Economics 2010b: Microeconomic Theory II

Michael Powell

Contract Theory

1 Introduction to Contract Theory

One of the important issues that we touched on briefly in our discussion of general equilibrium theory is the idea of market incompleteness and its consequences. When markets are incomplete—either in the sense that we talked about last time or in the sense that consumption involves unpriceable externalities—equilibrium allocations may not be constrained-efficient, opening up scope for some sort of third-party intervention. It may be government intervention via a system of taxation or rules, or it may be private intervention by an entrepreneur who sets up governance institutions. We were able to make some high-level claims last time about what happens when there are these "market failures," but without imposing more structure on the problem, it is difficult to make specific claims about how they should be managed.

For the last three weeks of the class, we will zoom in and study micro situations in which it could be said that markets are incomplete. We will focus on what is referred to as the Principal–Agent problem in which there are two players, a Principal P and an Agent A. The Principal needs the Agent to do something that she cannot do herself, so she hires the Agent and writes a contract that governs how the Agent will be paid. We can think of the Principal being an employer and the Agent an employee, where the Principal lacks the time or expertise to engage in production. We can think of the Principal being a patient and the Agent a doctor, where the doctor takes some actions that the patient does not know or understand. We can think of the Principal being a client and the Agent being a lawyer acting on the client's behalf. And so on.

When equilibrium outcomes arising from the Principal–Agent interaction are Pareto in-

efficient, we will say that there is a *moral hazard problem*, which is a term that originated in the insurance industry to describe situations in which someone increases their exposure to risk in response to buying insurance. Fundamentally, the moral hazard problem is a just an externality problem. Now, when we make a claim like "there are externalities, so outcomes will be inefficient" it is important to have in mind that whether or not externalities "matter" in the sense that they lead to Pareto inefficient equilibrium outcomes depends critically on the set of instruments parties have for managing those externalities: it depends on the contracting space. Over the next couple lectures, we will look at several different sources of *contractual frictions* that prevent the Principal and Agent from writing contracts with each other that result in Pareto optimal outcomes.

The first situation we will look at will occur when individual actions chosen by the Agent are not observed by the Principal but determine the distribution of a verifiable performance measure that can be written into a contract. The Agent may be more risk-averse than the Principal, so writing a high-powered contract on that noisy performance measure transfers risk onto the Agent and therefore leads to an inefficient allocation of risk between the two parties. As a result, there is a trade-off between incentive provision (and therefore what the Agent chooses to do) and inefficient risk allocation. This is the celebrated *risk–incentives trade-off*.

The second contracting friction that might arise is that an Agent is either liquidity-constrained or is subject to a limited-liability constraint. As a result, the Principal is unable to extract all the surplus the Agent generates and must therefore provide the Agent with *incentive rents* in order to motivate him. That is, offering the Agent a higher-powered contract induces him to work harder and therefore increases the total size of the pie, but it also leaves the Agent with a larger share of that pie. The Principal then, in choosing a contract, chooses one that trades off the creation of surplus with her ability to extract that surplus. This is the *motivation–rent extraction trade-off*.

A third contracting friction that might arise is that the Principal's objective simply

cannot be written into a formal contract. Instead, the Principal has to rely on imperfectly aligned performance measures. Increasing the strength of a formal contract that is based on imperfectly aligned performance measures may motivate the Agent to work hard toward the Principal's objectives, but it may also motivate him to work hard toward objectives that either hurt the Principal or at least do not help her. This is known as the *multi-task problem* (Holmström and Milgrom, 1991), and failure to account for the effects of using distorted performance measures is sometimes referred to as *the folly of rewarding A while hoping for B* (Kerr, 1975).

Finally, there may be multiple Agents who work together to produce something for the Principal. Their individual contributions may not be observable, so contracts may only be able to be written on the final output. This inability to distinguish individual contributions is what is referred to as the *moral hazard in teams problem* (Holmström, 1982).

All of these sources of contractual frictions lead to similar results—under the optimal contract, the Agent (or Agents) chooses an action that is not jointly optimal from his and the Principal's perspective. But in different applied settings, different assumptions regarding what is contractible and what is not are more or less plausible. As a result, it is useful to master at least elementary versions of models capturing these four sources of frictions, so that you are well-equipped to use them as building blocks.

2 The Risk-Incentives Trade-off

I will begin with a pretty general description of the standard principal-agent model, but I will shortly afterwards specialize the model quite a bit in order to focus on a single point—the risk–incentives trade-off.

