Optimal Technology Design[*]

Daniel F. Garrett, George Georgiadis, Alex Smolin and Balázs Szentes[†]

January 2023

Abstract

This paper considers a moral hazard model with agent limited liability. Prior to interacting with the principal, the agent designs the production technology, which is a specification of his cost of generating each output distribution. After observing the production technology, the principal offers a payment scheme and then the agent chooses a distribution over outputs. We show that there is an optimal design involving only binary distributions (i.e., the cost of any other distribution is prohibitively high), and we characterize the equilibrium technology defined on the binary distributions. Notably, the equilibrium payoff of both players is $1/e$.

JEL classification: D82, D86

Keywords: moral hazard, limited liability, contract theory

[*] We are grateful to Alessandro Bonatti, Florian Ederer, Uli Hege, Harry Di Pei, Luis Rayo, Phil Reny, Dan Spulber, Jean Tirole, as well as to participants at several seminars and conferences for helpful comments. In the initial stages of this project, D. Garrett was based at Toulouse School of Economics, University of Toulouse Capitole. He is grateful for funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 714147). D. Garrett and A. Smolin are grateful for funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program (grant ANR-17-EURE-0010).

[†] D. Garrett: University of Essex, Wivenhoe Park, Colchester CO4 3SQ, U.K., mailto:d.garrett@essex.ac.uk; G. Georgiadis: Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A., mailto:g-georgiadis@kellogg.northwestern.edu; A. Smolin: Toulouse School of Economics, University of Toulouse Capitole, 31080 Toulouse Cedex 06, France, mailto:alexey.v.smolin@gmail.com; B. Szentes: Department of Economics, London School of Economics, London, WC2A 2AE, U.K., mailto:b.szentes@lse.ac.uk.

1 Introduction

A central result in contract theory is that agency rents are a key source of economic welfare. When analyzing environments with asymmetric information, most microeconomic models take the determinants of these agency frictions as given. In hidden-information models, for example, the distribution of types, which determines information rents, is typically treated as exogenous. Similarly, in principal-agent problems with hidden actions, the production technology available to the agent, which governs the principal's cost of implementing various actions, is usually part of the model description. However, if an agent's payoff depends on agency frictions, then he is likely to pursue generating these frictions in a way that enhances his payoff. The goal of this paper is to reconsider the standard limited-liability moral hazard problem and understand how an agent might maximize rents by optimally designing the production technology.

For a potential application where such a problem may arise, consider an entrepreneur who is starting a business, and will eventually need venture capital backing to grow it. Prior to contracting with venture capitalists, he must make a host of choices pertaining to the product, the business model, the product market strategy, and so on.[1] If the venture capitalist has strong bargaining power, the entrepreneur benefits from making choices that exacerbate the moral hazard problem to increase agency rents. Even if there were more profitable alternatives, they may not be considered in the contractual negotiations if the venture capitalist is unaware of them.

Identifying the agent-optimal technology turns out to be useful even in those environments where the agent has no meaningful way of influencing the production technology and, from his viewpoint, it is given exogenously. Indeed, we are able to express predictions regarding surplus sharing in principal-agent models with limited liability that are robust to the set of available technologies. Specifically, we characterize the entire set of payoff combinations in such models which can arise for *some* production technology.[2] The agent-optimal technology corresponds to an extreme point in this set at which the agent's payoff is maximal. This exercise is similar in spirit to that of Bergemann et al. (2015), who characterize the set of consumer and seller payoffs in a model of third-degree price discrimination for some information of the seller.[3]

In the baseline setup, we consider a risk-neutral agent who can choose a production technology

[1] A sizable literature on entrepreneurship examines the decisions of entrepreneurs when developing a business. See Gans et al. (2019) for a recent review.

[2] To be precise, we characterize this set as a function of the largest possible output realization.

[3] See also Garrett (2021) for a related exercise in a setting with moral hazard and adverse selection.

(or "project") before interacting with a principal.[4] A production technology specifies the agent's cost of each output distribution with support contained in $[0, 1]$. That is, the only restriction on the available projects is that output is uniformly bounded. Such a bound may represent a physical constraint and is normalized to one. After observing the agent's project, the principal offers a wage contract, which is a mapping from output realizations to monetary compensation. We assume that the agent has limited liability and hence the payment must be non-negative. Finally, the agent chooses an output distribution at a cost determined by his first-stage choice.

In order to focus on the incentives to generate agency rents we consider a particularly stylized model which abstracts from a number of forces that may have first-order importance in applications. Perhaps most importantly, the agent in our model does not incur any costs in choosing a production technology. In practice, developing a project is likely to require a substantial amount of irreversible investment. In fact, the necessity of such investments may prevent the technology from being renegotiated at the contracting stage. The reason is that, even if both parties are aware of production technologies that are more profitable than the one put forward by the agent, they would not be implemented if modifying the agent's project is too expensive.[5] Since the agent chooses the production technology before the contracting stage, our model is a hold-up problem. The expense of modifications could also prevent the agent from secretly adjusting the technology after contracting to make the chosen output distribution less costly, helping to make the agent's commitment to the technology credible.[6]

We also recognize that, in practice, the agent's ability to shape the production technology may be severely limited. For example, some output distributions may be impossible to generate. The constraints on the set of available technologies are likely to be important determinants of the optimal design. However, even when the space of projects is restricted, our main result holds as long as this space includes the agent-optimal technology. Moreover, in the Discussion Section, we demonstrate that our analysis can accommodate certain moment constraints on the domain of production technologies.

Our first main result is that the optimal project involves only binary distributions on $\{0, 1\}$.[7]

[4] We extend our analysis to environments where the agent is risk averse in the Online Appendix.

[5] Gans et al. (2019) argue that many of the decisions an entrepreneur takes during business development cannot be revisited or renegotiated.

[6] Another consideration that may rule out secret deviations by the agent is monitoring by the principal. The principal may not permit the agent to adjust the project initially proposed, especially if she finds it costly to evaluate all the implications of the adjustments.

[7] While there is some multiplicity of optimal projects, we show in Appendix B that all optimal projects share the same essential attributes, so we write informally of "the" optimal project.

In other words, the cost of all other distributions can be assumed to be so high that the principal never wants to implement them, and the agent would never choose them irrespective of the payment scheme. This means that the equilibrium project can be thought of as a task which yields a positive payoff only if completed. The production technology specifies the cost of each probability of completion. The principal's wage contract can be viewed as a bonus paid for project completion, with payments set to zero otherwise.

Let us explain the optimality of binary projects. Just like in standard moral hazard problems, output plays a dual role in our model. On the one hand, it is the principal's revenue, and on the other hand, it is an informative signal about the output distribution chosen by the agent, which is used by the principal to incentivize the agent. By the Informativeness Principle, if this signal is made less informative, incentivizing the agent becomes more expensive. The key observation is that each binary distribution with support $\{0, 1\}$ can be viewed as a garbling of a distribution with the same mean. Consider now a transformation of each project so that, if the agent incurs a cost of a distribution, output is distributed according to the binary distribution with the same mean. This means that the agent's cost of inducing a given level of expected output remains the same in the transformed project but the principal's cost of implementing it goes up. In this sense, such a transformation exacerbates the moral hazard problem. We show how this observation can be used to replace any project with a binary one for which the agent's payoff is at least as high.

Our second main result is a full characterization of the optimal binary project. In this project, the cost of completing the task with probability $1/e$ is zero. That is, even if the agent incurs no cost, project completion can be achieved with probability $1/e$. In equilibrium, the principal offers a bonus which induces the agent to complete the project with probability one. Furthermore, the principal is indifferent between offering this bonus and anything less than that. This indifference condition pins down the cost to the agent of any probability of success between $1/e$ and one. Since the marginal cost of the success probability is less than one and the maximal output is produced surely, the equilibrium is ex-post efficient. That is, given the equilibrium production technology, the allocation is efficient. It turns out that the optimal project yields an equal split of surplus: both the principal and the agent earn payoff $1/e$.

The first-best social surplus in our model is one since projects that can generate output one at no cost are feasible. Of course, the agent does not choose such a project because then the principal could achieve the maximal output without making any payment. To earn rents, the agent designs the technology so that generating high expected output is artificially costly. In fact, since the equilibrium output is one with probability one, the only source of distortion induced by the optimal design relates to this cost. This might be considered a form of "cost padding", different from others identified in the literature.[8]

As mentioned above, an identifying feature of the optimal project is that the principal is indifferent between implementing a large range of completion probabilities. Let us explain the economic reasoning behind this feature. Note that, since the agent receives the bonus offered by the principal with the probability of completion, her marginal benefit at each completion probability is the bonus. Hence, at the agent's optimal completion probability, the marginal cost equals the bonus. This means that the principal's expected payment to implement a given completion probability is increasing in the marginal cost at that probability, so lowering this marginal cost makes it more attractive to the principal to implement the completion probability. However, if the principal strictly prefers to implement the equilibrium probability of completion to implementing some smaller probabilities, then the marginal costs at these smaller probabilities can be lowered without affecting the principal's equilibrium choice. Since the cost of a completion probability is the integral of the marginal costs of smaller probabilities, such a modification of the project decreases the agent's total cost and thus increases his overall payoff (while the principal is still willing to offer the same bonus that implements the equilibrium completion probability). The agent can improve any project in this way unless the principal is indifferent between implementing any completion probability which has a positive marginal cost.

We demonstrate that our main results remain valid even if the agent is risk averse.[9] In particular, the search for an optimal project can still be restricted to the set of binary projects. Moreover, the optimal binary project is still ex-post efficient; that is, the principal implements completion probability one. The optimal binary project is still characterized by the requirement that the principal must be indifferent between offering the equilibrium bonus and anything less than that. Of course, the equilibrium payoffs of the principal and the agent are no longer $1/e$ and they depend on the agent's concave utility function.

Despite the fact that our model is stylized, we believe that there are a number of lessons emerging from our analysis which may be useful in applications. First, a general message is that the agent in a technology design problem will often commit to a project that is not cost-efficient. By padding costs, he may be able to earn additional rents. Second, there is reason to expect that, where possible,

[8] For example, Averch and Johnson (1962) observe that a regulated firm has incentives to inflate capital costs.

[9] See the Online Appendix.

an agent's optimal technology will be binary. That is, the project essentially specifies a task and the agent either completes it or fails to solve it. More generally, the agent is likely to favor designs where the information content of output regarding his effort is limited. The reason is that this makes it more expensive for the principal to induce the target level of effort. A third observation that we expect to generalize is the importance of principal incentive constraints in the design of the agent's project. The agent's design must dissuade the principal from choosing ungenerous incentive schemes that give the agent little rent. Where such incentive constraints are slack, however, it may be possible for the agent to redesign the project to lower his equilibrium cost.

Finally, we characterize the payoff combinations that can arise in limited-liability moral hazard problems for some exogenously given technology where output is restricted to the interval $[0, 1]$. This follows by first identifying the largest payoff the agent can achieve as a function of a given profit of the principal. The domain of this function is the interval $[0, 1]$ because the principal can always guarantee a nonnegative profit by offering zero wage and she cannot get more than the maximal output. This function is shown to be strictly concave and zero at the boundaries of its domain. We argue that a payoff profile can be generated by some production technology if and only if it lies weakly below this curve.

The rest of the paper is as follows. Next we discuss related literature. Section 2 introduces the model, Section 3 provides our characterization of the agent's optimal project, and Section 4 provides discussion and extensions, including the characterization of the set of possible player payoffs. Appendix A provides proofs not given in the main text and Appendix B discusses the uniqueness of the agent's optimal project. The Online Appendix solves the agent's project design problem when the agent is risk averse.

Literature.— The limited liability model of moral hazard for a risk-neutral agent is a staple of introductory courses on contract theory, where a restriction to binary output (which emerges endogenously in our setting) is often made for tractability. A classic reference for limited-liability moral hazard is Innes (1990), who demonstrates the optimality of simple debt contracts in a model with a continuum of outputs. More recent treatments of moral hazard with limited liability include Poblete and Spulber (2012) for a model with a continuum of outputs and Ollier and Thomas (2013) for a model with binary output, but complicated by the presence of adverse selection.[10]

While the models discussed above feature a risk-neutral agent with limited liability, the al-

[10] Also related are Carroll (2015) who studies a principal facing ambiguity with respect to the agent's technology, Laux (2001) who considers agents with multiple projects, and Jewitt et al. (2008) who explore general bounds on payments beyond limited liability.

ternative friction commonly explored is risk aversion. Seminal work for the moral hazard model with a risk-averse agent includes Mirrlees (1976), Holmstrom (1979), Grossman and Hart (1983), Rogerson (1985) and MacLeod (2003); see Bolton and Dewatripont (2005) and Holmstrom (2017) for comprehensive treatments, and Georgiadis (2022) for a review. More recently, a strand of this literature has focused on models where the agent can shape the entire distribution of output under different assumptions about the cost of distributions (Hebert, 2018; Bonham and Riggs-Cragun, 2021, Georgiadis et al. , 2022, and Mattsson and Weibull, 2022). The focus of the moral hazard literature, then, has been on contract design taking the agent's technology as given. Our paper departs from this approach by viewing the production technology as a choice of the agent, raising the problem of technology design. We are unaware of this kind of problem being posed elsewhere in the moral hazard literature.

The question of project design is also related to work on how the primitive contractual environment affects payoffs in moral hazard problems. A relevant example in our context is the development of the Informativeness Principle by Holmstrom (1979), which was later refined for instance by Chaigneau et al. (2019). These papers clarify how additional information about the agent's action can reduce agency costs for the principal.

Another related paper is Condorelli and Szentes (2020) who study the problem of optimally generating information rents in the context of a bilateral trade model. Before interacting with the seller, the buyer can choose the distribution of her valuation for the seller's good. This choice is observed by the seller before she makes a take-it-or-leave-it offer. It turns out that the equilibrium distribution generates a unit-elastic demand, that is, it makes the seller indifferent between setting any price on its support. This is reminiscent of our optimal binary project which makes the principal indifferent across a range of bonuses.[11] This similarity might explain why, when the buyer is restricted to choose distributions with support in the interval $[0, 1]$, the equilibrium payoffs of the buyer and the seller are also $1/e$. We further elaborate on the relationship between these two results in Section 4.[12]

[11] Such an indifference argument has appeared in the contexts of incentivizing monitoring (Ortner and Chassang (2018)), optimal testing (Perez-Richet and Skreta (2018)) and monopoly pricing in the presence of ambiguity (Bergemann and Schlag (2011)).

[12] Also related, Roesler and Szentes (2017) consider a setting where signals inform an otherwise uninformed buyer of his value, and asks which signal structure yields the highest information rent.

2 Model

We consider a game between a principal (she) and an agent (he), which proceeds as follows. In the first stage, the agent chooses a cost function $c : \mathcal{F} \to \mathbb{R}_+ \cup \{+\infty\}$, where $\mathcal{F}$ denotes the set of CDFs with support on $[0, 1]$. We refer to such a function c as a project. Then, after observing c, the principal offers a payment scheme $w : [0, 1] \to \mathbb{R}_+$, which is restricted to be Borel-measurable.[13] Finally, after observing the offered payment scheme, the agent chooses a distribution $F \in \mathcal{F}$, and output is realized according to F.[14] If the realized output is x then the agent's and principal's payoffs are $w(x) - c(F)$ and $x - w(x)$, respectively. Both parties are expected payoff maximizers.

