rho'(t_0, \cdot)$, and

$$\hat{q}_0(t'_0) = 1 - \delta + \delta q_0(t'_0))$$

and

$$\hat{\rho}(t_0', \cdot) = \frac{\delta q_0(t_0')}{\hat{q}_0(t_0')}\rho'(t_0', \cdot) + \frac{1-\delta}{\hat{q}_0(t_0')}e(\cdot).$$

Assume δ is close to 1 such that $\text{supp}(\hat{q}_0) = \text{supp}(q_0)$.

At the belief $\hat{q}_0$ and the allocation $\hat{\rho}$, the agents expect to obtain the allocation f with probability δ, and the allocation e with probability $1-\delta$. Thus, at $\hat{\rho}$ the agents' constraints are satisfied strictly. Let $\overline{e}$ denote an allocation where each type $t_0 \in \text{supp}(q_0)\backslash\{t_0'\}$ obtains their maximum feasible payoff.

By construction, for $\epsilon > 0$ close to 0, the allocation $\epsilon\overline{e} + (1-\epsilon)\hat{\rho}$ satisfies the agents' constraints at belief $\hat{q}_0$ and, for all types in $\text{supp}(\hat{q}_0)$, is strictly preferred over the allocation ρ.

Now it remains to change $\hat{q}_0$ and $\hat{\rho}$ so that the principal's incentive constraints are reestablished; this works for all private-value environments: if a type t_0 in the support of $\hat{q}_0$ is attracted to what a type $\hat{t}_0$ offers, then let t_0 offer the average of what both types used to offer, and move the probability from type $\hat{t}_0$ to t_0. This procedure continues until incentive compatibility is satisfied for the types in the (remaining) support. Let the types outside the support choose their optimum among the offerings of the types in the support. Then we have a deviation as considered in the definition interim-optimality. □

Proof of Lemma 4. Note first that the feasibility set belongs to the set of realizable payoff vectors $[0,3]^2$. Koessler and Skreta (2023) show that all four points mentioned in the lemma are b-feasible; let μ^1 denote a mechanism that yields the payoff vector $(0,0)$, let μ^2 denote a mechanism that yields the payoff vector $(3,0)$, let $\mu^{3,b}$ denote a mechanism that yields the payoff vector $(3, \frac{3}{2}\frac{b}{1-b})$, and let $\mu^{4,b}$ denote a mechanism that yields the payoff vector $(2, 4\frac{b}{1-b})$. In the following, we fix $\mu^{4,b}$ as follows:

$$\mu^{4,b}(\underline{a}_0|G) = \mu^{4,b}(a_3|G) = \mu^{4,b}(a_3|I) = 0, \quad \mu^{4,b}(a_2|G) = 1,$$
$$\mu^{4,b}(\underline{a}_0|I) = \frac{1-3b}{1-b}, \quad \mu^{4,b}(a_2|I) = \frac{2b}{1-b}.$$

To further discuss the feasibility restrictions, we need to consider the obedience constraints. To this end, we must fix a payoff function for the agent that induces the agent's belief-dependent action as described in Koessler and Skreta (2023). Let

$$u_1(\underline{a}_0|G) = u_1(\underline{a}_0|I) = 0,\ u_1(a_2|G) = 2,\ u_1(a_2|I) = -1,\ u_1(a_3|G) = 3,\ u_1(a_3|I) = -3.$$

Define

$$\underline{g} = -2\frac{b}{1-b} \quad \text{and} \quad \overline{g} = \frac{5}{2}\frac{b}{1-b}.$$

We will show that, for all $g \in [\underline{g}, \overline{g}]$, the mechanism $\mu^{4,b}$ maximizes the weighted average of the principal-types' payoffs, where g denotes the weight on type t_G, while the payoff of type t_I has the weight 1. Formally, we show that (*) $\mu^{4,b}$ is a maximizer of the function

$$\phi_g(\mu) = g(2\mu(a_2|t_G) + 3\mu(a_3|t_G)) + 2\mu(a_2|t_I) + 3\mu(a_3|t_I)$$

subject to the b-feasibility constraints for the mechanism μ. This proves the lemma because

$$\phi_{\overline{g}}(\mu^{4,b}) = \phi_{\overline{g}}(\mu^{3,b}) \text{ and } \phi_{\underline{g}}(\mu^{4,b}) = \phi_{\underline{g}}(\mu^{1}).$$