2.1 The Model

There is a risk-neutral Principal (P) and a risk-averse Agent (A). The Agent chooses an **effort level** $e \in \mathcal{E} \subset \mathbb{R}_+$ and incurs a cost of $c(e)$, where $c : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing and strictly convex. If $\mathcal{E}$ is an interval, we will say that **effort is continuous**, and if $\mathcal{E}$ consists of a finite number of points, we will say that **effort is discrete**. We will assume $0 \in \mathcal{E}$, and $c(0) = 0$. The effort level affects the distribution over **output** $y \in \mathcal{Y}$, with y distributed according to CDF $F(\cdot|\, e)$. This output can be sold on the product market at price p, and the revenues py accrue to the Principal.

The Principal does not have any direct control over the Agent, but what she can do is write a contract that influences what the Agent will do. In particular, she can write a contract $w \in \mathcal{W} \subset \{w : \mathcal{Y} \times \mathcal{E} \to \mathbb{R}\}$, where $\mathcal{W}$ is the **contracting space**. The contract determines a transfer $w(y, e)$ that she is compelled to pay the Agent if output y is realized, and he chose effort e. If $\mathcal{W}$ does not allow for functions that depend directly on effort, we will say that **effort is noncontractible**, and abusing notation slightly, we will write the contractual payment the Principal is compelled to pay the Agent if output y is realized as $w(y, e) = w(y)$ for all $e \in \mathcal{E}$. We will be assuming throughout that effort is noncontractible, but I wanted to highlight that it is a real restriction on the contracting space, and it is one that we will impose as a primitive of the model.

The Agent can decline to work for the Principal and reject her contract, pursuing his outside option instead. This outside option provides utility $\bar{u}$ to the Agent and $\bar{\pi}$ to the Principal. If the Agent accepts the contract, the Principal's and Agent's preferences are, respectively,

$$\begin{aligned} \Pi(w, e) &= \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e) = E_y[py - w|\, e] \\ U(w, e) &= \int_{y \in \mathcal{Y}} u(w(y) - c(e))\, dF(y|\, e) = E_y[u(w - c(e))|\, e], \end{aligned}$$

where u is increasing and weakly concave.

We have described the players, what they can do, and what their preferences are. We still need to describe the timing of the game that the players play, as well as the solution concept. Explicitly describing the timing of the model is essential to remove any ambiguity about what players know when they make their decisions. In this model, the timing of the game is:

1. P offers A a contract $w \in \mathcal{W}$. w is commonly observed.

2. A accepts the contract ($d = 1$) or rejects it ($d = 0$), in which case he receives $\bar{u}$, and the game ends. d is commonly observed.

3. If A accepts the contract, A chooses effort level e and incurs cost $c(e)$. e is privately observed by A.

4. Output y is drawn from distribution with CDF $F(\cdot \mid e)$. y is commonly observed.

5. P pays A an amount $w(y)$. The payment is commonly observed.

A couple remarks are in order at this point. First, behind the scenes, there is an implicit assumption that there is a third-party contract enforcer (a judge or arbitrator) who can costlessly detect when agreements have been broken and costlessly exact harsh punishments on the offender.

Second, much of the literature assumes that the Agent's effort level is privately observed by the Agent and therefore refers to this model as the "hidden action" model. Ultimately, though, the underlying source of the moral-hazard problem is that contracts cannot be conditioned on relevant variables, not that the relevant variables are unobserved by the Principal. Many papers assume effort is unobservable to justify it being noncontractible. While this is a compelling justification, in our framework, the contracting space itself is a primitive of the model. Later in the course, we will talk a bit about the microfoundations for different assumptions on the contracting space.

Finally, let us describe the solution concept. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an **acceptance decision** $d^* : \mathcal{W} \to \{0, 1\}$, and an **effort choice** $e^* : \mathcal{W} \times \{0, 1\} \to \mathcal{E}$ such that, given the contract w^*, the Agent optimally chooses d^* and e^*, and given d^* and e^*, the Principal optimally offers contract w^*. We will say that the optimal contract **induces** effort e^*.

2.2 First-Best Benchmark

If we want to talk about the inefficiencies that arise in equilibrium in this model, it will be useful first to establish a benchmark against which to compare outcomes. In this model, a **feasible outcome** is a distribution over payments from the Principal to the Agent as well as an effort level $e \in \mathcal{E}$. We will say that a feasible outcome is **Pareto optimal** if there is no other feasible outcome that both players weakly prefer and one player strictly prefers. If an effort level e is part of a Pareto optimal outcome, we will say that it is a **first-best** effort level, and we will denote it by e^{FB}.