Notation.— For each $F \in \mathcal{F}$, let μ_F denote the expected value of F, that is, $\mu_F = \int_0^1 x dF(x)$. The set of projects and the set of Borel-measurable payment schemes are denoted by $\mathcal{C}$ and $\mathcal{W}$, respectively. We refer to a triple $(c, w, F) \in \mathcal{C} \times \mathcal{W} \times \mathcal{F}$ as an outcome. Let U and Π denote the expected payoffs of the agent and the principal defined on the outcomes; that is,

$$U(c, w, F) = \int_0^1 w(x) dF(x) - c(F), \text{ and } \Pi(w, F) = \int_0^1 [x - w(x)] dF(x).$$

Equilibrium in a Project.—Loosely speaking, we wish to call a pair $(w, F) \in (\mathcal{W}, \mathcal{F})$ an *equilibrium* in project $c \in \mathcal{C}$ if it satisfies the following two requirements. First, the distribution F is incentive compatible (or a best response) for the agent given project c and the payment schedule w. Second, the payment scheme w is incentive compatible for the principal given project c. Formal statements of the incentive compatibility constraints and the equilibrium definition are provided below.

Optimal Projects.— We define the project c to be *optimal* if there is an equilibrium (w, F) in c such that the agent's payoff in outcome (c, w, F) is larger than in any outcome (c', w', F') such that (w', F') is an equilibrium in c'. Thus, we assess the optimality of a project c that induces multiple equilibria by considering those which give the highest payoff to the agent. This is in line with the approach prevalent in mechanism design, where the designer is permitted to pick the most favorable equilibrium.

Incentive Compatibility.— We say that choosing F is incentive compatible for the agent in the

[13] Non-negativity of payments encodes the limited-liability constraint.

[14] That the agent can choose any distribution on $[0, 1]$ departs from much of the moral hazard literature, where the output distribution is parameterized by a one-dimensional variable called "effort". We view the agent's chosen distribution as synonymous with his action, an approach also taken for instance by Carroll (2015).

subgame (c,w) if

$$U(c,w,F) \geq U(c,w,F') \text{ for all } F' \in \mathcal{F}. \tag{1}$$

To describe the principal's incentive constraint is harder because the agent may not have a best response in a subgame generated by a pair (c,w). In turn, this can make it difficult to assess the profitability of certain deviations. To circumvent this problem, we define the *value of the agent*, $u(c,w)$, in each subgame (c,w), by

$$u(c,w) \equiv \sup_{F\in\mathcal{F}} U(c,w,F).$$

We aim to define the *value of the principal* in a subgame (c,w) by reference to sequences of distributions along which the agent's payoff converges to his value. In general, there may be many such sequences, potentially generating different limit payoffs to the principal. Let $\mathbf{F}^{c,w}$ denote the set of sequences of distributions (F_n) along which the agent's payoff converges to $u(c,w)$. Formally, $(F_n) \in \mathbf{F}^{c,w}$ if and only if $\lim_{n\to\infty} U(c,w,F_n) = u(c,w)$. Then, the principal's value in subgame (c,w) is given as

$$\pi(c,w) \equiv \sup\left\{\limsup_{n\to\infty} \Pi(w,F_n) \quad : \quad (F_n) \in \mathbf{F}^{c,w}\right\}.$$

Evaluating the principal's value by reference to the supremum is again in the spirit of the approach prevalent in mechanism design, where the principal is permitted to pick the most favorable best response of the agent. The principal's incentive compatibility constraint guaranteeing that she offers payment schedule w in project c can be stated as follows: for all $w' \in \mathcal{W}$,

$$\pi(c,w) \geq \pi(c,w'). \tag{2}$$

That is, the principal cannot gain by deviating to a payment schedule w', whether or not the agent has a best response to w'.

Equipped with the incentive compatibility constraints, we are ready to define equilibria formally.

Definition. The pair (w,F) is said to be an *equilibrium in project c* if

(i) the distribution F satisfies (1),

(ii) the payment scheme w satisfies (2), and

(iii) $\Pi(w,F) = \pi(c,w)$.

Observe that Part (iii) requires that the distribution F indeed generates the principal's value in subgame (c, w).

Binary Projects and Linear Contracts.— As mentioned in the Introduction, binary projects play an important role in our analysis. Next, we formally define these projects. We call a distribution in $\mathcal{F}$ *binary* if its support is contained in $\{0, 1\}$. For each $\mu \in [0, 1]$, let B_μ denote the binary CDF which specifies an atom of size μ at one. Note that the mean of B_μ is also μ. Let $\mathcal{B}$ denote the set of binary distributions; that is, $\mathcal{B} = \{B_\mu : \mu \in [0, 1]\}$. We call a project c *binary* if $c(F) = +\infty$ whenever $F \notin \mathcal{B}$.

In each binary project, the principal always finds it optimal to offer a compensation scheme which pays zero if output is zero. So, the optimal payment scheme can be summarized by a single bonus, b, which is paid to the agent if output is one. If the project is binary, the wage at output $x \notin \{0, 1\}$ is irrelevant, so without loss of generality, such a wage contract can be assumed to be *linear*, denoted by w_b, that is, $w_b(x) = bx$ for all x.

If the principal offers a linear contract, w_b, the output distribution affects the payoffs of the agent and the principal only through its mean. That is, whether or not the project is binary,

$$U(c, w_b, F) = \mu_F b - c(F), \text{ and } \Pi(w, F) = \mu_F(1 - b). \tag{3}$$

This implies that, if the principal offers a linear contract to which the agent has a best response, this best response must involve a distribution which is the least costly among those with the same mean.

3 Main Results

This section is devoted to our two main results. In the next section, we show that it suffices to restrict attention to binary projects and, in Section 3.2, we fully characterize an optimal binary project.

3.1 Binary Projects

In this section, we fix a project c^* and an equilibrium (w^*, F^*) in c^*. Our aim is to construct a binary project $\widetilde{c}$ and an equilibrium $(\widetilde{w}, \widetilde{F})$ in $\widetilde{c}$ so that the outcome $(\widetilde{c}, \widetilde{w}, \widetilde{F})$ Pareto dominates the outcome (c^*, w^*, F^*). Since c^* can be an optimal project, this result implies that there exists an optimal project in the class of binary projects.

As explained in the Introduction, the key observation for this result is that an output realization not only determines the principal's payoff, but also serves as an informative signal about the agent's action. If this signal is made less informative in the sense of Blackwell, incentivizing the agent becomes harder for the principal. To see how an output distribution can be made less informative, consider the following garbling: instead of observing output x, the principal observes output one with probability x and output zero otherwise. That is, the garbling of each $F \in \mathcal{F}$ is B_{μ_F}, so the expected output is unaffected. In fact, B_{μ_F} is the least informative garbling of F, as the same transformation can be applied to any other garbling of F which would again result in B_{μ_F}. So, if the principal could contract only on the realization of B_{μ_F} but not on that of F, her wage cost of implementing F would increase. We next explain how this observation can be used to transform the project c^* to a binary one which is more beneficial for the agent.

The idea behind the construction of the binary project, $\widetilde{c}$, is as follows. First, define the agent's cost of any binary distribution to be the cost of the cheapest distribution in project c^* with the same mean. In this binary project, the principal's wage cost of attaining any level of expected output is higher than in project c^*. In fact, the wage cost of generating μ_{F^*} may be so high that the principal prefers to implement a distribution with a lower mean, thus saving on payments to the agent. In this case, the payoffs of both parties can be lower. Therefore, we further modify the binary project by reducing the agent's cost of $B_{\mu_{F^*}}$ so that the principal can implement it at exactly the same wage cost as that of F^* in c^*.

Before stating the main result of this section, let us introduce an additional piece of notation. Note that the expected payment in outcome (c^*, w^*, F^*) is $\mathbb{E}_{F^*}[w^*]$. If $\mu_{F^*} > 0$ (as must be the case if the outcome is optimal for the agent), we can define b^* to equal $\mathbb{E}_{F^*}[w^*]/\mu_{F^*}$. We can then observe that[15]

$$\mathbb{E}_{F^*}[w^*] = \mu_{F^*} b^* = \mathbb{E}_{F^*}[w_{b^*}] = \mathbb{E}_{B_{\mu_{F^*}}}[w_{b^*}]. \tag{4}$$

That is, the expected payment induced by the pair (w^*, F^*) is the same as that induced by $(w_{b^*}, B_{\mu_{F^*}})$.

Proposition 1. *Suppose that (w^*, F^*) is an equilibrium in project c^* with $\mu_{F^*} > 0$. Then there exists a binary project, $\widetilde{c}$, such that*

(i) $(w_{b^*}, B_{\mu_{F^*}})$ is an equilibrium in $\widetilde{c}$,

(ii) $U(c^*, w^*, F^*) \leq U(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$, and

[15] Recall that $w_{b^*}(x) = b^* x$ for all x.

(iii) $\Pi\left(w^{*},F^{*}\right)=\Pi\left(w_{b^{*}},B_{\mu_{F^{*}}}\right)$.

Let us describe the binary project $\widetilde{c}$ and the main arguments in the proof of the proposition. It turns out that, in this binary project, the agent's rent can be ensured by making it hard for the principal to dissuade the agent from deviating downwards (i.e., to distributions with lower means). Upwards deviations need not play a role, so we specify the agent's cost of each B_{μ} with $\mu>\mu_{F^{*}}$ to be infinity throughout the construction. We now explain the two steps of constructing $\widetilde{c}$ from c^{*} in more detail. In the first step, we take the agent's cost of $B_{\mu_{F^{*}}}$ to also be infinite. For each $\mu<\mu_{F^{*}}$, we specify the cost of B_{μ} to be the cost of the cheapest distribution in project c^{*} with expectation μ.[16] We then prove that, in order to achieve any expected output, the principal must make a higher expected payment in this binary project than in c^{*}. In the second step, we redefine the agent's cost of $B_{\mu_{F^{*}}}$ so that the principal's wage cost of implementing $B_{\mu_{F^{*}}}$ is exactly $\mathbb{E}_{F^{*}}\left[w^{*}\right]=\mathbb{E}_{B_{\mu_{F^{*}}}}\left[w_{b^{*}}\right]$, thus obtaining the project $\widetilde{c}$. We show that the agent's cost of $B_{\mu_{F^{*}}}$ in $\widetilde{c}$ is less than $c^{*}\left(F^{*}\right)$. This means that, by Equation (4), Parts (ii) and (iii) of the proposition are satisfied.

We now explain how to obtain Part (i). Note that, by our choice of the agent's cost of the distribution $B_{\mu_{F^{*}}}$ in project $\widetilde{c}$, the agent best responds to $w_{b^{*}}$ by choosing $B_{\mu_{F^{*}}}$. Therefore, to prove that $\left(w_{b^{*}},B_{\mu_{F^{*}}}\right)$ is an equilibrium in this project, we need to demonstrate only that offering $w_{b^{*}}$ is incentive compatible for the principal. By construction, if the principal wants to implement $B_{\mu_{F^{*}}}$, she offers payment schedule $w_{b^{*}}$. She therefore receives a payoff of $\Pi\left(w^{*},F^{*}\right)$. As explained, attaining any other expected output μ ($\mu\neq\mu_{F^{*}}$) is more expensive for the principal in $\widetilde{c}$ than in c^{*}. Therefore, since the principal found it optimal to implement F^{*} in c^{*}, she optimally chooses to implement $B_{\mu_{F^{*}}}$ in $\widetilde{c}$ by offering $w_{b^{*}}$; that is, $w_{b^{*}}$ is incentive compatible in $\widetilde{c}$.

Towards the first step described above, let us define the binary project, $\widehat{c}$, as follows:

$$\widehat{c}\left(B_{\mu}\right)=\begin{cases}\inf\left\{c^{*}\left(F\right):\mu_{F}=\mu\right\} & \text{if } \mu<\mu_{F^{*}},\\ \infty & \text{otherwise.}\end{cases}$$

We next formalize the aforementioned implication of the Informativeness Principle; in particular, we demonstrate that the principal is worse off in $\widehat{c}$ than in c^{*}. In fact, we show that, from the principal's point of view, the transformed project $\widehat{c}$ is worse than being restricted to linear contracts in c^{*} in the sense that each contract w_{b} generates weakly more profit to the principal in project c^{*} than in $\widehat{c}$.

[16] We take the infimum in case no cheapest distribution exists.

Lemma 1. *For all $b \in [0,1]$, $\pi(\widehat{c}, w_b) \leq \pi(c^*, w_b)$.*

Proof. See the Appendix. *QED*

Let us illustrate the argument behind the proof of this lemma for the case where the principal's value in subgame (c^*, w_b), $\pi(c^*, w_b)$, is generated by a best response of the agent. Since the agent's expected payment generated by the linear contract w_b depends only on the expected output, he chooses a distribution only if it is the cheapest among those with the same mean. Therefore, the agent's value in subgame (c^*, w_b) is

$$\sup_{\mu\in[0,1]} \left\{\mu b - \inf\left\{c^*(F) : F \in \mathcal{F},\ \mu_F = \mu\right\}\right\}. \tag{5}$$

This is the same problem as the one which determines the agent's value in the subgame $(\widehat{c}, w_b)$, except that in the latter, the domain is effectively restricted to be $[0, \mu_{F^*})$. Suppose now that F is incentive compatible in (c^*, w_b) and generates the principal's value. That is, μ_F solves the problem in (5) and $\pi(c^*, w_b) = \Pi(w_b, F)$. If $\mu_F < \mu_{F^*}$, then μ_F also solves the agent's problem with the restricted domain, implying that B_{μ_F} is incentive compatible in $(\widehat{c}, w_b)$. In this case, $\pi(\widehat{c}, w_b) = \mu_F(1-b) = \pi(c^*, w_b)$. If $\mu_F \geq \mu_{F^*}$, then the principal's value is at least $\mu_{F^*}(1-b)$ in the subgame (c^*, w_b), so $\pi(c^*, w_b) \geq \mu_{F^*}(1-b) \geq \pi(\widehat{c}, w_b)$, where the second inequality holds because, in project $\widehat{c}$, the agent never chooses a distribution which has mean larger than μ_{F^*}.