To show (*), we verify that the Karush-Kuhn-Tucker first-order conditions are satisfied at the point $\mu = \mu^{4,b}$ for all $g \in [\underline{g}, \overline{g}]$. Let $\lambda_{3,0}$ denote the Lagrangian multiplier for the obedience constraint that the agent when the recommended action is a_3 cannot gain from taking action $\underline{a}_0$ instead. Let $\lambda_{3,2}$, $\lambda_{2,0}$, $\lambda_{2,3}$, $\lambda_{0,2}$, and $\lambda_{0,3}$, denote the Lagrangian multipliers of the other obedience constraints, where in all variables the first lower index indicates the recommended action and the second indicates a non-recommended action. The KKT conditions require that

$$\lambda_{2,3} = \lambda_{0,2} = \lambda_{0,3} = 0$$

because the corresponding obedience conditions are not binding at $\mu^{4,b}$. The other three obedience constraints are binding at $\mu^{4,b}$, implying that the KKT conditions require the inequalities

$$\lambda_{3,0} \geq 0, \quad \lambda_{3,2} \geq 0, \quad \lambda_{2,0} \geq 0. \tag{35}$$

Let λ_G and λ_I denote the Lagrangian multipliers for the probability constraints

$$\begin{aligned} \mu(\underline{a}_0|t_G) + \mu(a_2|t_G) + \mu(a_3|t_G) &= 1, \\ \mu(\underline{a}_0|t_I) + \mu(a_2|t_I) + \mu(a_3|t_I) &= 1. \end{aligned}$$

We will not introduce Lagrangian multipliers for the probability-boundary constraints $0 \leq \mu(a|t) \leq 1$, but will write the KKT conditions as appropriate inequalities.

Differentiating the Lagrangian with respect to $\mu(\underline{a}_0|t_I)$ yields the first-order condition

$$\lambda_I = 0.$$

(Note that the condition is an equality because $0 < \mu^{4,b}(\underline{a}_0|t_I) < 1$, implying that the probability-boundary constraints are not binding.)

Differentiating the Lagrangian with respect to $\mu(\underline{a}_0|t_G)$ yields the first-order condition

$$\lambda_G \leq 0.$$

Differentiating the Lagrangian with respect to $\mu(a_2|t_I)$ yields the first-order condition

$$2 + \lambda_{2,0}(1-b)(-1) = 0,$$

where we have already used that $\lambda_I = 0$. Differentiating the Lagrangian with respect to $\mu(a_2|t_G)$ yields the first-order condition

$$r \cdot 2 + \lambda_G + \lambda_{2,0} b \cdot 2 \geq 0. \tag{36}$$

Differentiating the Lagrangian with respect to $\mu(a_3|t_I)$ yields the first-order condition

$$3 + \lambda_{3,0}(1-b)(-3) + \lambda_{3,2}(1-b)(-2) \leq 0.$$

Differentiating the Lagrangian with respect to $\mu(a_3|t_G)$ yields the first-order condition

$$r \cdot 3 + \lambda_G + \lambda_{3,0} b(-3) + \lambda_{3,2} b \cdot 1 \leq 0. \tag{37}$$