Lemma 1. The first-best effort level satisfies

$$e^{FB} \in \underset{e \in \mathcal{E}}{\operatorname{argmax}} E_y \left[py \middle| e \right] - c\left(e\right).$$

Proof of Lemma 1. In any Pareto-optimal outcome, payments to the agent are deterministic. Since the Agent is risk averse, given an outcome involving stochastic payments to the Agent, there is another outcome in which the Agent chooses the same effort level and receives the certainty equivalent wage instead. This outcome yields the same utility for the Agent, and since the Agent is risk averse, the certainty equivalent payment is smaller in expectation, so the Principal is strictly better off. Next, given constant deterministic wages, any Pareto-optimal outcome must solve

$$\max_{w \in \mathbb{R}, e \in \mathcal{E}} E_y \left[py \middle| e \right] - w$$

subject to

$$u\left(w - c\left(e\right)\right) \geq \bar{u},$$

for some $\bar{u}$. In any solution to this problem, the constraint must bind, since u is increasing. Moreover, since u is increasing, it is invertible, so we can write

$$w = u^{-1}\left(\bar{u}\right) + c\left(e\right),$$

and therefore the first-best effort level must solve the problem specified in the Lemma.■

This Lemma shows that the first-best effort level maximizes expected revenues net of effort costs. If effort is fully contractible, so that the Principal could offer any contract w that depended nontrivially on e, then the first-best effort would be implemented in equilibrium. In particular, the Principal could offer a contract that pays the Agent $u^{-1}\left(\bar{u}\right) + c\left(e^{FB}\right)$ if he choose e^{FB}, and pays him a large negative amount if he chooses any $e \neq e^{FB}$. That the first-best effort level can be implemented in equilibrium if effort is contractible is an illustration of a version of the *Coase Theorem*: if the contracting space is sufficiently rich, equilibrium outcomes will be Pareto optimal.

If effort is noncontractible, and $e^{FB} > 0$, then equilibrium will not involve Pareto optimal outcomes. For an outcome to be Pareto optimal, it has to involve a deterministic wage payment to the Agent. But if the Agent's wage is independent of output, then it must also be independent of his effort level. He will therefore receive no benefit from choosing a costly effort level, and so he will choose $e = 0 < e^{FB}$. The question to which we will now turn is: what effort will be implemented in equilibrium when effort is noncontractible?

2.3 Equilibrium Effort

Since the Agent's effort choice affects the Principal's payoffs, the Principal would ideally like to directly choose the Agent's effort. But, she has only indirect control: she can offer different contracts, and different contracts may get the Agent to optimally choose different

effort levels. We can think of the Principal's problem as choosing an effort level e as well as a contract for which e is *incentive compatible* for the Agent to choose and for which it is *individually rational* for the Agent to accept. As a loose analogy, we can connect the Principal's problem to the social planner's problem from general equilibrium theory. We can think of e as analogous to an allocation the Principal would like to induce, and the choice of a contract as analogous to setting "prices" so as to decentralize e as an equilibrium allocation.

Formally, the Principal offers a contract $w \in \mathcal{W}$ and "proposes" an effort level e in order to solve

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|e)$$

subject to two constraints. The first constraint is that the agent actually prefers to choose effort level e rather than any other effort level $\hat{e}$. This is the **incentive-compatibility constraint**:

$$e \in \operatorname*{argmax}_{\hat{e} \in \mathcal{E}} \int_{y \in Y} u(w(y) - c(\hat{e}))\, dF(y|\hat{e}).$$

The second constraint ensures that, given that the agent knows he will choose e if he accepts the contract, he prefers to accept the contract rather than to reject it and receive his outside utility $\bar{u}$. This is the **individual-rationality constraint** or **participation constraint**:

$$\int_{y \in Y} u(w(y) - c(e))\, dF(y|e) \geq \bar{u}.$$

At this level of generality, the model is not very tractable. We will need to impose more structure on it in order to highlight some its key trade-offs and properties.

CARA-Normal Case with Affine Contracts In order to highlight one of the key trade-offs that arise in this class of models, we will make a number of strong simplifying assumptions.