We are now ready to define project $\widetilde{c}$. Our aim is to modify $\widehat{c}$ at $B_{\mu_{F^*}}$ so that $\left(w_{b^*}, B_{\mu_{F^*}}\right)$ is an equilibrium in project $\widetilde{c}$. On the one hand, this requires the cost of $B_{\mu_{F^*}}$ to be sufficiently small to guarantee that $B_{\mu_{F^*}}$ is a best response to w_{b^*}. On the other hand, this cost cannot be too small, for otherwise $B_{\mu_{F^*}}$ could be implemented with a bonus smaller than b^*. Therefore, we specify the cost of $B_{\mu_{F^*}}$ to be the largest cost at which the agent still best responds to w_{b^*} by choosing $B_{\mu_{F^*}}$. This cost, denoted by $\overline{c}$, satisfies $\mu_{F^*}b^* - \overline{c} = \sup\left\{\mu b^* - \widehat{c}(B_\mu)\right\}$. The binary project $\widetilde{c}$ is defined as follows:

$$\widetilde{c}(F) = \begin{cases} \overline{c} & \text{if } F = B_{\mu_{F^*}}, \\ \widehat{c}(F) & \text{if } F \neq B_{\mu_{F^*}}. \end{cases}$$

Next, we demonstrate that the outcome $\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right)$ Pareto dominates (c^*, w^*, F^*). To this end, we first argue that the cost of $B_{\mu_{F^*}}$ in project $\widetilde{c}$ is weakly smaller than $c^*(F^*)$. Suppose, for a contradiction, that $\overline{c} > c^*(F^*)$. Then,

$$\mu_{F^*}b^* - c^*(F^*) > \mu_{F^*}b^* - \overline{c} = \sup\left\{\mu b^* - \widehat{c}(B_\mu)\right\} = \sup\left\{\mu_F b^* - c^*(F) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\right\},$$

where the two equalities follow from the definitions of $\overline{c}$ and $\widehat{c}$, respectively. By continuity, this chain implies the existence of $b < b^*$ such that

$$\mu_{F^*} b - c^* \left(F^*\right) > \sup \left\{\mu_F b - c^* \left(F\right) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\right\}.$$

This means that offering the linear contract w_b in project c^* provides the principal with a value at least $\mu_{F^*}\left(1 - b\right)$, which is strictly more than her equilibrium payoff, $\Pi\left(w^*, F^*\right) = \mu_{F^*}\left(1 - b^*\right)$. This contradicts the incentive compatibility of w^* in c^*.

We are now ready to show that the agent is weakly better off in the outcome $\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right)$ than in $\left(c^*, w^*, F^*\right)$. Indeed,

$$U\left(c^*, w^*, F^*\right) = \mu_{F^*} b^* - c^*\left(F^*\right) \leq \mu_{F^*} b^* - \widetilde{c}\left(B_{\mu_{F^*}}\right) = U\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right), \tag{6}$$

where the equalities follow from (4) and the inequality follows from $\widetilde{c}\left(B_{\mu_{F^*}}\right) = \overline{c} \leq c^*\left(F^*\right)$. Also note that (4) implies that the principal's payoffs are the same in these two outcomes:

$$\Pi\left(w^*, F^*\right) = \mu_{F^*} - \mathbb{E}_{F^*}\left[w^*\right] = \mu_{F^*}\left(1 - b^*\right) = \Pi\left(w_{b^*}, B_{\mu_{F^*}}\right). \tag{7}$$

We defined $\widetilde{c}\left(B_{\mu_{F^*}}\right)$ so that the payment schedule w_{b^*} implements $B_{\mu_{F^*}}$ in project $\widetilde{c}$. Next, we confirm that the principal cannot implement $B_{\mu_{F^*}}$ in project $\widetilde{c}$ with any payment schedule w_b such that $b < b^*$. This means that the principal's value from offering w_b in project $\widetilde{c}$ is the same as in project $\widehat{c}$.

Lemma 2. *For all $b \in [0, b^*)$, $\pi\left(\widetilde{c}, w_b\right) = \pi\left(\widehat{c}, w_b\right)$.*

Proof. See the Appendix. *QED*

Let us illustrate the argument of the proof for the case where the agent has a best response to w_{b^*} in project $\widehat{c}$, say $B_{\mu'}$ (with $\mu' < \mu_{F^*}$); that is, $\mu' b^* - \widehat{c}\left(B_{\mu'}\right) = \sup\left\{\mu b^* - \widehat{c}\left(B_\mu\right)\right\}$. By the definition of $\widetilde{c}$, this means that $\mu' b^* - \widetilde{c}\left(B_{\mu'}\right) = \mu_{F^*} b^* - \widetilde{c}\left(B_{\mu_{F^*}}\right)$. Since $\mu' < \mu_{F^*}$, this equality implies that, for all $b \in [0, b^*)$, $\mu' b - \widetilde{c}\left(B_{\mu'}\right) > \mu_{F^*} b - \widetilde{c}\left(B_{\mu_{F^*}}\right)$, implying that $B_{\mu_{F^*}}$ is not incentive compatible in $\left(\widetilde{c}, w_b\right)$. Since the projects $\widetilde{c}$ and $\widehat{c}$ are identical on the rest of their domains, the statement of the lemma follows.

Finally, we are ready to prove Proposition 1.

Proof of Proposition 1. Observe that the outcome $\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right)$ satisfies Parts (ii) and (iii) of the proposition by Equations (6) and (7), respectively. Therefore, we only need to establish Part (i); that is, we need to argue that $(w_{b^*}, B_{\mu_{F^*}})$ is an equilibrium in project $\widetilde{c}$. By the definition of $\overline{c}$, $B_{\mu_{F^*}}$ is incentive compatible in the subgame $(\widetilde{c}, w_{b^*})$. Next, we prove that w_{b^*} is incentive compatible in $\widetilde{c}$.

If $b > b^*$, then

$$\pi(\widetilde{c}, w_b) \leq \mu_{F^*}(1-b) < \mu_{F^*}(1-b^*) = \Pi(w_{b^*}, B_{\mu_{F^*}}),$$

where the first inequality follows from $\widetilde{c}(B_\mu) = \infty$ for all $\mu > \mu_{F^*}$, and the second inequality is implied by $b > b^*$.

If $b < b^*$, then

$$\pi(\widetilde{c}, w_b) \leq \pi(c^*, w_b) \leq \Pi(w^*, F^*) = \Pi(w_{b^*}, B_{\mu_{F^*}})$$

where the first inequality follows from Lemmas 1 and 2, the second one follows from (w^*, F^*) being an equilibrium outcome in project c^*, and the equality is implied by the definition of b^*. *QED*

3.2 Optimal Project

Each binary project, c, can be described by specifying the cost of each probability of success through a function $C : [0,1] \to \mathbb{R}_+$. In particular, we set $C(\mu) = c(B_\mu)$ for all μ, and we keep in mind that the cost of a non-binary distribution is infinity. Recall that, in binary projects, it is without loss of generality to restrict attention to bonus contracts, w_b, where the agent is paid b if output is one. Finally, the agent's choice of distribution B_μ can be identified by its mean, μ (equivalently, by its "completion probability"). In what follows, we describe each binary outcome (c, w_b, B_μ) by a triple (C, b, μ).

We are ready to state the main result of this section.

Proposition 2. *There is an optimal binary project, C^*, and an agent-optimal equilibrium in C^*, (b^*, μ^*), such that*

(i) $C^{*\prime}(\mu) = 1 - 1/(e\mu)$ if $\mu \geq 1/e$ and zero otherwise,

(ii) $b^* = 1 - 1/e$, and

(iii) $\mu^* = 1$.

Proposition 2 describes an optimal binary project in terms of the marginal costs of the possible completion probabilities. Of course, adding a fixed cost has no impact on incentives, so $C^*(0) = 0$.

We explain below that the functional form of the marginal cost in Part (i) is pinned down by the requirement that the principal be indifferent between implementing a large range of completion probabilities. To this end, we sketch an argument for Proposition 2 that restricts the agent to choose cost functions C that are non-decreasing, differentiable, and such that equilibrium can be determined using the first-order approach. The formal proof in the Appendix uses an envelope type argument to determine the agent's payoff and does not rely on any such restrictions.

We view the agent's problem of finding an optimal binary project as a maximization problem subject to incentive compatibility constraints. Our argument can be understood in terms of backward induction. We first determine a condition relating the principal's choice of bonus b to the agent's optimal choice of completion probability μ. We can then determine a condition on the project C for the principal to implement a given completion probability μ. Finally, we consider optimizing the agent's payoff over the project C and the completion probability μ to be implemented by the principal, subject to the constraints determined in the previous steps.

Let us then describe the agent's incentive constraint in a subgame (C, b). Note that, at this stage, the agent's problem is to solve $\max_{\widehat{\mu}\in[0,1]} \{\widehat{\mu}b - C(\widehat{\mu})\}$. Then the requirement on the reward b ensuring the agent chooses a given completion probability μ can be described by the first-order condition

$$b = C'(\mu) . \tag{8}$$

We now turn to the incentive constraint of the principal. In project C, the principal's problem is

$$\max_{\mu\in[0,1], b\in\mathbb{R}_+} \mu(1-b)$$

subject to the constraint that μ and b satisfy Equation (8). Plugging the constraint into the maximand, the principal's problem can be expressed solely in terms of the completion probability she wants to implement; that is, $\max_{\mu\in[0,1]} \mu(1-C'(\mu))$. So, in project C, the principal's choice of μ must satisfy

$$\mu(1-C'(\mu)) \geq \widetilde{\mu}(1-C'(\widetilde{\mu})) , \tag{9}$$

for all $\widetilde{\mu} \in [0,1]$.

We are now ready to state the project selection problem. As anticipated above, we include the completion probability μ as a choice variable alongside the cost function C. This is important because the principal may be indifferent between implementing various completion probabilities, generating different payoffs for the agent. The interpretation of including μ as a choice is that, after designing the project, the agent makes a recommendation to the principal regarding which μ to implement. The agent's first-stage problem can be written

$$\max_{C,\mu,b} \mu b - C(\mu)$$
$$s.t.\ (8)\text{ and }(9).$$

Alternatively, plugging the constraint (8) into the maximand, it can be written as

$$\max_{C,\mu} \mu C'(\mu) - C(\mu) \tag{10}$$
$$s.t.\ (9).$$

Now let us explain that, if $(\widehat{C}, \widehat{\mu})$ solves this problem, the constraint (9) must bind at each $\widetilde{\mu} < \widehat{\mu}$ such that $\widehat{C}'(\widetilde{\mu}) > 0$. To understand this observation, notice first that the agent's cost of $\widehat{\mu}$ must be given by $\widehat{C}(\widehat{\mu}) = \int_0^{\widehat{\mu}} \widehat{C}'(\widetilde{\mu})\, d\widetilde{\mu}$, since $\widehat{C}(0) = 0$.[17] Hence, reducing the marginal cost $\widehat{C}'(\widetilde{\mu})$ for any $\widetilde{\mu} < \widehat{\mu}$ implies a reduction in the cost $\widehat{C}(\widehat{\mu})$. However, the possibility to reduce the marginal costs $\widehat{C}'(\widetilde{\mu})$ is restricted by the constraint (9) evaluated at $(C, \mu) = (\widehat{C}, \widehat{\mu})$. We can conclude that, for the optimum $(\widehat{C}, \widehat{\mu})$, the constraint (9) binds for all $\widetilde{\mu} < \widehat{\mu}$ as long as $\widehat{C}'(\widetilde{\mu}) > 0$, implying that the principal is indifferent between implementing any completion probability below $\widehat{\mu}$ which has a strictly positive marginal cost.

Our next aim is to use this indifference condition to reduce the agent's problem in (10) to a two-dimensional problem by replacing the domain of projects by the principal's possible equilibrium payoffs. To this end, note that if $(\widehat{C}, \widehat{\mu})$ solves (10), then $\widehat{C}'$ can be expressed in terms of the principal's equilibrium payoff, $\widehat{\pi} \equiv \widehat{\mu}\left(1 - \widehat{C}'(\widehat{\mu})\right)$. Indeed, the binding constraint (9) evaluated at $(C, \mu) = (\widehat{C}, \widehat{\mu})$ can be written as

$$\widehat{C}'(\widetilde{\mu}) = \begin{cases} 0 & \text{if } \widetilde{\mu} < \widehat{\pi} \\ 1 - \frac{\widehat{\pi}}{\widetilde{\mu}} & \text{if } \widetilde{\mu} \in [\widehat{\pi}, \widehat{\mu}]\,. \end{cases} \tag{11}$$

Consequently, the agent's problem in (10) can be rewritten as

$$\max_{\widehat{\mu}, \widehat{\pi} \in [0,1]} \widehat{\mu}\left(1 - \frac{\widehat{\pi}}{\widehat{\mu}}\right) - \int_{\widehat{\pi}}^{\widehat{\mu}} \left(1 - \frac{\widehat{\pi}}{\widetilde{\mu}}\right) d\widetilde{\mu}. \tag{12}$$

[17] Note that $\widehat{C}(0) = 0$ follows because, if $\widehat{C}(0) > 0$, we could reduce all costs by this amount, keeping the players' incentives unchanged, but increasing the agent's equilibrium payoff.

To conclude Proposition 2, observe that

$$\widehat{\mu}\left(1-\frac{\widehat{\pi}}{\widehat{\mu}}\right)-\int_{\widehat{\pi}}^{\widehat{\mu}}\left(1-\frac{\widehat{\pi}}{\widetilde{\mu}}\right)d\widetilde{\mu}=\int_{\widehat{\pi}}^{\widehat{\mu}}\frac{\widehat{\pi}}{\widetilde{\mu}}d\widetilde{\mu}=\widehat{\pi}\left[\log\widehat{\mu}-\log\widehat{\pi}\right], \tag{13}$$

which is maximized at $\widehat{\mu}=1$ and $\widehat{\pi}=1/e$. This explains how we obtain Parts (ii) and (iii) of the proposition. In particular, the principal's profit in an optimal outcome is $\pi^*=\mu^*(1-b^*)=1/e$, and since $\mu^*=1$, we have $b^*=1-1/e$. Finally, note that evaluating the right-hand side of (11) at $\widehat{\mu}=1$ and $\widehat{\pi}=1/e$ yields Part (i) of the proposition.

Finally, we compute the payoffs of the agent and the principal in the outcome (C^*, b^*, μ^*). As mentioned above, the principal's equilibrium payoff is $\mu^*(1-b^*)=1/e$. The agent's payoff is pinned down by evaluating (13) at $(\mu^*,\pi^*)=(1,1/e)$, also yielding $1/e$.

4 Discussion

Payoff Possibility Set.—We can use our results to characterize the set of possible payoff combinations which can arise in principal-agent models for some arbitrary production technology. To be more specific, we still consider an environment where the agent is risk neutral and has limited liability. However, the production technology is exogenously given and satisfies the constraint that the largest output cannot exceed one. Moreover, the agent has an outside option we take to be zero. We wish to characterize those payoff profiles which can arise in equilibrium for some production technology. We next show that there exists a production technology where the principal receives $\widehat{\pi}$ and the agent receives $\widehat{u}$ if and only if $\widehat{\pi}\in[0,1]$ and $\widehat{u}\in[0,-\widehat{\pi}\log\widehat{\pi}]$; see Figure 1 for illustration.

First note that the principal's payoff, $\widehat{\pi}$, must be between zero and one. The reason is that the contract w_0 guarantees at least zero profit. Since output is less than one, limited liability implies that the profit cannot exceed one. Next, we identify the frontier of the payoff possibility set. That is, we compute the largest payoff the agent can get if the principal's payoff is $\widehat{\pi}$. By Proposition 1, we know that this largest payoff is achieved in a binary project. Furthermore, recall that in Section 3.2 we rewrote the problem of designing the optimal binary project as a maximization problem with respect to the equilibrium probability of success, $\widehat{\mu}$, and the principal's equilibrium profit, $\widehat{\pi}$; see (12). So the problem of designing the agent-optimal binary technology which generates $\widehat{\pi}$ can be reduced to a similar maximization problem except $\widehat{\pi}$ is treated as a parameter instead of a choice variable. By Equation (13), the agent's maximal payoff is $-\widehat{\pi}\log\widehat{\pi}$.