Here we can interpret (36) as an upper bound for $-\lambda_G$ and (37) as a lower bound (essentially we are applying the Fourier-Motzkin algorithm in the following). Thus we can remove λ_G from the first-order system of equations and inequalities and conclude that a solution to the Karush-Kuhn-Tucker condition exists if and only if the following system has a solution:

$$\begin{array}{rcl} 2 + \lambda_{2,0}(1-b)(-1) & = & 0, \\ r \cdot 2 + \lambda_{2,0} b \cdot 2 & \geq & 0, \\ 3 + \lambda_{3,0}(1-b)(-3) + \lambda_{3,2}(1-b)(-2) & \leq & 0, \\ r \cdot 2 + \lambda_{2,0} b \cdot 2 & \geq & r \cdot 3 + \lambda_{3,0} b(-3) + \lambda_{3,2} b, \end{array}$$

and (35). The first equation implies $\lambda_{2,0} = 2/(1-b) > 0$. Plugging this into the remaining conditions yields the simplified system

$$\begin{aligned} r \cdot 2 + 4\frac{b}{1-b} &\geq 0, \\ \frac{b}{1-b}4 - r &\geq 3b\lambda_{3,0} + b\lambda_{3,2}, \\ 3 &\leq 3(1-b)\lambda_{3,0} + 2(1-b)\lambda_{3,2}, \\ \lambda_{3,0} &\geq 0, \\ \lambda_{3,2} &\geq 0. \end{aligned}$$

This can be rearranged as follows.

$$\begin{aligned} r &\geq \underline{g}, \\ \lambda_{3,2} &\leq \frac{1}{1-b}4 - \frac{1}{b}r - 3\lambda_{3,0}, \\ \frac{3}{2}\frac{1}{1-b} - \frac{3}{2}\lambda_{3,0} &\leq \lambda_{3,2}, \\ \lambda_{3,0} &\geq 0, \\ \lambda_{3,2} &\geq 0. \end{aligned}$$

Now we can remove $\lambda_{3,2}$ and simplify to

$$\begin{aligned} r &\geq \underline{g}, \\ 0 &\leq \frac{1}{1-b}4 - \frac{1}{b}r - 3\lambda_{3,0}, \\ \frac{3}{2}\frac{1}{1-b} - \frac{3}{2}\lambda_{3,0} &\leq \frac{1}{1-b}4 - \frac{1}{b}r - 3\lambda_{3,0}, \\ \lambda_{3,0} &\geq 0. \end{aligned}$$

After rearranging terms,

$$\begin{aligned} r &\geq \underline{g}, \\ \lambda_{3,0} &\leq \frac{1}{1-b}\frac{4}{3} - \frac{1}{b}r\frac{1}{3}, \\ \lambda_{3,0} &\leq \frac{1}{1-b}\frac{5}{3} - \frac{1}{b}r\frac{2}{3}, \\ \lambda_{3,0} &\geq 0. \end{aligned}$$

Thus, a solution exists if and only if

$$\begin{aligned} r &\geq \underline{g}, \\ 0 &\leq \frac{1}{1-b}\frac{4}{3} - \frac{1}{b}r\frac{1}{3}, \\ 0 &\leq \frac{1}{1-b}\frac{5}{3} - \frac{1}{b}r\frac{2}{3}, \end{aligned}$$

which is equivalent to $\underline{g} \leq r \leq \overline{g}$. □

Appendix B: examples of mechanism-design by a privately informed seller with interdependent values

For another illustration of Kakutani PBE and neo-optimum, let the principal be a seller who proposes a mechanism for determining the probability $q \in [0,1]$ that she receives a unit of an indivisible good, and let $p \in \mathbb{R}$ denote the buyer's paymentm as in the informed-seller example in Dosis (2022). The principal has one of two (non-negative) cost types, $T = \{c_1, c_2\}$. Any belief b can be identified with the probability of the type c_2, that is, $b \in B = [0,1]$. Let $p - c_i q$ $(i = 1, 2)$ denote the principal's payoff if she has the type c_i, sells with probability q, and receives the payment p; the buyer then obtains the payoff $-p + v_i q$, where $v_1 > c_1$, $v_2 > c_2$, and $v_2 > v_1$. If the payments were bounded, the analysis would remain the same and the setting would be a Bayesian incentive problem, implying that a neo-optimum exists (cf. Corollary 2).