Assumption A1 (CARA). The Agent has CARA preferences over wealth and effort costs,

which are quadratic:

$$u\left(w\left(y\right)-c\left(e\right)\right)=-\exp\left\{-r\left(w\left(y\right)-\frac{c}{2}e^{2}\right)\right\},$$

and his outside option yields utility $-\exp\left\{-r\bar{u}\right\}$.

Assumption A2 (Normal Output). Effort shifts the mean of a normally distributed random variable. That is, $y\sim N\left(e,\sigma^{2}\right)$.

Assumption A3 (Affine Contracts). $\mathcal{W}=\left\{w:\mathcal{Y}\rightarrow\mathbb{R},w\left(y\right)=s+by\right\}$. That is, the contract space permits only affine contracts.

Assumption A4 (Continuous Effort). Effort is continuous and satisfies $\mathcal{E}=\mathbb{R}_{+}$.

In principle, we should not impose exogenous restrictions on the *functional form* of $w\left(y\right)$. There is an important class of applications, however, that restrict attention to affine contracts, $w\left(y\right)=s+by$, and a lot of the basic intuition that people have for the comparative statics of optimal contracts come from imposing this restriction.

In many environments, an optimal contract does not exist if the contracting space is sufficiently rich, and situations in which the agent chooses the first-best level of effort, and the principal receives all the surplus can be arbitrarily approximated with a sequence of sufficiently perverse contracts (Mirrlees, 1974; Moroni and Swinkels, 2014). In contrast, the optimal affine contract often results in an effort choice that is lower than the first-best effort level, and the principal receives a lower payoff.

There are then at least three ways to view the exercise of solving for the optimal affine contract.

1. From an applied perspective, many pay-for-performance contracts in the world are affine in the relevant performance measure—franchisees pay a franchise fee and receive a constant fraction of the revenues their store generates, windshield installers receive a base wage and a constant piece rate, fruit pickers are paid per kilogram of fruit they pick. And so given that many practitioners seem to restrict attention to this class

of contracts, why not just make sure they are doing what they do optimally? Put differently, we can brush aside global optimality on purely pragmatic grounds.

2. Many pay-for-performance contracts in the world are affine in the relevant performance measure. Our models are either too rich or not rich enough in a certain sense and therefore generate optimal contracts that are inconsistent with those we see in the world. Maybe the aspects that, in the world, lead practitioners to use affine contracts are orthogonal to the considerations we are focusing on, so that by restricting attention to the optimal affine contract, we can still say something about how real-world contracts ought to vary with changes in the underlying environment. This view presumes a more positive (as opposed to normative) role for the modeler and hopes that the theoretical analogue of the omitted variables bias is not too severe.

3. Who cares about second-best when first-best can be attained? If our models are pushing us toward complicated, non-linear contracts, then maybe our models are wrong. Instead, we should focus on writing down models that generate affine contracts as the optimal contract, and therefore we should think harder about what gives rise to them. (And indeed, steps have been made in this direction—see Holmström and Milgrom (1987), Diamond (1998) and, more recently, Carroll (2013) and Barron, Georgiadis, and Swinkels (2017)) This perspective will come back later in the course when we discuss the Property Rights Theory of firm boundaries.

Given Assumptions $(A1) - (A3)$, for any contract $w(y) = s + by$, the income stream the agent receives is normally distributed with mean $s + be$ and variance $b^2\sigma^2$. His expected utility over monetary compensation is therefore a moment-generating function for a normally distributed random variable, (recall that if $X \sim N(\mu, \sigma^2)$, then $E[\exp\{tX\}] = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}$), so his preferences can be written as

$$E[-\exp\{-r(w(y) - c(e))\}] = -\exp\left\{-r\left(s + be - \frac{r}{2}b^2\sigma^2 - \frac{c}{2}e^2\right)\right\}.$$

We can take a monotonic transformation of his utility function ($f(x) = -\frac{1}{r}\log(-x)$) and represent his preferences as:

$$\begin{aligned} U(e,w) &= E[w(y)] - \frac{r}{2}Var(w(y)) - \frac{c}{2}e^2 \\ &= s + be - \frac{r}{2}b^2\sigma^2 - \frac{c}{2}e^2. \end{aligned}$$