Since the agent's outside option is zero, his equilibrium payoff cannot be negative. It remains to

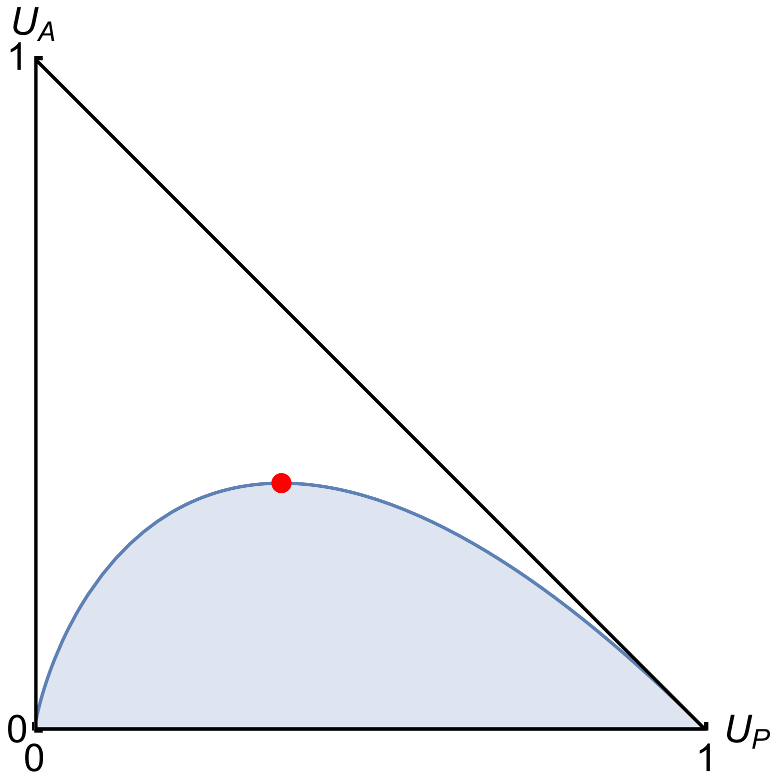

Figure 1: The shaded area represents the set of payoffs for the principal and the agent that arise for *some* technology of the agent. The players' payoffs under the optimal project characterized in Proposition 2 are illustrated by the red dot.

argue that given the principal's payoff, $\widehat{\pi}$, for each $\widehat{u} \in [0, -\widehat{\pi} \log \widehat{\pi})$, there is a production technology so that the equilibrium payoff profile is $(\widehat{\pi}, \widehat{u})$. To do so, consider the production technology generating $(\widehat{\pi}, -\widehat{\pi} \log \widehat{\pi})$ and add a fixed cost of $-\widehat{\pi} \log \widehat{\pi} - \widehat{u}$. That is, the agent's cost of each distribution is increased by this quantity. Adding this fixed cost does not change the agent's incentives, but it lowers his payoff to $\widehat{u}$.

Competition Among Agents.— As is standard in the literature, we maintained the assumption that the principal has full bargaining power and makes a take-it-or-leave-it offer to the agent. The principal's strong bargaining power is often motivated by fierce competition among the agents, which is usually left unmodeled. At first glance, such a justification might be questioned in our setting because competition could limit the rent an agent can earn from contracting with any principal. It is therefore of interest to examine which projects may arise in markets with imperfect competition among agents.

One way to understand competition among agents is to consider embedding our model into a standard search and matching framework in which one-to-one matches occur over time. Each steady state equilibrium in such a model is associated with a continuation value of an unmatched principal. So, when an agent designs a project, he must bear in mind that any principal prefers to remain unmatched whenever her value from contracting with the agent is less than the continuation value from further search. As before, the agent's problem of designing his equilibrium project can be understood in terms of maximizing the expression in Equation (13) by choice of the completion probability and principal expected profits. The difference, however, is that the principal profit, $\widehat{\pi}$, is bounded from below by a constant which makes a principal indifferent between contracting with the agent and remaining unmatched. The solution to that problem has features similar to those of the optimal project described by Proposition 2. In particular, the equilibrium completion probability is one and the principal is indifferent between offering the equilibrium bonus and anything less than that.

Other Constraints on the Technology.— While we believe there are a range of situations where the agent can determine the technology in advance of contracting, in reality the agent's choice of projects would typically be more limited than in our setting where the only constraint is the bounds on output. Such constraints on technology design may ultimately be responsible for the key properties of the selected project. An example is where noise in the environment prevents distributions that put mass at the extremes of the possible output values, ruling out binary projects. Note also that imposing additional constraints on the set of available technologies is related to the possibility that some designs are more costly than others, e.g. binary technologies might be feasible but too costly to set up. We have treated in this paper the case where the agent is only subject to output bounds, and where all designs are costless to the agent. This was intended as a natural first step and helps shed light on the forces at play in the agent's design problem, while future work could examine the choice of projects given different constraints or costly technology choice.

In spite of these observations, the analysis in the paper does already allow determination of the agent's optimal project in some situations where he is more constrained. For instance, suppose that the choice of project c is restricted to come from a set that contains c^*, a project determined to maximize the agent's payoff in our setting with only bounds on output. Then the same project c^* remains optimal in the more constrained setting. An example is where there is, in addition to the output bounds, a lower bound $L(\mu)$ on the cost of generating mean output μ. Recall that C^* is the optimal binary project defined in Proposition 2. Provided $L(\mu) \leq C^*(\mu)$ for all $\mu \in [0, 1]$, then C^* remains an optimal project for the agent.

There are also cases where constraints on project selection bind, but the optimal project follows closely from our analysis. Suppose that, in addition to output being bounded between zero and one, mean output is restricted to be no greater than some $m \in (0,1)$. This constraint does not affect any of the arguments needed to obtain Proposition 1, so there is an optimal project for the agent which is binary.[18] Also, the agent's maximal payoff in an equilibrium where the probability of output one is $\widehat{\mu}$ and the principal earns payoff $\widehat{\pi}$ is still given by the expression in Equation (13), following the arguments in the proof of Proposition 2. An optimal project is then defined by Equation (11) with $\widehat{\mu} = m$ and $\widehat{\pi} = m/e$. There is an equilibrium given this binary project where the principal offers a bonus payment $1 - 1/e$ for output one (and pays zero otherwise), and the agent chooses probability m of output one. We have thus seen that, in the mean-constrained version of the problem, the agent's equilibrium probability of a high output is less than one.

Risk Aversion.— Our results can be generalized to the case where the agent is risk averse. To explain it more formally, let us modify our model so that, if the agent receives payment w and chooses F at cost $c(F)$, then his payoff is $v(w) - c(F)$, where v is an increasing, concave and continuously differentiable function. It can be shown that, even in this case, binary projects remain optimal. More precisely, the following version of the statement of Proposition 1 remains valid. For each project c^* and any equilibrium (w^*, F^*) in c^*, there is a binary project $\widetilde{c}$ with an equilibrium $(\widetilde{w}, B_1)$ such that the outcome $(\widetilde{c}, \widetilde{w}, B_1)$ Pareto dominates (c^*, w^*, F^*). The characterization of optimal binary projects follows the same steps as for the risk-neutral case. The marginal cost of any completion probability $\mu > \pi^*$ is $v(1 - [\pi^*/\mu])$, where π^* is the equilibrium payoff of the principal. These claims are proved in the Online Appendix.

Relation to Another Hold-up Problem.— As mentioned in the Introduction, Condorelli and Szentes (2020) study a different hold-up problem in the context of bilateral trade. In their model, the buyer first chooses the distribution of her valuation for the seller's good. Then the seller, after observing the value distribution, makes a take-it-or-leave-it price offer. It turns out that, when the support of any value distribution must be in $[0,1]$, the buyer's equilibrium CDF, F^*, has the same functional form as $C^{*\prime}$ described in Proposition 2. More precisely, $F^*(v) = 1 - 1/(ev)$ on $(1/e, 1)$. In what follows, we attempt to illuminate this similarity and the relationships between the two models by transforming the buyer's problem into a strategically equivalent one and show that the transformed problem is identical to the constrained maximization problem (10).

First note that by choosing a value distribution, F, the buyer determines his demand curve,

[18] Given a possibly non-binary equilibrium distribution F^*, we showed that agent upward deviations to distributions with means above μ_{F^*} could be ignored in the analysis (recall the construction of the binary project $\widetilde{c}$).

$1-F$. That is, the probability of trade at price p is $1-F(p)$. The idea behind the transformation is that we replace the buyer's choice set by inverse demand curves and assume that the seller sets a quantity instead of a price, i.e., the probability of trade. Since there is a bijection between demand curves and inverse demand curves and the price is pinned down by inverse demand curve for each quantity, the transformed model is strategically equivalent to that of Condorelli and Szentes (2020). Since both the buyer's willingness-to-pay and the probability of trade are in $[0,1]$, the domain as well as the range of the inverse demand curve must also be in $[0,1]$. Let $\mathcal{P}$ denote the set of such inverse demand functions. So, if the buyer chooses $P \in \mathcal{P}$ and the seller sets Q then the price is $P(Q)$. Next, we explain that the buyer's problem can be written as

$$\begin{aligned} &\max_{q\in[0,1],P\in\mathcal{P}} \int_0^Q \left(P(x) - P(Q)\right) dx \\ &\text{s.t. } QP(Q) \geq \widetilde{Q}P\left(\widetilde{Q}\right) \text{ for all } \widetilde{Q} \in [0,1]. \end{aligned}$$

Note that the objective function is the familiar expression for the buyer's payoff from Consumer Theory, for a given inverse demand curve P and quantity Q. In addition, the constraint guarantees that when the buyer's inverse demand curve is P, the seller indeed prefers to set Q to any smaller quantity. Now, observe that replacing P by $1-C'$ in this problem yields (10), that is, the two problems are equivalent. This means that the buyer-optimal inverse demand curve, P^*, is given by $P^*(Q) = 1 - C^{*\prime}(Q)$, where C^* is characterized Proposition 2, that is, $P^*(Q) = 1/[eQ]$ on $[1/e, 1]$. Finally, notice that for each inverse demand curve P, the buyer's value distribution is determined by the following equation: $P(1-F(v)) = v$. Coincidently, applying this formula for P^* yields $F^*(v) = 1 - 1/(ev)$ on $(1/e, 1)$, confirming the main result in Condorelli and Szentes (2020).

Our discussion about the similarity of Proposition 2 to the main result in Condorelli and Szentes (2020) has followed the heuristic argument presented in Section 3.2. This *assumes* the applicability of the first-order approach to derive that the agent's bonus is equal to his marginal cost, i.e., to derive Equation (8). Importantly, the validity of this approach must be verified, and hence the proof of Proposition 2 in the appendix relies on a different argument, one which has no parallel in Condorelli and Szentes.

References

Averch, H. and Johnson, L.L., 1962. Behavior of the firm under regulatory constraint. *American Economic Review*, 52(5), pp.1052-1069.

Bergemann, D. and Schlag, K., 2011. Robust monopoly pricing. *Journal of Economic Theory*, 146(6), pp.2527-2543.

Bergemann, D., Brooks, B. and Morris, S., 2015. The limits of price discrimination. *American Economic Review*, 105(3), pp.921-57.

Bolton, P. and Dewatripont, M., 2005. *Contract theory.* MIT Press.

Bonham, J. and Riggs-Cragun, A., 2021. Contracting on what firm owners value. Working Paper.

Carroll, G., 2015. Robustness and linear contracts. *American Economic Review*, 105(2), pp. 536-563.

Chaigneau, P., Edmans, A. and Gottlieb, D., 2019. The informativeness principle without the first-order approach. *Games and Economic Behavior*, 113, pp.743-755.

Condorelli, D. and Szentes, B., 2020. Information design in the holdup problem. *Journal of Political Economy*, 128(2), pp.681-709.

Gans, J. S., Stern, S. and Wu, J., 2019. Foundations of entrepreneurial strategy. *Strategic Management Journal*, 40(5), pp.736-756.

Garrett, D., 2021. Payoff implications of incentive contracting. *Theoretical Economics*, 16(4), pp. 1281-1312.

Georgiadis, G. 2022. Contracting with Moral Hazard: A Review of Theory & Empirics.

Georgiadis, G., Ravid, D. and Szentes, B., 2022. Flexible Moral Hazard Models. Working Paper.

Grossman, S.J. and Hart, O.D., 1983. An analysis of the principal-agent problem. *Econometrica*, 51(1), pp.7-46.

Hebert, B. 2018. Moral Hazard and the Optimality of Debt. *Review of Economic Studies*, 85 (4), pp. 2214–2252.

Holmstrom, B., 1979. Moral hazard and observability. *Bell Journal of Economics*, 10(1), pp.74-91.

Holmstrom, B., 2017. Pay for performance and beyond. *American Economic Review,* 107(7), pp.1753-77.

Innes, R.D., 1990. Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory*, 52(1), pp.45-67.

Jewitt, I., Kadan, O. and Swinkels, J., 2008. Moral hazard with bounded payments. *Journal of Economic Theory*, 143(1), pp.59-82.

Laux, C., 2001. Limited-liability and incentive contracting with multiple projects. *RAND Journal of Economics*, 32(3), pp.514-526.

MacLeod, W.B., 2003. Optimal contracting with subjective evaluation. *American Economic Review*, 93(1), pp.216-240.

Mattsson, Lars-GÃ¶ran and Jorgen Weibull. 2022. An Analytically Solvable Principal- Agent Model. Available at SSRN 4252495.

Milgrom, P. and Segal, I., 2002. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2), pp.583-601.

Mirrlees, J.A., 1976. The optimal structure of incentives and authority within an organization. *Bell Journal of Economics*, 7(1), pp. 105-131.

Ollier, S. and Thomas, L., 2013. Ex post participation constraint in a principal–agent model with adverse selection and moral hazard. *Journal of Economic Theory*, 148(6), pp.2383-2403.

Ortner, J. and Chassang, S., 2018. Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy*, 126(5), pp.2108-2133.

Perez-Richet, E. and Skreta, V., 2018. Test design under falsification. Working Paper.

Poblete, J. and Spulber, D., 2012. The form of incentive contracts: agency with moral hazard, risk neutrality, and limited liability. *RAND Journal of Economics*, 43(2), pp.215-234.

Roesler, A.K. and Szentes, B., 2017. Buyer-optimal learning and monopoly pricing. *American Economic Review*, 107(7), pp.2072-80.

Rogerson, W.P., 1985. The first-order approach to principal-agent problems. *Econometrica*, 53(6), pp.1357-1367.