A mechanism is a game form in which the seller and the buyer play, and each end node is a probability-payment pair $(q,p) \in [0,1] \times \mathbb{R}$. The mechanism is played if the buyer accepts it. By the revelation principle, given any belief $b \in B$, the mechanism (or, more precisely, the action of proposing the mechanism) implements a probability-payment pair[17] $(q_i(b), p_i(b))$ for each type c_i such that incentive compatibility is satisfied,

$$p_1 - c_1 q_1 \geq p_2 - c_1 q_2 \quad \text{and} \quad p_2 - c_2 q_2 \geq p_1 - c_2 q_1, \tag{38}$$

[17] Given the linearity of the payoff functions, probability distributions over probability-payment pairs need not be considered.

and the buyer's participation constraint is satisfied,

$$(1-b)(v_1q_1 - p_1) + b(v_2q_2 - p_2) \geq 0. \tag{39}$$

Thus,

$$K = \{(b, (p_1 - c_1q_1, p_2 - c_2q_2)) \mid (38), \quad (39)\}.$$

We now distinguish two cases. In the first, $c_1 < c_2$; this is consistent with the interpretation that the seller has a low-quality good if her type is c_1 and has a high-quality good if her type is c_2. In the second case, $c_1 > c_2$; recalling $v_1 < v_2$, this case can be interpreted in terms of fit: if the seller's type is c_1 then the good is a relatively better fit for the seller, and otherwise fits the buyer relatively better.

For any set of payoff vectors S, denote the payoff vectors that are below some payoff vector in S along a 45 degree ray by

$$\text{diagray}(S) = \{U \mid \exists x \geq 0 : \; U + (x, x) \in S\}.$$

As is clear from the constraints (38) and (39), if all payoff vectors in a set S are b-feasible for some $b \in B$, then all payoff vectors in diagray(S) are b-feasible as well because the payments to all types of the seller may be changed by the same amount x.

Consider now the quality-uncertainty case $c_1 < c_2$. After standard manipulations of the constraints (38) and (39), we obtain the following characterization,

$$\begin{aligned} K \;\; = \;\; & \bigcup_{b \in B} \{b\} \times \text{diagray}\left(\text{conv}\left\{((1-b)v_1 + bv_2 - c_1, (1-b)v_1 + bv_2 - c_2)\right),\right. \\ & \left.\left. \qquad\qquad ((1-b)(v_1 - c_1), (1-b)(v_1 - c_1))\right\}\right). \end{aligned}$$

where conv$\{\dots\}$ denotes the line section between two "vertex" payoff vectors. The first of these arises from both types selling with probability 1, and the payment is equal to the buyer's expected valuation. The second vertex payoff vector arises from type c_1 selling with probability 1 and type c_2 selling with probability 0, where the payments are determined such that the incentive constraint of type c_1 (that is, the left constraint in (38)) and the agent's participation constraint (39) are binding (in particular, whenever $b < 1$ so that the buyer is not certain to face type c_2, a seller of this type obtains a payment although she is not selling anything).

As illustrated in Figure 4, the bounding line sections tilt around a common point $\underline{U}$ that is feasible at all beliefs. This point is characterized as follows. The payoff of a type-c_1 seller is maximized across all points in the 0-feasible set $K(0)$, that is, under the constraint that the buyer is sure to face type c_1 and accepts the mechanism. This is achieved via the type-c_1 seller selling the good for sure (i.e., $q_1 = 1$) and have the buyer pay $p_1 = v_1$. The vertical component of the tilting point can be computed by the type-c_2 seller maximizing her payoff across all points in the 1-feasible set that satisfy the incentive constraint of type c_1 given her outcome (q_1, p_1). This is achieved at the outcome (q_2, p_2) such that $-c_1 + v_1 = -q_2 c_1 + p_2$ and $q_2 v_2 - p_2 = 0$, that is,

$$q_2 = \frac{v_1 - c_1}{v_2 - c_1}, \quad p_2 = \frac{v_1 - c_1}{v_2 - c_1} v_2.$$

(In particular, $q_2 < 1$ so that the high-quality seller keeps the good with positive probability.)