The Principal's program is then

$$\max_{s,b,e} pe - (s + be)$$

subject to incentive-compatibility

$$e \in \operatorname*{argmax}_{\hat{e}} b\hat{e} - \frac{c}{2}\hat{e}^2$$

and individual-rationality

$$s + be - \frac{r}{2}b^2\sigma^2 - \frac{c}{2}e^2 \geq \bar{u}.$$

Solving this problem is then relatively straightforward. Given an affine contract $s + be$, the Agent will choose an effort level $e(b)$ that satisfies his first-order conditions

$$e(b) = \frac{b}{c},$$

and the Principal will choose the value s to ensure that the Agent's individual-rationality constraint holds with equality. If it did not hold with equality, the Principal could reduce s, making herself better off without affecting the Agent's incentive-compatibility constraint, while still respecting the Agent's individual-rationality constraint. That is,

$$s + be(b) = \frac{c}{2}e(b)^2 + \frac{r}{2}b^2\sigma^2 + \bar{u}.$$

In other words, the Principal has to ensure that the Agent's total expected monetary compensation, $s + be(b)$, fully compensates him for his effort costs, the risk costs he has to bear if he accepts this contract, and his opportunity cost. Indirectly, then, the Principal bears these costs when designing an optimal contract.

The Principal's remaining problem is to choose the incentive slope b to solve

$$\max_b pe(b) - \frac{c}{2}e(b)^2 - \frac{r}{2}b^2\sigma^2 - \bar{u}.$$

This is now an unconstrained problem with proper convexity assumptions, so the Principal's optimal choice of incentive slope solves her first-order condition

$$0 = p\underbrace{e'(b^*)}_{1/c} - c\underbrace{e^*(b^*)}_{b^*/c}\underbrace{e'(b^*)}_{1/c} - rb^*\sigma^2,$$

and therefore the optimal incentive slope satisfies

$$b^* = \frac{p}{1 + rc\sigma^2}.$$

Moreover, given b^* and the individual-rationality constraint, we can back out s^*.

$$s^* = \bar{u} + \frac{1}{2}\left(rc\sigma^2 - 1\right)\frac{(b^*)^2}{c}.$$

Depending on the parameters, it may be the case that $s^* < 0$. That is, the Agent would have to pay the Principal if he accepts the job and does not produce anything.

Now, how does the effort that is induced in this optimal affine contract compare to the **first-best effort**? Using the result from Lemma 1, we know that first-best effort in this setting solves

$$\max_{e \in \mathbb{R}_+} pe - \frac{c}{2}e^2,$$

and therefore $e^{FB} = p/c$.

Even if effort is noncontractible, the Principal could in principle implement exactly this same level of effort by writing a contract only on output. To do so, she would choose $b = p$, since this would get the Agent to choose $e(p) = p/c$. Why, in this setting, does the Principal not choose such a contract? Let us go back to the Principal's problem of choosing the incentive slope b.

$$\max_b pe(b) - \frac{c}{2}e(b)^2 - \frac{r}{2}b^2\sigma^2 - \bar{u}$$

Often, when an economic model can be solved in closed form, we jump right to the solution. Only when a model cannot be solved in closed form do we typically stop to think carefully about what economic properties its solution must possess. I want to spend a couple minutes *partially* characterizing this model's solution, even though we already completely characterized it above, just to highlight how this kind of reasoning can be helpful in developing intuition that might generalize beyond the present setting. In particular, many fundamental features of models can be seen as a comparison of first-order losses or gains against second-order gains or losses, so it is worth going through this first-order–second-order logic. Suppose the Principal chooses $b = p$, and consider a marginal reduction in b away from this value. The change in the Principal's profits would be

$$\begin{aligned} & \frac{d}{db}\left(pe(b) - \frac{c}{2}e(b)^2 - \frac{r}{2}b^2\sigma^2\right)\Big|_{b=p} \\ = & \underbrace{\frac{d}{db}\left(pe(b) - \frac{c}{2}e(b)^2\right)\Big|_{b=p}}_{=0} - rp\sigma^2 < 0. \end{aligned}$$

This first term is zero, because $b = p$ in fact maximizes $pe(b) - \frac{c}{2}e(b)^2$, since it induces the first-best level of effort. This is just an application of the envelope theorem you learned in Ec 2010a. The second term in this expression is strictly negative. This implies that, relative to the contract that induces first-best effort, a reduction in the slope of the incentive contract yields a first-order gain to the Principal resulting from a decrease in the risk costs the Agent bears, while it yields a second-order loss in terms of profits resulting from moving away from

the effort level that maximizes revenues minus effort costs. The optimal contract balances the incentive benefits of higher-powered incentives with these risk costs, and these risk costs are higher if the Agent is more risk averse and if output is noisier.