Appendix A: Omitted proofs

Proofs of results in Section 3.1

Proof of Lemma 1. Consider a sequence (μ_n), with $\mu_n \in [0, \mu_{F^*})$ for all n, such that

$$u\left(\widehat{c}, w_b\right) = \lim_{n\to\infty} U\left(\widehat{c}, w_b, B_{\mu_n}\right) \text{ and } \pi\left(\widehat{c}, w_b\right) = \lim_{n\to\infty} \Pi\left(w_b, B_{\mu_n}\right).$$

For each $k \in \mathbb{N}$, there exists n_k such that, for all $\mu \in [0,1]$,

$$\mu_{n_k} b - \widehat{c}\left(B_{\mu_{n_k}}\right) + \frac{1}{k} \geq \mu b - \widehat{c}\left(B_{\mu}\right).$$

Equivalently, for all k, and all $\mu \in [0, \mu_{F^*})$,

$$\mu_{n_k} b - \inf\left\{c^*\left(F\right) : \mu_F = \mu_{n_k}\right\} + \frac{1}{k} \geq \mu b - \inf\left\{c^*\left(F\right) : \mu_F = \mu\right\}.$$

For each k, we can pick a distribution F_{n_k} with mean μ_{n_k} such that

$$c^*\left(F_{n_k}\right) < \inf\left\{c^*\left(F\right) : \mu_F = \mu_{n_k}\right\} + \frac{1}{k}.$$

Then, for all k and all $F \in \mathcal{F}$ with mean $\mu_F \in [0, \mu_{F^*})$,

$$\mu_{n_k} b - c^*\left(F_{n_k}\right) + \frac{2}{k} > \mu_F b - c^*\left(F\right). \tag{14}$$

There are then two cases. In the first, the inequality (14) holds for all k and all $F \in \mathcal{F}$ (not only those F with $\mu_F < \mu_{F^*}$). Then

$$u\left(c^*, w_b\right) = \lim_{k\to\infty} U\left(c^*, w_b, F_{n_k}\right)$$

and hence

$$\pi\left(c^*, w_b\right) \geq \lim_{k\to\infty} \Pi\left(w_b, F_{n_k}\right) = \lim_{k\to\infty} \Pi\left(w_b, B_{\mu_{n_k}}\right) = \pi\left(\widehat{c}, w_b\right)$$

as desired. In the second, the inequality (14) fails to hold for some k and some $F \in \mathcal{F}$ with $\mu_F \geq \mu_{F^*}$, which implies

$$u\left(c^*, w_b\right) = \sup\left\{\mu_F b - c^*\left(F\right) : F \in \mathcal{F}\right\} > \sup\left\{\mu_F b - c^*\left(F\right) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\right\}.$$

This means that there is a sequence of distributions in $\mathcal{F}$ along which the agent's payoff converges to his value $u\left(c^{*},w_{b}\right)$ and for which every distribution has mean at least $\mu_{F^{*}}$. By the definition of the principal's value, we have

$$\pi\left(c^{*},w_{b}\right)\geq\mu_{F^{*}}\left(1-b\right)\geq\pi\left(\widehat{c},w_{b}\right),$$

where the second inequality follows because any distribution with mean at least $\mu_{F^{*}}$ is assigned an infinite cost in the project $\widehat{c}$. *QED*

Proof of Lemma 2. Let us fix $b\in[0,b^{*})$. We first show that w_{b} does not implement $B_{\mu_{F^{*}}}$ in $(\widetilde{c},w_{b})$. Suppose for a contradiction that $B_{\mu_{F^{*}}}$ satisfies the agent's incentive constraint in $(\widetilde{c},w_{b})$, that is,

$$\mu_{F^{*}}b-\overline{c}\geq\sup_{\mu<\mu_{F^{*}}}\left\{\mu b-\widetilde{c}\left(B_{\mu}\right)\right\}. \tag{15}$$

Therefore,

$$\overline{c}\leq-\sup_{\mu<\mu_{F^{*}}}\left\{\left(\mu-\mu_{F^{*}}\right)b-\widetilde{c}\left(B_{\mu}\right)\right\}\leq-\sup_{\mu<\mu_{F^{*}}}\left\{\left(\mu-\mu_{F^{*}}\right)b^{*}-\widetilde{c}\left(B_{\mu}\right)\right\}=\overline{c},$$

where the first inequality is just the previous displayed inequality rearranged, the second inequality follows from $b<b^{*}$, and the equality is the definition of $\overline{c}$. Since the farthest left term and the farthest right term are equal in the previous chain, all inequalities must be equalities. Note that the second inequality is an equality only if the supremum in Equation (15) is approached along a sequence of $\mu^{'}$s converging to $\mu_{F^{*}}$. Since $\widetilde{c}\left(B_{\mu}\right)=\widehat{c}\left(B_{\mu}\right)$ whenever $\mu\neq\mu_{F^{*}}$, and since $\widehat{c}\left(B_{\mu}\right)=\infty$ for $\mu\geq\mu_{F^{*}}$, it follows that the supremum of $\mu b-\widehat{c}\left(B_{\mu}\right)$ is approached by the same sequence. Hence, $\pi\left(\widehat{c},w_{b}\right)=\mu_{F^{*}}\left(1-b\right)$. We can conclude that

$$\pi\left(c^{*},w_{b}\right)\geq\pi\left(\widehat{c},w_{b}\right)=\mu_{F^{*}}\left(1-b\right)>\mu_{F^{*}}\left(1-b^{*}\right)=\Pi\left(w^{*},F^{*}\right),$$

where the first inequality follows from Lemma 1, the strict inequality is implied by $b<b^{*}$ and the second equality follows from the definition of b^{*}. This inequality implies that w^{*} is not incentive compatible in project c^{*}, a contradiction.

Since w_{b} does not implement $B_{\mu_{F^{*}}}$ in $\widetilde{c}$, $U\left(\widetilde{c},w_{b},B_{\mu_{F^{*}}}\right)<u\left(\widetilde{c},w_{b}\right)$. We next show that $\mathbf{F}^{\widetilde{c},w_{b}}=\mathbf{F}^{\widehat{c},w_{b}}$.[19] Note that, for each $(F_{n})\in\mathbf{F}^{\widetilde{c},w_{b}}\cup\mathbf{F}^{\widehat{c},w_{b}}$, there exists $K\in\mathbb{N}$ such that $F_{k}\neq B_{\mu_{F^{*}}}$

[19] Recall that $(F_{n})\in\mathbf{F}^{c,w}$ if and only if $\lim_{n\to\infty}U\left(c,w,F_{n}\right)=u\left(c,w\right)$.

if $k > K$. If $(F_n) \in \mathbf{F}^{\widehat{c},w_b}$, it follows from $\widehat{c}\left(B_{\mu_{F^*}}\right) = \infty$. If $(F_n) \in \mathbf{F}^{\widetilde{c},w_b}$, it is implied by $U\left(\widetilde{c}, w_b, B_{\mu_{F^*}}\right) < u\left(\widetilde{c}, w_b\right)$. Since $\widetilde{c}(F) = \widehat{c}(F)$ whenever $F \neq B_{\mu_{F^*}}$, this means that, for each $(F_n) \in \mathbf{F}^{\widetilde{c},w_b} \cup \mathbf{F}^{\widehat{c},w_b}$,

$$\lim_{n\to\infty} U\left(\widetilde{c}, w_b, F_n\right) = \lim_{n\to\infty} U\left(\widehat{c}, w_b, F_n\right),$$

implying that $\mathbf{F}^{\widetilde{c},w_b} = \mathbf{F}^{\widehat{c},w_b}$. Consequently,

$$\begin{aligned}\pi\left(\widetilde{c}, w_b\right) &\equiv \sup\left\{\limsup_{n\to\infty} \Pi\left(w_b, F_n\right) \ : \ (F_n) \in \mathbf{F}^{\widetilde{c},w_b}\right\} \\ &= \sup\left\{\limsup_{n\to\infty} \Pi\left(w_b, F_n\right) \ : \ (F_n) \in \mathbf{F}^{\widehat{c},w_b}\right\} = \pi\left(\widehat{c}, w_b\right).\end{aligned}$$

QED

Proof of Proposition 2

This section proves Proposition 2. We begin by considering an arbitrary binary project C. Let us drop the dependence on C and write the value for the agent when the bonus is b as $u(b)$. The value is given by

$$u(b) = \sup_{\mu\in[0,1]} \left\{b\mu - C(\mu)\right\}.$$

Note that u is non-decreasing. Moreover, as the upper envelope of linear functions, it is convex and hence continuous.

Let $\Gamma(b)$ be the set of values μ such that there is a sequence (μ_n) with $\mu_n \to \mu$ and $b\mu_n - C(\mu_n) \to u(b)$. Take $\bar{\mu}(b) = \max \Gamma(b)$, and note that the maximum is attained. Similarly, let $\underline{\mu}(b)$ be the minimum of $\Gamma(b)$ (also attained). Note that, if the principal offers a bonus $b \in [0,1]$ in project C, then she obtains value $\bar{\mu}(b)(1-b)$.

For any $b \geq 0$, let $u'_+(b)$ be the right derivative of u at b. For any $b > 0$, let $u'_-(b)$ be the left derivative of u at b. We next show a result that is analogous to Theorem 1 of Milgrom and Segal (2002), but adjusted for the possibility that the agent's payoff $u(b)$ is not attained by values $\mu \in \Gamma(b)$.

Lemma 3. *For all $b \geq 0$, $u'_+(b) \geq \bar{\mu}(b)$. For all $b > 0$, $u'_-(b) \leq \underline{\mu}(b)$.*

Proof. Consider a sequence (μ_n) with $\mu_n \to \bar{\mu}(b)$ and $b\mu_n - C(\mu_n) \to u(b)$. Then, for any $b' > b$, we have

$$(b' - b)\,\bar{\mu}(b) = \lim_{n\to\infty} \left\{b'\mu_n - C(\mu_n)\right\} - \lim_{n\to\infty} \left\{b\mu_n - C(\mu_n)\right\} \leq u(b') - u(b).$$

Dividing by $b' - b$ and taking limits as b' approaches b from above yields $u'_+ (b) \geq \bar{\mu} (b)$.

Let $b > 0$ and consider a sequence (μ_n) with $\mu_n \to \underline{\mu}(b)$ and $b\mu_n - C(\mu_n) \to u(b)$. For any $b' < b$, we have

$$(b - b') \underline{\mu} (b) = \lim_{n\to\infty} \{b\mu_n - C(\mu_n)\} - \lim_{n\to\infty} \{b'\mu_n - C(\mu_n)\} \geq u(b) - u(b').$$

Dividing by $b - b'$ and taking limits as b' approaches b from below yields $u'_- (b) \leq \underline{\mu} (b)$. *QED*

We can further use the convexity of u to determine its right derivative in terms of the completion probability attainable with a given bonus.

Lemma 4. *For all $b \geq 0$, $u'_+ (b) = \bar{\mu} (b)$.*

Proof. Fix $b \geq 0$ and suppose for a contradiction that $u'_+ (b) > \bar{\mu} (b)$. By convexity of u and the previous lemma

$$\bar{\mu} (b) < u'_+ (b) \leq u'_- (b') \leq \underline{\mu} (b')$$

for all $b' > b$. For each $n \in \mathbb{N}$, let

$$b_n \in \left(b, b + \frac{1}{n}\right)$$

and let $\mu_n \in \left[\frac{\bar{\mu}(b)+u'_+(b)}{2}, 1\right]$ and such that

$$b_n\mu_n - C(\mu_n) > u(b_n) - \frac{1}{n}$$

(that such a choice is possible follows because $\frac{\bar{\mu}(b)+u'_+(b)}{2} < u'_- (b_n) \leq \underline{\mu} (b_n)$ for all n). Consider a subsequence (b_{n_k}) such that $\mu_{n_k} \to \mu^* \geq \frac{\bar{\mu}(b)+u'_+(b)}{2}$ for some μ^*. We have

$$\lim \{b\mu_{n_k} - C(\mu_{n_k})\} = \lim \{b_{n_k}\mu_{n_k} - C(\mu_{n_k})\} = \lim u(b_{n_k}) = u(b)$$

where the final equality follows by continuity of u. The fact that $\mu^* > \bar{\mu} (b)$ contradicts the definition of $\bar{\mu} (b)$. *QED*

Note now that, because u is convex, it is absolutely continuous and hence differentiable almost everywhere. This means that

$$u(b) = u(0) + \int_0^b \bar{\mu} (s)\, ds. \tag{16}$$

It is immediate from the agent's problem that we must have $u(0) \leq 0$; i.e., the agent cannot obtain a strictly positive payoff if the bonus is set to zero.

Consider now a project C with an equilibrium in which the principal offers bonus $\widehat{b}$ for project completion, the agent chooses completion probability $\widehat{\mu}$, and therefore the principal's payoff is given by $\widehat{\pi} = \widehat{\mu}(1 - \widehat{b})$. Note that, if $\left(C, \widehat{b}, \widehat{\mu}\right)$ is an optimal outcome for the agent, then we must have $\widehat{\mu} > 0$. Incentive compatibility of the principal offering bonus $\widehat{b}$ requires that, for all b,

$$\begin{aligned} \widehat{\pi} &\geq \bar{\mu}(b)(1-b) \\ &= u'_+(b)(1-b). \end{aligned} \tag{17}$$

Hence, if the agent is to get positive rent in outcome $\left(C, \widehat{b}, \widehat{\mu}\right)$, we must have also $\widehat{\pi} > 0$. Assume from now on that $\widehat{\mu}, \widehat{\pi} > 0$.

Now let us determine the highest agent value, across projects C, that can occur for an equilibrium in which the principal offers bonus $\widehat{b}$ for completion and the agent chooses completion probability $\widehat{\mu}$. Consider then the problem of maximizing the agent's equilibrium payoff

$$u(\widehat{b}) = u(0) + \int_0^{\widehat{b}} u'_+(b)\, db$$

by choice of convex function $u : \mathbb{R}_+ \to \mathbb{R}$ satisfying (i) $u(0) \leq 0$, and (ii) $\widehat{\pi} \geq u'_+(b)(1-b)$ for all b. The first requirement reflects the above observation that the agent cannot obtain a positive payoff if the bonus is zero. The second condition is a re-statement of Condition (17). Any solution to this problem involves $u(0) = 0$ and

$$u'_+(b) = \frac{\widehat{\pi}}{1-b}$$

for all $b \in [0, \widehat{b}]$. In other words, the constraint (ii), or equivalently (17), holds with equality over $b \in [0, \widehat{b}]$ (in which case, the principal must obtain payoff $\widehat{\pi}$ from all such bonuses b). The agent's value function is therefore given on $[0, \widehat{b}]$ by

$$u(b) = \int_0^b \frac{\widehat{\pi}}{1-z} dz. \tag{18}$$

Now, recall that $\widehat{\pi} = \widehat{\mu}(1 - \widehat{b})$, or $\widehat{b} = 1 - \frac{\widehat{\pi}}{\mu}$. The agent's equilibrium payoff can then be written as

$$\int_0^{1-\widehat{\frac{\pi}{\mu}}} \frac{\widehat{\pi}}{1-z} dz = \left[-\widehat{\pi} \log(1-z)\right]_0^{1-\widehat{\frac{\pi}{\mu}}} = \widehat{\pi}\left(\log(\widehat{\mu}) - \log(\widehat{\pi})\right) \tag{19}$$

which is the expression given in Equation (13) (hence establishing also the one in Equation (12)). As explained in the main text, this payoff is maximized across feasible equilibrium values of $\widehat{\mu}$ and $\widehat{\pi}$ by $\widehat{\mu} = 1$ and $\widehat{\pi} = \frac{1}{e}$. The corresponding equilibrium bonus must be $\widehat{b} = 1 - 1/e$.