By Lemma 1, any Kakutani PBE is above $\underline{U}$. Thus, whenever the 1st vertex does not dominate $\underline{U}$ (that is, at all interior prior beliefs b^* sufficiently close to 0), the point $\underline{U}$ is the unique Kakutani PBE which, by Proposition 1 then also is the unique neo-optimum. Using Corollary 1, we conclude that, for all interior priors, the set of Kakutani PBEs is equal to the points above $\underline{U}$ that are feasible given the prior. Using Remark 1, the 1st vertex is the unique neo-optimum if it dominates $\underline{U}$ strictly. In the non-generic case where a seller of type c_2 is indifferent between $\underline{U}$ and the 1st vertex, the line section connecting these two points is the set of neo-optima.

Consider now the fit-uncertainty case $c_1 > c_2$. After standard manipulations of the constraints (38) and (39), we obtain the following characterization,

$$\begin{aligned} K \;=\; & \bigcup_{b \in B} \{b\} \times \text{diagray}\left(\text{conv}\left\{((1-b)v_1 + bv_2 - c_1, (1-b)v_1 + bv_2 - c_2),\right.\right. \\ & \qquad\qquad\qquad \left.\left.(b(v_2 - c_2), b(v_2 - c_2))\right\}\right). \end{aligned}$$

The first "vertex" payoff vector argument in conv$\{\dots\}$ is the same as in the quality-uncertainty case (i.e., both seller types sell with probability 1 at a price equal to the buyer's expected valuation). The second vertex payoff vector arises from type c_1 selling with probability 0 and type c_2 selling with probability 1, where the payments are determined such that the incentive

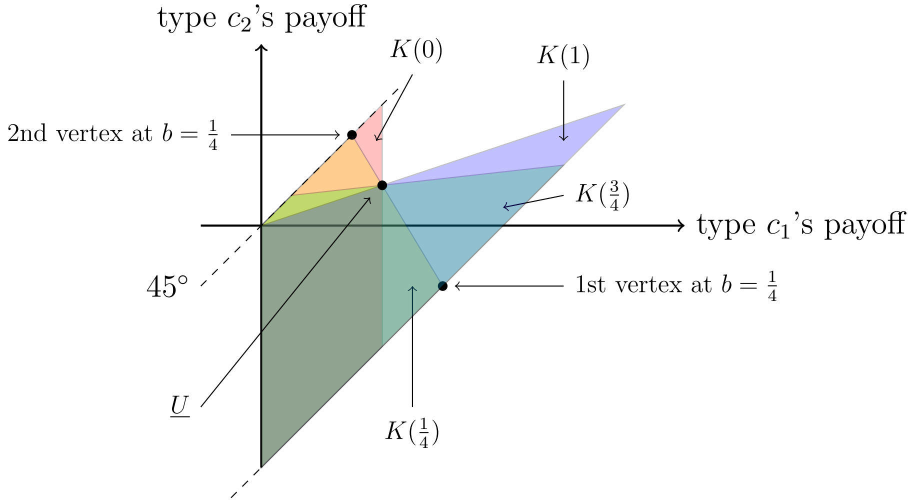

Figure 4: The sets of b-feasible payoff vectors at different beliefs b in the informed-seller example with quality-uncertainty (we omit the parts of the feasible sets that are to the left of the vertical axis). In contrast to the Spence example, the feasibility sets are not nested across different beliefs. Each b-feasibility set is bounded by the line section between the corresponding 1st and 2nd vertices, and includes all points below this line section along 45 degree rays. As b changes, the line sections tilt around a common point $\underline{U}$ that is feasible at all beliefs. (The diagram corresponds to the case $c_1 = 1$, $c_2 = 3$, $v_1 = 2$, $v_2 = 4$.)