Note then that, if the project is C^* as given in the proposition (with $C^*(0) = 0$, as explained in the main text), and the principal offers any $b \in [0, 1 - 1/e]$, the agent best responds by choosing μ such that

$$b = 1 - \frac{1}{e\mu},$$

i.e. $\mu = \frac{1}{e(1-b)}$. All such bonuses therefore generate profit $1/e$ for the principal. Hence, it is indeed an equilibrium of project C^* for the principal to offer bonus $b^* = 1 - 1/e$, and the agent to choose completion probability equal to $\mu^* = 1$. This completes the proof of the proposition.

QED

Appendix B: Discussion of uniqueness of optimal projects

The optimal project, C^*, described in Proposition 2 is not unique. To see this, note that C^* can be arbitrarily modified at completion probabilities strictly below $1/e$. Since $C^*(1/e) = 0$, any of these probabilities are weakly dominated by $1/e$, so the modified project is still optimal. The goal of this appendix is to argue that optimal projects differ only in non-essential ways.

Uniqueness of the Optimal Binary Project.— Recall that the optimal project C^* in Proposition 2 is determined via the principal's incentive constraint, (9), evaluated now at $(C, \mu) = (C^*, \mu^*)$ and taken to hold with equality for completion probabilities $\widetilde{\mu}$ above π^*. This means that the principal can also generate her equilibrium payoff by setting any bonus smaller than b^*. Appendix A showed that these conclusions are valid for each optimal binary project. We also showed that, in any optimal binary project, the agent's optimal payoff is determined by an equilibrium which satisfies Parts (ii) and (iii) of the proposition. We now state formally these observations. We abuse notation and write the players' values, u and π, as functions of a binary outcome.

Remark 1. In any optimal binary project C,

(i) $(b^*, \mu^*) = (1 - 1/e, 1)$ is an agent-optimal equilibrium,

(ii) for all $b \in [0, 1 - 1/e]$, $u(C, b) = \int_0^b 1/[e(1-z)]\,dz$, and

(iii) for all $b \in [0, 1 - 1/e]$, $\pi(C, b) = 1/e$.

Proof of Remark 1. The proof of Remark 1 recalls the proof of Proposition 2 in Appendix A.

Part (i) of the remark follows because the expression in Equation (19) represents the agent's highest possible expected payoff in *any* project in which the completion probability is $\widehat{\mu}$ and the principal's expected payoff is $\widehat{\pi}$. Moreover, it is uniquely maximized by $\widehat{\mu} = \mu^* = 1$ and $\widehat{\pi} = \pi^* = 1/e$. Part (ii) of Remark 1 then follows from Equation (18). Part (iii) of Remark 1 follows because the constraint (17) holds with equality in an optimal binary project over the relevant range of bonus payments. That is, because

$$\pi^* = \bar{\mu}(b)(1-b)$$

for all $b \in [0, 1 - 1/e]$, where recall $\bar{\mu}(b)(1-b)$ coincides with the principal's value $\pi(C, b)$ in any subgame (C, b).

QED

We now argue that Remark 1 implies optimal binary projects are "close" to uniquely determined. We begin with a remark which compares any optimal binary project to the project C^* of Proposition 2.

Remark 2. In any optimal binary project C,

(i) for all $\mu \in [0, 1]$, $C(\mu) \geq C^*(\mu)$, and

(ii) for all $\mu \in [1/e, 1]$, there is a sequence (μ_n) with $\mu_n \to \mu$ and $C(\mu_n) \to C^*(\mu)$.

Proof of Remark 2. To show Part (i) of Remark 2, consider the optimal project C^* in Proposition 2. Notice that, as we reduce the bonus b from b^* to zero, the agent's best response in the subgame (C^*, b) decreases continuously from one to zero. Suppose now that C is a project with $C(\mu') < C^*(\mu')$ for some μ', and let $b' \in (0, b^*]$ be the bonus that implements μ' in project C^*. Then we have

$$u(C, b') \geq b'\mu' - C(\mu') > b'\mu' - C^*(\mu') = u(C^*, b') = \int_0^{b'} 1/[e(1-z)]\,dz.$$

The first equality follows because μ' is a best response in the subgame (C^*, b'). The second equality follows because C^* is an optimal project, and by Part (ii) of Remark 1. Hence, by Part (ii) of Remark 1, C is not an optimal project.

To show Part (ii) of Remark 2, fix an optimal binary project C and consider a completion probability $\mu \in [1/e, 1]$. By Part (iii) of Remark 1, it can be attained by a bonus $b = 1 - 1/(\mu e) \in [0, 1 - 1/e]$. Formally, we mean that there exists a sequence (μ_n) convergent to μ with $\mu_n b - C(\mu_n) \to u(C, b)$. Note that the bonus b also implements μ in project C^*. By Part (ii) of Remark 1, the agent's value in subgame (C, b) is the same as in subgame (C^*, b); that is, $u(C, b) = u(C^*, b)$. In particular, considering the aforementioned sequence (μ_n), we have

$$\mu_n b - C(\mu_n) \to u(C^*, b) = \mu b - C^*(\mu),$$

implying $\lim_{n\to\infty} C(\mu_n) = C^*(\mu)$, which is what we wanted to show.

QED

We argue that, in spite of possible differences between any optimal binary project C and the project C^* of Proposition 2, such projects can be viewed almost equivalently from the agent's perspective. First, note that Remark 2 admits that some completion probabilities in $[1/e, 1]$ may be more costly under C than C^*. In this case, however, there are arbitrarily close probabilities which are as affordable to the agent as in C^*, or for which the difference in costs is negligible (this follows by Part (ii) of Remark 2). In addition, it turns out that, for any optimal binary project C, the specification of costs for probabilities below $1/e$ is irrelevant. The reason is that, using Part (ii) of Remark 2, the agent can generate a completion probability at least $1/e$ at negligible cost.[20] Hence, the agent's value can be approached by completion probabilities at least $1/e$ irrespective of the bonus.

We now formalize further the equivalence of optimal binary projects C on $[1/e, 1]$. Note first that if C is restricted to be continuous, then C is unique on $[1/e, 1]$ and equal to C^* on this interval by Part (ii) of Remark 2. For cost functions C that are not continuous, we now demonstrate formally that the completion probabilities μ that are relevant for the agent's problem are only those for which the cost is close to the one given by C^*.

To achieve our goal we let, for any $\varepsilon > 0$,

$$P(\varepsilon) \equiv \{(\mu, C(\mu)) : \mu \in [1/e, 1],\ C(\mu) - C^*(\mu) \leq \varepsilon\}.$$

Suppose the agent can choose only probabilities μ with $(\mu, C(\mu)) \in P(\varepsilon)$, and suppose the associated costs are $C(\mu)$. Then $P(\varepsilon)$ describes the agent's technology in project C but after removing his ability to choose probabilities μ that we anticipate being redundant, either because they are less than $1/e$ or because their costs exceed $C^*(\mu)$ by more than ε. By Parts (i) and (ii) of Remark 2, for a fixed bonus payment b, the agent's value is the same for the restricted technology $P(\varepsilon)$ as if

[20] Formally, there is a sequence (μ_n) that approaches $1/e$ from above, and for which $C(\mu_n) \to 0$. This follows from Part (ii) of Remark 2 and the continuity of C^*.

the agent could choose any completion probability with a cost specified by C. In fact, the feature of $P(\varepsilon)$ that is relevant in determining the agent's value is the lower boundary of its closure. For any $\varepsilon > 0$, any $\mu \in [1/e, 1]$, we have

$$\min\{y : (\mu, y) \in \mathrm{cl}(P(\varepsilon))\} = C^*(\mu),$$

which again can be seen directly from Parts (i) and (ii) of Remark 2.[21] This demonstrates a further sense of equivalence between C and C^*.

Uniqueness beyond binary projects.— The above discussion describes a qualified sense in which optimal binary projects are uniquely determined. Still, the possibility of optimal but non-binary projects may also be a source of non-uniqueness. Nonetheless, we show that properties analogous to those described above continue to hold, even among non-binary projects.

We first explain that output is one in any agent-optimal equilibrium of any optimal project. To this end, let c^* be an optimal project and (w^*, F^*) an agent-optimal equilibrium in c^*. Proposition 1 applied to the optimal outcome (c^*, w^*, F^*) implies that the corresponding binary outcome $(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$ is also optimal. Then, Part (i) of Remark 1 implies that we must have $\mu_{F^*} = 1$, and hence $F^* = B_1$.

Next we argue that output realizations in $(0, 1)$ are redundant in the sense that replacing these realizations by output zero has no impact on equilibrium behavior. To this end, we first show that it can be assumed that $w^*(x) = 0$ for all $x < 1$. The intuition is that if the principal wants to implement output one, she should not reward the agent for any other output realization by offering a positive payment. To state it formally, let us define the payments scheme, ρ_b, for each b such that $\rho_b(1) = b$ and $\rho_b(x) = 0$ for $x \neq 1$. Observe that replacing w^* by $\rho_{w^*(1)}$ makes choosing B_1 no less attractive to the agent, so $\left(\rho_{w^*(1)}, B_1\right)$ is also an agent-optimal equilibrium in c^*.

Provided that the compensation scheme is $\rho_{w^*(1)}$, when the agent is contemplating choosing a distribution, all that matters is the probability that output is one. So, moving all the probability mass from $(0, 1)$ to zero has no impact on the agent's choice of a distribution. Moreover, these new distributions generate smaller expected outputs, hence the principal still prefers to implement B_1. To make these claims precise, let us define a binary project $\widetilde{C}$ such that $\widetilde{C}(\mu) = \inf\{c^*(F) : \Delta(F) = \mu\}$, where $\Delta(F)$ denotes the atom at one specified by F.[22] Given the payment schedule $\rho_{w^*(1)}$, the change in cost function from c^* to $\widetilde{C}$ does not affect the agent's

[21] Here, "cl(·)" refers to the closure of the set.

[22] That is, $\Delta(F) = F(1) - \lim_{x \nearrow 1} F(x)$.

willingness to choose completion probability one. We conclude that the binary project $\widetilde{C}$ is optimal and $(w^*(1), 1)$ is an agent-optimal equilibrium in $\widetilde{C}$.

To give a further sense in which the results in the previous section are robust to the considerations of non-binary projects, we state the following.

Remark 3. If c^* is an optimal project and (w^*, F^*) is an agent-optimal equilibrium in c^*, then

(i) $F^* = B_1$,

(ii) $\Pi(w^*, F^*) = U(c^*, w^*, F^*) = 1/e$,

(iii) $u(c^*, \rho_b) = \int_0^b 1/[e(1-z)]\, dz$ for all $b \in [0, 1-1/e]$, and

(iv) $\pi(c^*, \rho_b) = 1/e$ for all $b \in [0, 1-1/e]$.

We have already established Part (i) and that the binary outcome $(\widetilde{C}, w^*(1), 1)$ is agent-optimal. So, by Part (iii) of Remark 1, $w^*(1) = 1-1/e$, and the principal's payoff in project c^* must be $1/e$, establishing Part (ii). By the construction of $\widetilde{C}$, $u(c^*, \rho_b)$ is equal to $u(\widetilde{C}, b)$ for all b. Optimality of $\widetilde{C}$ and Part (ii) of Remark 1 then imply Part (iii) of Remark 3. Part (iii) of Remark 1 implies $\pi(\widetilde{C}, b) = 1/e$ for any $b \in [0, 1-1/e]$ and hence $\pi(c^*, \rho_b) \geq 1/e$. Part (ii) of Remark 3 then implies $\pi(c^*, \rho_b) = 1/e$ for all $b \in [0, 1-1/e]$ establishing Part (iv).

An interpretation of Remark 3 is that the conclusions reached in Remark 1 remain valid even when projects are not restricted to be binary. For instance, consider any optimal project c^* and an optimal binary project C. The players' values are the same across both projects for any payment schedule ρ_b, $b \in [0, 1-1/e]$, as follows from Parts (iii) and (iv) of Remark 3. This can be compared to Parts (ii) and (iii) of Remark 1.

Online Appendix to Optimal Technology Design: Risk-averse Agent

Daniel Garrett, George Georgiadis, Alex Smolin, and Balázs Szentes

January 2023

Abstract

This appendix solves the agent's project design problem when the agent is risk averse, and where the agent's risk preferences are known to the principal. We build on several results provided in the main document. We show that the agent still finds it optimal to choose a binary project, and we characterize the optimal binary project.

A Risk-averse agent

We now study the natural possibility that the agent is risk averse, with a concave utility over payments. In particular, while the principal is still risk neutral and has the same preferences as in the paper, the agent has a payoff $v(w) - c(F)$ where w is the payment, v is a concave utility function, c represents the agent's technology (i.e., his cost function), and F is the agent's choice of output distribution.

Establishing the optimality of a binary project for the agent is more challenging in this environment for the following reason. The idea of the proof of Proposition 1 for a risk-neutral agent was that garbling output to determine binary output distributions with the same mean makes the agent more difficult to incentivize, so the agent receives a (weakly) higher expected payment to generate the same expected output. Using this idea, we showed that, starting with an arbitrary project c^*, it is possible to find a binary project in which the principal implements the same expected output and the agent is better off. With a risk-averse agent, garbling output still increases the expected payments that must be made to incentivize the agent. However, because the associated payments may be more risky, risk aversion could leave the agent worse off than without the garbling. The argument required to establish the optimality of binary technologies is therefore more complicated. It involves first studying optimal binary projects, and then combining the insights from this analysis with arguments that are similar in spirit to the proof of Proposition 1.

Given the structure of our arguments, we begin our analysis in Section A.2 below by characterizing optimal binary pojects, and determining the relationship between maximum agent payoffs and the profits that are attainable by the principal. Section A.3 then makes use of this analysis in establishing that binary projects are optimal.

We recap at the end of Section A.3 how this online appendix demonstrates the conclusions reported in Section 4 of the main text, under the heading "Risk Aversion". First, the arguments in Section A.3 will show that, for each project c^* and any equilibrium (w^*, F^*) in c^*, there is a binary project and equilibrium in that project in which (a) the agent generates output one with certainty, and (b) the outcome Pareto dominates the outcome in the original equilibrium. Moreover, if F^* does not put all its mass on output one, we will see that the Pareto improvement is strict, a conclusion which in turn implies that binary projects are optimal. In Section A.2, we will see that, in an optimal binary project, where the principal has some equilibrium payoff π^*, the cost function is defined by a cost of zero for generating output one with probability π^* and a marginal cost $v\left(1 - \pi^*/\mu\right)$ for probabilities μ of ouput one with $\mu > \pi^*$.

A.1 Preliminaries

The setting is the same as in the main document, except we generalize to allow that the agent's payoff is given by $v\left(w\right) - c\left(F\right)$, where $w \geq 0$ is any payment to the agent, F is any distribution on $[0, 1]$, and $v : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing, weakly concave and continuously differentiable function. For any $w \geq 0$, we refer to $v\left(w\right)$ as the agent's "felicity of consumption" from consumption w. We adopt the normalization that $v\left(0\right) = 0$. The inverse of v is given by the function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$. A central objective will be to establish the optimality of binary projects for a risk-averse agent, as follows immediately from Proposition A1 below. We characterize optimal binary projects in the process.

Due to the agent's risk aversion, lotteries over payments are no longer payoff equivalent to deterministic payments with the same mean. In fact, the principal always finds it optimal to choose payments that are deterministic conditional on output. This is because the agent's incentives are determined by the expected utility conditional on output, and a given level of expected utility is most cheaply provided (for the risk-neutral principal) by a deterministic payment. It is then easy to see that considering only equilibria in which the principal offers deterministic payments comes at no loss in the agent's project design problem.

In spite of the previous observation, the proof below makes use of payment schemes that are random conditional on output. We find this convenient in order to mimic the logic of Lemma 1 in the main document. Lemma 1 considered payments offered in a possibly non-binary original project c^* that were linear in output, thus rendering the agent's payoff linear in output. Given the agent's risk aversion, such linearity of the expected payoff in output can be obtained by considering instead payments that are randomized over a binary support, with the probability of the higher payment linear in output.

Let us then introduce the notation for random payments up front, while the details of how they are used in the argument appear in Section A.3 below. For any output $x \in [0,1]$, we let the random payment conditional on x be determined by a cdf $G_x : \mathbb{R}_+ \to [0,1]$, where $G_x\left(w\right)$ is the probability of a payment no greater than w conditional on output realization x. A collection of distributions is denoted $\mathcal{G} = \left(G_x\right)_{x\in[0,1]}$, which is the representation of the payment schedule when random payments conditional on output are permitted.

The notation introduced in Section 2 of the main document updates straightforwardly to random payments and a risk-averse agent. The agent's expected payoff in project c, when choosing output distribution F, given payments $\mathcal{G}$, is

$$U\left(c,\mathcal{G},F\right) = \int_0^1 \int_0^\infty v\left(\widetilde{w}\right) dG_x\left(\widetilde{w}\right) dF\left(x\right) - c\left(F\right).$$

The principal's expected payoff from output distribution F and payments $\mathcal{G}$ is

$$\Pi\left(\mathcal{G},F\right) = \int_0^1 \left[x - \int_0^\infty \widetilde{w} dG_x\left(\widetilde{w}\right)\right] dF\left(x\right).$$

The value of an agent in project c for payments $\mathcal{G}$ is

$$u(c, \mathcal{G}) = \sup_{F \in \mathcal{F}} U(c, \mathcal{G}, F).$$

We can then define $\mathbf{F}^{c,\mathcal{G}}$ by $(F_n) \in \mathbf{F}^{c,\mathcal{G}}$ if and only if $\lim_{n\to\infty} U(c, \mathcal{G}, F_n) = u(c, \mathcal{G})$. The principal's value in project c given payment policy $\mathcal{G}$ is then

$$\pi(c, \mathcal{G}) \equiv \sup \left\{ \limsup_{n\to\infty} \Pi(\mathcal{G}, F_n) \quad : \quad (F_n) \in \mathbf{F}^{c,\mathcal{G}} \right\}.$$

If $w : [0,1] \to \mathbb{R}_+$ is a deterministic payment schedule, we continue to use the notation of the main document and write w in place of $\mathcal{G}$. So, for instance, the players' values in project c, given deterministic payment schedule w, are $u(c, w)$ and $\pi(c, w)$.

A.2 Analysis of binary projects

Similar to Section 3.2 of the main document, we aim at a characterization of the players' payoffs in equilibria of binary projects. We begin by determining the highest payoff the agent can obtain in an outcome in which the principal earns payoff π. That is, we consider all binary projects in which there is an equilibrium of the subgame following project selection where the principal earns payoff π, and ask what is the highest payoff that can be obtained by the agent in such an equilibrium?

Lemma A1. *Conditional on any payoff $\pi \in [0,1]$ for the principal, the agent can obtain a maximal payoff, in any binary project, of*

$$\int_0^{v(1-\pi)} \frac{\pi}{1-\gamma(z)} dz = v(1-\pi) - \int_\pi^1 v\left(1 - \frac{\pi}{z}\right) dz.$$

Proof. Consider binary projects in which there is an equilibrium involving the principal offering bonus $\widehat{b} \in [0,1]$ and the agent choosing probability of output one equal to $\widehat{\mu} \in [0,1]$. Let us now determine an upper bound on the payoff the agent can obtain in such an equilibrium.

Let $C : [0,1] \to \mathbb{R}_+$ be a binary project specifying the cost of obtaining each probability of output one. Write the agent's value when the bonus delivers felicity of consumption ω as

$$u(\omega) = \sup_{\mu \in [0,1]} \{\omega\mu - C(\mu)\}. \tag{1}$$

Note that the function u is identical to that for the risk-neutral case (where ω is equal to the bonus).

Let $\Gamma(\omega)$ denote the values μ such that there is a sequence (μ_n) with $\mu_n \to \mu$ and $\omega\mu_n - C(\mu_n) \to u(\omega)$. Take $\bar{\mu}(\omega) = \max \Gamma(\omega)$. From the same arguments as in the proof of Proposition 2 in the main document we have, for any $\omega \geq 0$,

$$u(\omega) = u(0) + \int_0^{\omega} \bar{\mu}(z)\, dz.$$

As in the main document, note that $u(0) \leq 0$; i.e., the agent cannot obtain a strictly positive payoff if the bonus is zero. Also, $u(\cdot)$ is non-decreasing and convex.

Denote $\widehat{\pi} = \widehat{\mu}\left(1 - \widehat{b}\right)$. Incentive compatibility of the principal offering bonus $\widehat{b}$ requires that, for all $z \geq 0$,

$$\begin{aligned} \widehat{\pi} &\geq \bar{\mu}(z)(1 - \gamma(z)) \\ &= u'_+(z)(1 - \gamma(z)). \end{aligned} \tag{2}$$

(Note that equality of $\bar{\mu}$ and u'_+ is established in Lemma 4 of the main document.)

Now let us determine an upper bound on the agent's payoff, across binary projects C, that can occur for an equilibrium in which the principal offers bonus $\widehat{b}$ and the agent achieves output one with probability $\widehat{\mu}$. Letting $\widehat{\omega} = v(\widehat{b})$, we can write the agent's equilibrium payoff as

$$u(\widehat{\omega}) = u(0) + \int_0^{\widehat{\omega}} u'_+(z)\, dz.$$

Consider maximizing this value by choice of convex function $u : \mathbb{R}_+ \to \mathbb{R}$ subject to the constraints (i) $u(0) \leq 0$, and (ii) $\widehat{\pi} \geq u'_+(z)(1 - \gamma(z))$ for all z. The first requirement reflects the above observation that the agent cannot obtain a positive payoff if the bonus is zero. The second condition is a re-statement of Condition (2). Any solution to this problem involves $u(0) = 0$ and

$$u'_+(z) = \frac{\widehat{\pi}}{1-\gamma(z)}$$

for all $z \in [0, \widehat{\omega})$. In other words, Constraint (ii) above holds with equality over $z \in [0, \widehat{\omega})$. The optimal choice of u in the above problem therefore satisfies

$$u(\omega) = \int_0^{\omega} \frac{\widehat{\pi}}{1-\gamma(z)} dz$$

on $[0, \widehat{\omega}]$.[1]

If $\widehat{\mu} = 0$, then both principal and agent must earn payoff zero. So suppose that $\widehat{\mu} > 0$. We have $\widehat{b} = 1 - \widehat{\pi}/\widehat{\mu}$ and so $\widehat{\omega} = v(1 - \widehat{\pi}/\widehat{\mu})$. An upper bound on the agent's payoff can then be written as

$$\int_0^{v\left(1-\frac{\widehat{\pi}}{\widehat{\mu}}\right)} \frac{\widehat{\pi}}{1-\gamma(z)} dz.$$

This is maximized by taking $\widehat{\mu} = 1$. It evaluates to zero if $\widehat{\pi} \in \{0, 1\}$ (i.e., the agent must obtain a payoff zero for these values of the principal's payoff), so suppose that $\widehat{\pi} \in (0, 1)$.

Let us now demonstrate that the agent can obtain the above payoff. Consider the binary project

$$C(\mu; \widehat{\pi}) = \begin{cases} \int_{\widehat{\pi}}^{\mu} v\left(1 - \frac{\widehat{\pi}}{\widetilde{\mu}}\right) d\widetilde{\mu} & \text{if } \mu \in [\widehat{\pi}, 1] \\ +\infty & \text{otherwise.} \end{cases}$$

The function $C(\cdot; \widehat{\pi})$ is strictly convex on $[\widehat{\pi}, 1]$. If the principal offers a bonus $b \in [0, 1 - \widehat{\pi}]$, generating felicity $\omega = v(b)$, the agent solves $\max_{\mu \in [0,1]} \{\mu\omega - C(\mu)\}$. The solution, $\mu^*(\omega)$ is unique and characterized by the first-order condition

$$\omega = C'(\mu^*(\omega); \widehat{\pi}) = v\left(1 - \frac{\widehat{\pi}}{\mu^*(\omega)}\right).$$

Hence,

$$\mu^*(\omega) = \frac{\widehat{\pi}}{1-\gamma(\omega)}.$$

The principal obtains payoff $\widehat{\pi}$ from offering every bonus in $[0, 1 - \widehat{\pi}]$. So there is indeed an

[1] Note that the integrand in the expression for $u(\omega)$ remains bounded, so $u(\omega)$ is necessarily well-defined and finite. To see this, we only need to consider the case where $\widehat{\omega}$ takes its highest value, i.e. $\widehat{\omega} = v(1)$. In this case, $\widehat{\pi} = 0$ and so we immediately conclude $u(\omega) = 0$ on $[0, \widehat{\omega}]$.

equilibrium in $C(\cdot;\widehat{\pi})$ with the principal offering bonus $1-\widehat{\pi}$ and the agent choosing output one with certainty. The agent's value in this project, using the definition in (1), satisfies $u(0)=0$ and $u'_+(\omega)=\mu^*(\omega)$ for all $\omega\in[0,v(1-\widehat{\pi})]$. Hence, the agent indeed obtains a payoff

$$\int_0^{v\left(1-\widehat{\pi}\right)}\frac{\widehat{\pi}}{1-\gamma(z)}dz.$$

This is also equal to the felicity of consumption $1-\widehat{\pi}$ less the agent's cost; i.e., $v(1-\widehat{\pi})-C(1;\widehat{\pi})$. This yields the expressions in the lemma. *QED*

Next consider the cost the agent incurs in an equilibrium of a binary project that yields the highest possible agent payoff, given principal expected payoffs π (this value of the agent's payoff is obtained in the previous result). We show that this cost is strictly convex in the principal's profits, a fact that will be important below.

Lemma A2. *The function $h(\pi)\equiv\int_\pi^1 v(1-\pi/z)\,dz$ is strictly convex over $\pi\in[0,1]$.*

Proof. Let $\pi\in(0,1)$ and consider the change of variables

$$x=1-\frac{\pi}{z}.$$

Note that

$$\frac{dx}{dz}=\frac{\pi}{z^2}=\frac{(1-x)^2}{\pi}.$$

Therefore, we have

$$h(\pi)=\pi\int_0^{1-\pi}\frac{v(x)}{(1-x)^2}dx.$$

Differentiating with respect to π yields

$$h'(\pi)=\int_0^{1-\pi}\frac{v(x)}{(1-x)^2}dx-\frac{v(1-\pi)}{\pi}$$

and

$$h''(\pi)=\frac{v'(1-\pi)}{\pi}>0.$$

The result follows. *QED*

A.3 Optimality of binary projects

We now state our main result, which implies the optimality of binary projects. For this purpose, recall that B_x denotes a binary distribution with probability mass x on output 1.

Proposition A1. *In any optimal outcome for the agent, (c^*, w^*, F^*), we have $F^* = B_1$.*

The rest of this section proves Proposition A1. Consider an agent-optimal outcome (c^*, w^*, F^*) and suppose for a contradiction that $F^* \neq B_1$. Note that, because the agent can secure a strictly positive payoff (as demonstrated in the previous section), we have $\mu_{F^*} > 0$ as well as $\mathbb{E}_{F^*}[w^*] > 0$.

We reach a contradiction by determining a project (and equilibrium in that project) in which the agent has a strictly higher payoff than for the outcome (c^*, w^*, F^*). Our first aim is to construct a binary project $\check{c}$ in which the principal offers bonus

$$b^* = \frac{\mathbb{E}_{F^*}[w^*]}{\mu_{F^*}}$$

to implement B_1 and earns a profit $1 - b^* = (\mu_{F^*} - \mu_{F^*} b^*) / \mu_{F^*} = (\mu_{F^*} - \mathbb{E}_{F^*}[w^*]) / \mu_{F^*}$ (we complete this task by Lemma A5 below). That is, the principal's profit is $1/\mu_{F^*}$ times the profit in the original outcome, and the agent's payment is also $1/\mu_{F^*}$ times the expected payment in the original. Note however that, because possibly $\check{c}(B_1) > c^*(B_1)$, we are unable to guarantee that the agent be better off in the binary project $\check{c}$. So the project $\check{c}$ will need to be further modified.

As with the proof of Proposition 1 in the main document, our first step is to define a binary project

$$\widehat{c}(B_\mu) = \begin{cases} \inf\{c^*(F) : \mu_F = \mu\} & \text{if } \mu < \mu_{F^*}. \\ \infty & \text{otherwise.} \end{cases}$$

As explained in the main document, in binary projects, the players view equivalently a payment schedule that pays a bonus b only for output one and the payment schedule w_b (with $w_b(x) = bx$ for $x \in [0,1]$, as defined in the main document). Different to the main document, we introduce a particular random payment policy, denoted $\mathcal{G}^b$. The policy $\mathcal{G}^b$ specifies, for each output realization $x \in [0,1]$, the payment distribution $G^b_x(w) = 1 - x + x\mathbf{1}_{w \geq b}(w)$

where $\mathbf{1}_{w \geq b}$ is the indicator function that takes value 1 when $w \geq b$ and zero otherwise. That is, G_x^b is the distribution that puts mass $1-x$ on payment zero and mass x on payment b. Note that the expected payment to the agent given policy $\mathcal{G}^b$ depends only on the mean of output: the expected payment when mean output is μ is equal to $b\mu$.

We now provide a result that is analogous to Lemma 1 in the main document.