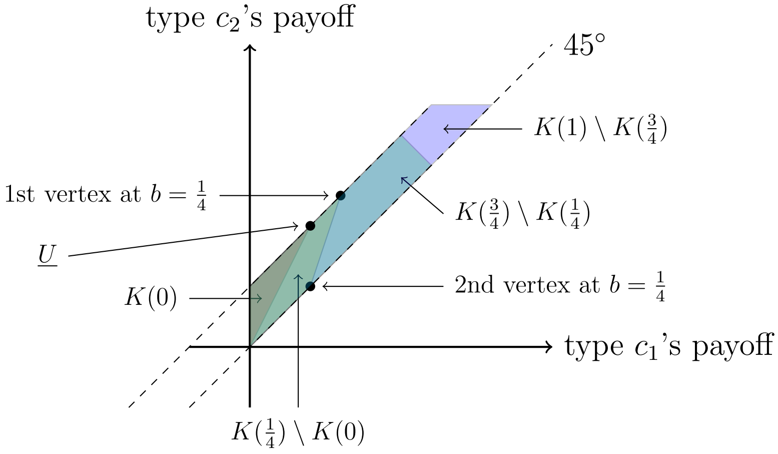
type c_2's payoff
45°
$K(1) \setminus K(\frac{3}{4})$
1st vertex at $b = \frac{1}{4}$
$K(\frac{3}{4}) \setminus K(\frac{1}{4})$
$\underline{U}$
2nd vertex at $b = \frac{1}{4}$
$K(0)$
type c_1's payoff
$K(\frac{1}{4}) \setminus K(0)$

Figure 5: The sets of b-feasible payoff vectors at different beliefs b in the informed-seller example with fit-uncertainty (we omit the parts of the feasible sets that are to the left of the vertical axis). The feasibility sets are nested across different beliefs. Each b-feasibility set is bounded by the line section between the corresponding 1st and 2nd vertices, and includes all points below this line section along 45 degree rays. The largest payoff vector that is feasible at all beliefs is $\underline{U}$. (The diagram corresponds to the case $c_1 = 1$, $c_2 = 0$, $v_1 = 2$, $v_2 = 4$.)

constraint of type c_2 (that is, the right constraint in (38)) and the agent's participation constraint (39) are binding (in particular, whenever $b > 0$ so that the buyer is not certain to face type c_1, a seller of this type obtains a payment although she is not selling anything).

As illustrated in Figure 5, the b-feasibility sets are nested, with a larger probability b of the type c_2 leading to more feasible points. The largest point that is common to all b-feasibility sets, $\underline{U}$, arises from the 1st vertex at $b = 0$, that is, both seller types sell with probability 1 at the price v_1.

Given any interior prior belief b^*, the set of Kakutani PBE is easily characterized. By Lemma 1, any Kakutani PBE is above $\underline{U}$. On the other hand, any b^*-feasible point above $\underline{U}$ is a Kakutani PBE because the "pessimistic" off-path belief $b = 0$ can always be used.

Using Remark 1, the 1st vertex is the unique neo-optimum if it dominates all other b^*-feasible points strictly (which happens for all b^* sufficiently close to 0 and is the case if $b^* = 1/4$ in Figure 5).

For all b^* sufficiently close to 1 (such as in the case $b^* = 3/4$ in Figure 5), there exist multiple b^*-feasible points above $\underline{U}$ that are on the b^*-weak-Pareto frontier. Thus, neither Proposition 1 nor Remark 1 are useful to determine which of these points is a neo-optimum. Similar arguments as in our Spence example show that the 1st vertex is the unique neo-optimum.

We conclude that the 1st vertex, where the good is sold with probability 1 by both types, is the unique neo-optimum at all interior prior beliefs.

It is straightforward to show that the 1st vertex also is the ex-ante optimum. However, given any sufficiently large b^*, a trivial change of the setting will let the ex-ante optimum switch to the 2nd vertex while the neo-optimum remains at the 1st vertex: just assume that the payoff function of the type c_1-seller is instead given by $\alpha(p - c_1 q)$, where α is large; this will stretch Figure 5 horizontally by the factor α. The example is noteworthy because then the 2nd vertex is both a Kakutani PBE and yields a higher ex-ante expected payoff than the 1st vertex, yet the 1st vertex is the unique neo-optimum.