Lemma A3. *For all $b \in [0,1]$, $\pi\left(\widehat{c}, w_b\right) \leq \pi\left(c^*, \mathcal{G}^b\right)$.*

Proof. By the definition of $u\left(\widehat{c}, w_b\right)$ and $\pi\left(\widehat{c}, w_b\right)$, there is a sequence $\left(\mu_n\right)$ such that

$$\mu_n v\left(b\right) - \widehat{c}\left(B_{\mu_n}\right) \to u\left(\widehat{c}, w_b\right)$$

and

$$\Pi\left(w_b, B_{\mu_n}\right) = \mu_n\left(1-b\right) \to \pi\left(\widehat{c}, w_b\right).$$

For each $k \in \mathbb{N}$, there exists n_k such that

$$\mu_{n_k} v\left(b\right) - \widehat{c}\left(B_{\mu_{n_k}}\right) + \frac{1}{k} \geq \mu v\left(b\right) - \widehat{c}\left(B_\mu\right)$$

for all $\mu \in [0,1]$. Therefore, for all k, and all $\mu \in [0, \mu_{F^*})$,

$$\mu_{n_k} v\left(b\right) - \inf\left\{c^*\left(F\right) : \mu_F = \mu_{n_k}\right\} + \frac{1}{k} \geq \mu v\left(b\right) - \inf\left\{c^*\left(F\right) : \mu_F = \mu\right\}.$$

Hence, there is a sequence $\left(F_{n_k}\right)$ with means μ_{n_k} (for each k) such that, for every F with mean in $[0, \mu_{F^*})$,

$$\mu_{n_k} v\left(b\right) - c^*\left(F_{n_k}\right) + \frac{2}{k} \geq \mu_F v\left(b\right) - c^*\left(F\right). \tag{3}$$

There are then two cases. The first is where the inequality (3) holds for all k and *all* F. In this case,

$$\lim_{k\to\infty} U\left(c^*, \mathcal{G}^b, F_{n_k}\right) = u\left(c^*, \mathcal{G}^b\right)$$

and hence

$$\pi\left(c^*, \mathcal{G}^b\right) \geq \lim_{k\to\infty} \Pi\left(\mathcal{G}^b, F_{n_k}\right) = \lim_{k\to\infty} \Pi\left(w_b, B_{\mu_{n_k}}\right) = \pi\left(\widehat{c}, w_b\right)$$

as desired. In the second case, the inequality (3) does not hold for some k and F, which implies

$$u\left(c^*,\mathcal{G}^b\right)=\sup\left\{\mu_F v\left(b\right)-c^*\left(F\right):F\in\mathcal{F}\right\}>\sup\left\{\mu_F v\left(b\right)-c^*\left(F\right):F\in\mathcal{F},\mu_F<\mu_{F^*}\right\}.$$

This means that there is a sequence of distributions along which the agent's payoff converges to $u\left(c^*,\mathcal{G}^b\right)$ and for which every distribution has mean at least μ_{F^*}. We therefore have

$$\pi^*\left(c^*,\mathcal{G}^b\right)\geq\mu_{F^*}\left(1-b\right)\geq\pi\left(\widehat{c},w_b\right),$$

where the second inequality follows because any distribution with mean at least μ_{F^*} is assigned an infinite cost in project $\widehat{c}$. *QED*

Our next goal is to define a binary project $\widetilde{c}$ and an equilibrium in $\widetilde{c}$ in which the principal offers the bonus payment b^*, and the agent chooses distribution $B_{\mu_{F^*}}$. Note that the principal's profit is then the same as in the original equilibrium: $\Pi\left(w_{b^*},B_{\mu_{F^*}}\right)=\Pi\left(w^*,F^*\right)=\mu_{F^*}\left(1-b^*\right)$.

Let then $\bar{c}$ be determined by

$$\mu_{F^*}v\left(b^*\right)-\bar{c}=\sup_{\mu\in[0,1]}\left\{\mu v\left(b^*\right)-\widehat{c}\left(B_\mu\right)\right\}.$$

Define the binary project $\widetilde{c}$ by

$$\widetilde{c}\left(F\right)=\begin{cases}\bar{c} & \text{if } F=B_{\mu_{F^*}}\\ \widehat{c}\left(F\right) & \text{if } F\neq B_{\mu_{F^*}}.\end{cases}$$

Note that the agent has a best response in project $\widetilde{c}$ to bonus b^* equal to the distribution $B_{\mu_{F^*}}$.

As in the case of a risk-neutral agent, we can show that $\bar{c}\leq c^*\left(F^*\right)$. Suppose for a contradiction that $\bar{c} > c^*(F^*)$. Then

$$\begin{aligned}\mu_{F^*} v(b^*) - c^*(F^*) &> \mu_{F^*} v(b^*) - \bar{c} \\ &= \sup\{\mu v(b^*) - \widehat{c}(B_\mu) : \mu \in [0,1]\} \\ &= \sup\{\mu_F v(b^*) - c^*(F) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\}.\end{aligned}$$

By continuity of v, there is then $b < b^*$ such that

$$\mu_{F^*} v(b) - c^*(F^*) > \sup\{\mu_F v(b) - c^*(F) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\}.$$

This means that, if the principal offers payment schedule $\mathcal{G}^b$ in project c^*, the principal's value must be at least $\mu_{F^*}(1-b) > \mu_{F^*}(1-b^*) = \Pi(w^*, F^*)$. This contradicts the incentive compatibility of the payment schedule w^* in c^*.

We now want to show that $\left(w_{b^*}, B_{\mu_{F^*}}\right)$ is an equilibrium of project $\widetilde{c}$. To do so, we will rely on the following lemma.

Lemma A4. *For all $b \in [0, b^*)$, $\pi(\widetilde{c}, w_b) = \pi(\widehat{c}, w_b)$.*

Proof. The argument is identical to Lemma 2 in the main document, after noting that the agent has a felicity $v(b)$ (rather than b) when receiving bonus payment b. *QED*

Now let us show that $\left(w_{b^*}, B_{\mu_{F^*}}\right)$ is an equilibrium of project $\widetilde{c}$. We already saw that $B_{\mu_{F^*}}$ is incentive compatible for the agent in subgame $(\widetilde{c}, w_{b^*})$ (by choice of its cost $\bar{c}$). So we need to show that w_{b^*} is incentive compatible in $\widetilde{c}$. It is immediate that the principal does not want to deviate to a bonus $b > b^*$ (because the principal cannot attain expected output higher than μ_{F^*}; see also the argument in the main document). If $b < b^*$, then

$$\pi(\widetilde{c}, w_b) = \pi(\widehat{c}, w_b) \leq \pi\left(c^*, \mathcal{G}^b\right) \leq \Pi(w^*, F^*) = \Pi\left(w_{b^*}, B_{\mu_{F^*}}\right).$$

The first equality follows by Lemma A4. The first inequality follows by Lemma A3. The second inequality follows because w^* is incentive compatible for the principal in c^*. The final equality follows by definition of b^*. Thus, the principal does not gain by deviating to $b < b^*$.

As we observed, the principal obtains the same payoff in outcome $\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right)$ as in the outcome (c^*, w^*, F^*). We have been unable, however, to determine whether the agent is better off in $\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right)$. Although the expected payment is the same in both outcomes, we have not ruled out that the agent could earn a lower payoff in outcome $\left(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}\right)$ due to the uncertainty in payments (i.e., due to worse insurance). Therefore, we make further modifications to the project.

First, we determine the project $\check{c}$ mentioned above, together with an equilibrium in which the agent chooses distribution B_1 and payoffs are those in the equilibrium $\left(w_{b^*}, B_{\mu_{F^*}}\right)$ of project $\widetilde{c}$, multiplied by $1/\mu_{F^*}$. This project is defined by $\check{c}\left(B_\mu\right) = \widetilde{c}\left(B_{\mu\mu_{F^*}}\right)/\mu_{F^*}$ for all $\mu \in [0,1]$. We show the following.

Lemma A5. *For all $b \geq 0$, $\pi\left(\check{c}, w_b\right) = \pi\left(\widetilde{c}, w_b\right)/\mu_{F^*}$.*

Proof. We first show $\pi\left(\check{c}, w_b\right) \geq \pi\left(\widetilde{c}, w_b\right)/\mu_{F^*}$ for any $b \geq 0$. Fix any such b and suppose that (μ_n) is a sequence for which $U\left(\widetilde{c}, w_b, B_{\mu_n}\right) \to u\left(\widetilde{c}, w_b\right)$ and $\Pi\left(w_b, B_{\mu_n}\right) \to \pi\left(\widetilde{c}, w_b\right)$. Then there is a subsequence (μ_{n_k}) such that, for all k and all $\mu \in [0,1]$,

$$\mu_{n_k} v(b) - \widetilde{c}\left(B_{\mu_{n_k}}\right) + \frac{1}{k} \geq \mu v(b) - \widetilde{c}\left(B_\mu\right).$$

Then, for all $\mu \in [0, \mu_{F^*}]$,

$$\frac{\mu_{n_k}}{\mu_{F^*}} v(b) - \frac{1}{\mu_{F^*}}\widetilde{c}\left(B_{\mu_{n_k}}\right) + \frac{1}{\mu_{F^*}k} \geq \frac{\mu}{\mu_{F^*}} v(b) - \frac{1}{\mu_{F^*}}\widetilde{c}\left(B_\mu\right).$$

Hence, for all $\mu \in [0,1]$,

$$\frac{\mu_{n_k}}{\mu_{F^*}} v(b) - \check{c}\left(B_{\frac{\mu_{n_k}}{\mu_{F^*}}}\right) + \frac{1}{\mu_{F^*}k} \geq \mu v(b) - \check{c}\left(B_\mu\right).$$

This implies that $U\left(\check{c}, w_b, B_{\frac{\mu_{n_k}}{\mu_{F^*}}}\right) \to u\left(\check{c}, w_b\right)$ and $\Pi\left(w_b, B_{\mu_{n_k}/\mu_{F^*}}\right) \to \pi\left(\widetilde{c}, w_b\right)/\mu_{F^*}$. This establishes the claim.

We now show that $\pi\left(\check{c}, w_b\right) \leq \pi\left(\widetilde{c}, w_b\right)/\mu_{F^*}$. Suppose that (μ_n) is a sequence for which $U\left(\check{c}, w_b, B_{\mu_n}\right) \to u\left(\check{c}, w_b\right)$ and $\Pi\left(w_b, B_{\mu_n}\right) \to \pi\left(\check{c}, w_b\right)$. Then there is a subsequence (μ_{n_k})

such that, for all k and all $\mu \in [0,1]$,

$$\mu_{n_k} v\left(b\right) - \frac{\widetilde{c}\left(B_{\mu_{n_k}\mu_{F^*}}\right)}{\mu_{F^*}} + \frac{1}{k} \geq \mu v\left(b\right) - \frac{\widetilde{c}\left(B_{\mu\mu_{F^*}}\right)}{\mu_{F^*}}.$$

Then, for all k and all $\mu \in [0,1]$,

$$\mu_{F^*}\mu_{n_k} v\left(b\right) - \widetilde{c}\left(B_{\mu_{n_k}\mu_{F^*}}\right) + \frac{\mu_{F^*}}{k} \geq \mu_{F^*}\mu v\left(b\right) - \widetilde{c}\left(B_{\mu\mu_{F^*}}\right).$$

Letting, for each k, $\mu'_{n_k} = \mu_{F^*}\mu_{n_k}$, we have

$$\mu'_{n_k} v\left(b\right) - \widetilde{c}\left(B_{\mu'_{n_k}}\right) + \frac{\mu_{F^*}}{k} \geq \mu v\left(b\right) - \widetilde{c}\left(B_{\mu}\right)$$

for all $\mu \in [0, \mu_{F^*}]$, and hence all $\mu \in [0,1]$. Therefore, $U\left(\widetilde{c}, w_b, B_{\mu_{F^*}\mu_{n_k}}\right) \to u\left(\widetilde{c}, w_b\right)$, while $\Pi\left(w_b, B_{\mu_{F^*}\mu_{n_k}}\right) \to \mu_{F^*}\pi\left(\check{c}, w_b\right)$, which establishes the claim. *QED*

Lemma A5 implies that the principal's incentives to offer different bonuses in project $\check{c}$ are the same as in $\widetilde{c}$. The optimal bonus for the agent that is incentive-compatible for the principal is b^*, with the agent best responding in subgame $(\check{c}, b^*)$ with the distribution B_1. To see this, recall that $\left(b^*, B_{\mu_{F^*}}\right)$ is an equilibrium in project $\widetilde{c}$. Therefore, for all $\mu \in [0,1]$,

$$\mu_{F^*} v\left(b^*\right) - \widetilde{c}\left(B_{\mu_{F^*}}\right) \geq \mu v\left(b^*\right) - \widetilde{c}\left(B_{\mu}\right).$$

The claim follows because this is equivalent to the statement that, for all $\mu \in [0,1]$,

$$v\left(b^*\right) - \check{c}\left(B_1\right) \geq \mu v\left(b^*\right) - \check{c}\left(B_{\mu}\right).$$

We have established then that $\check{c}$ is a binary project in which the agent (in the agent-optimal equilibrium) receives a payment equal to $1/\mu_{F^*}$ times the expected payment $\mathbb{E}_{F^*}\left[w^*\right]$ of the original equilibrium and achieves output one with certainty. The principal's profits are $1 - b^*$. Consider now an *optimal* binary project for the agent in which the principal earns profits $1 - b^*$. Recalling the analysis in the previous section, there is an optimal binary project where the principal pays b^* and the agent achieves output one with probability one.

Recalling the definition in Lemma A2, the agent's cost is $h(1-b^*)$. The agent's payoff satisfies

$$v(b^*) - h(1-b^*) \geq v(b^*) - \check{c}(B_1) \geq v(b^*) - \frac{c^*(F^*)}{\mu_{F^*}}.$$

The first inequality follows because $\check{c}$ is a (not-necessarily optimal) binary project. The second inequality follows by construction of $\check{c}(B_1)$.

We can conclude that

$$h(1-b^*) \leq \frac{c^*(F^*)}{\mu_{F^*}}.$$

Because h is strictly convex, and because $h(1) = 0$, we have

$$h(1-\mu_{F^*}b^*) < c^*(F^*).$$

Therefore,

$$v(\mu_{F^*}b^*) - h(1-\mu_{F^*}b^*) > v(\mu_{F^*}b^*) - c^*(F^*) \geq U(c^*, w^*, F^*),$$

where the second inequality follows by Jensen's inequality and concavity of v (as well as the observation that the expected payment determined by distribution F^* and payment schedule w^* is $\mu_{F^*}b^*$). The left-hand side represents the agent's payoff in an agent-optimal binary project in which the principal obtains payoff $1-\mu_{F^*}b^*$ (as determined in the previous section). The right-hand side is the agent's expected payoff in the original project. That the agent does better in the aforementioned binary project contradicts the optimality for the agent of the original, as desired. This completes the proof of Proposition A1.

Let us conclude by relating the claims in this online appendix to those made at the end of Section 4 (under the heading "Risk Aversion"). First note that Proposition A1 implies that binary projects are optimal. While the claim in the proposition only states that it is optimal for the agent to choose a project where the principal implements distribution B_1, recall from the discussion in Appendix B (under the heading "Uniqueness beyond binary projects") that any such project can be converted to a binary one. The proof of Proposition A1 showed in particular that, for an outcome (c^*, w^*, F^*) where (w^*, F^*) is an equilibrium of c^*, and where

$F^* \neq B_1$, there exists a binary project and equilibrium of that project which represents a strict Pareto improvement for both players. In particular, we constructed a binary outcome where the principal obtains a payoff $1 - \mu_{F^*} b^* = 1 - \mathbb{E}_{F^*}[w^*] > \mu_{F^*} - \mathbb{E}_{F^*}[w^*]$. For non-binary projects where B_1 is implemented in equilibrium, the previous observation that the project can be converted to a binary project applies. In this case, the equilibrium payoffs of the players are unaffected by the conversion. Finally, the claim that the marginal costs of probabilities μ above π^* are given by $v\left(1 - \frac{\pi^*}{\mu}\right)$ in an optimal project follows from the specification of $C(\mu; \pi^*)$ in the proof of Lemma A1.