References

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

Balzer, B. (2017): "Collusion in auctions: an informed principal perspective," *University of Technology, Sydney.*

Bergemann, D., and S. Morris (2019): "Information design: A unified perspective," *Journal of Economic Literature*, 57(1), 44–95.

Cho, I.-K., and D. M. Kreps (1987): "Signaling games and stable equilibria," *The Quarterly Journal of Economics*, 102(2), 179–221.

Dosis, A. (2022): "On the informed principal model with common values," *The RAND Journal of Economics*, 53(4), 792–825.

Farrell, J. (1993): "Meaning and Credibility in Cheap-Talk Games," *Games and Economic Behavior*, 5, 514–531.

Izmalkov, S., and F. Balestrieri (2012): "The informed seller problem: The case of horizontal differentiation," *mimeo.*

KOCHERLAKOTA, N. R. (2006): “Advances in dynamic optimal taxation,” *Econometric Society Monographs*, 41, Vol I, ed. Richard Blundell, Chapter 7.

KOESSLER, F., AND V. SKRETA (2016): “Informed seller with taste heterogeneity,” *Journal of Economic Theory*, 165, 456–471.

——— (2019): “Selling with evidence,” *Theoretical Economics*, 14(2), 345–371.

——— (2023): “Informed information design,” *Journal of Political Economy*, 131(11), 3186–3232.

LAFFONT, J.-J. (2000): *Incentives and political economy.* OUP Oxford.

LAFFONT, J.-J., AND J. TIROLE (1993): *A Theory of Incentives in Procurement and Regulation.* MIT Press.

MAILATH, G. J. (1987): “Incentive compatibility in signaling games with a continuum of types,” *Econometrica: Journal of the Econometric Society*, pp. 1349–1365.

MAILATH, G. J., M. OKUNO-FUJIWARA, AND A. POSTLEWAITE (1993): “Belief-Based Refinements in Signalling Games,” *Journal of Economic Theory*, 60(2), 241–276.

MASKIN, E., AND J. TIROLE (1990): “The principal-agent relationship with an informed principal: The case of private values,” *Econometrica*, 58(2), 379–409.

——— (1992): “The principal-agent relationship with an informed principal, II: Common values,” *Econometrica*, 60(1), 1–42.

MYERSON, R. B. (1983): “Mechanism design by an informed principal,” *Econometrica*, 51(6), 1767–1798.

——— (1985): “Analysis of Two Bargaining Problems with Incomplete Information,” in *Game Theoretic Models of Bargaining*, ed. by A. Roth, pp. 59–69. Cambridge University Press.

Mylovanov, T., and T. Tröger (2012): "Informed-principal problems in environments with generalized private values," *Theoretical Economics*, 7(3), 465–488.

——— (2014): "Mechanism Design by an Informed Principal: Private Values with Transferable Utility," *Review of Economic Studies*, 81(4), 1668–1707.

Nishimura, T. (2022): "Informed principal problems in bilateral trading," *Journal of Economic Theory*, 204, 105498.

Pęski, M. (2022): "Bargaining with mechanisms," *American Economic Review*, 112(6), 2044–2082.

Rabin, M. (1990): "Communication between rational agents," *Journal of Economic Theory*, 51(1), 144–170.

Skreta, V. (2009): "On the informed seller problem: optimal information disclosure," *Review of Economic Design*, 15(1), 1–36.

Spence, A. M. (1973): "Job Market Signaling," *The Quarterly Journal of Economics*, 87(3), 355–74.

Tan, G. (1996): "Optimal Procurement Mechanisms for an Informed Buyer," *Canadian Journal of Economics*, 29(3), 699–716.

Tröger, T. (2025): "Optimal testing and social distancing of individuals with private health signals," *Review of Economic Design*, pp. 1–41.

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

Yilankaya, O. (1999): "A note on the seller's optimal mechanism in bilateral trade with two-sided incomplete information," *Journal of Economic Theory*, 87(1), 125–143.

Zhao, X. (2023): "Auction design by an informed seller: A foundation of reserve price signalling," *Canadian Journal of Economics/Revue canadienne d'économique*, 56(3), 1161–1190.