This trade-off seems first-order in some settings (e.g., insurance contracts in health care markets, some types of sales contracts in industries in which individual sales are infrequent, large, and unpredictable) and for certain types of output. There are many other environments in which contracts provide less-than-first-best incentives, but the first-order reasons for these low-powered contracts seem completely different, and we will turn to these environments next week.

Exercise 18 (Adapted from MWG 14.B.4). Suppose there are three possible effort levels, $\mathcal{E} = \{e_1, e_2, e_3\}$, and two possible output levels, $\mathcal{Y} = \{0, 10\}$, and the output price is $p = 1$. The probability that $y = 10$ conditional on each of the effort levels is given by the probability mass function $f(10|\, e_1) = 2/3$, $f(10|\, e_2) = 1/2$, and $f(10|\, e_3) = 1/3$. The Agent's effort cost function satisfies $c(e_1) = 5/3$, $c(e_2) = 8/5$, and $c(e_3) = 4/3$. Finally, the Agent's utility function is given by $u(w) = \sqrt{w}$, and his outside option yields utility $\bar{u} = 0$.

(a) What is the optimal contract for the Principal when effort is contractible?

(b) Show that if effort is noncontractible, and $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}\}$, then there is no contract w for which the Agent will choose e_2. For what levels of $c(e_2)$ would there exist a contract w under which the Agent would choose e_2?

(c) What is the optimal contract when effort is noncontractible, and $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}\}$?

(d) Suppose instead that $c(e_1) = \sqrt{8}$, and let $f(10|\, e_1) = x \in (0, 1)$. If effort is noncontractible, and $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}\}$, what is the optimal contract for the Principal as x approaches 1? Is the level of effort implemented higher or lower than when effort is contractible?

Exercise 19. Suppose the Agent can allocate time to two different tasks. Let e_i be the amount of time spent on task $i \in \{1, 2\}$. The Principal cares only about task 1 and obtains payoff $y = e_1 + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. The Agent, however, derives a benefit $v(e_2)$ from spending time on task 2. The Agent has CARA preferences with utility function

$$u(w, e_1, e_2) = -\exp\{-r[w - c(e_1 + e_2) + v(e_2)]\},$$

where $c(e_1 + e_2)$ is the cost of time, with $c'(\cdot) > 0$, $c''(\cdot) > 0$, and $c(0) = 0$. Assume also that $v'(\cdot) > 0$, $v''(\cdot) < 0$, and $v(0) = 0$, and that optimization with respect to (e_1, e_2) results in an interior solution. Let $\bar{w}$ denote the wage the Agent receives from his outside option, so $\bar{u} = -\exp\{-r\bar{w}\}$.

(a) What is the first-best outcome in this setting?

(b) Suppose effort e_1 is noncontractible, and the Principal can write a contract that is an affine function of output and can also allow the Agent to engage in task 2 or not. Under these assumptions, what is the contracting space?

(c) Suppose the Principal must pay the Agent $s = 1$ if $y = 0$. Will the Principal allow the Agent to engage in task 2? Compare this to your answer in part (a). What if $s < 1$ is set exogenously? Find the difference in the Principal's utility under the two policies, as a function of s.

Exercise 20. This exercise goes through a two-period version of Holmström and Milgrom's (1987) linear contracts argument. In each of two periods, $t \in \{1, 2\}$, the Agent chooses whether to "work" or to "shirk": $e_t \in \{0, 1\}$ at cost ce_t with $c > 0$. Output is binary, so that $y_t \in \{0, 1\}$, and the price of output is normalized to 1. Effort increases the probability that $y_t = 1$:

$$1 > \Pr\left[y_t = 1 \middle| e_t = 1\right] = p_H > p_L = \Pr\left[y_t = 1 \middle| e_t = 0\right] > 0.$$

The Agent's Bernoulli utility function is

$$u\left(w, e_1, e_2\right) = -\exp\left\{-r\left(w - ce_1 - ce_2\right)\right\},$$

and his outside option yields utility $-\exp\{-r \cdot 0\}$. The Agent can observe the realization of y_1 before choosing y_2.

The Principal's payoff is $y_1 + y_2 - w$, and the payment w can depend on each period's output and is paid at the end of period 2 (i.e., after both realizations of output). Assume it is optimal to induce the Agent to work hard in both periods. Show that a least-cost (optimal) contract that implements $e_1 = e_2 = 1$ has the form

$$w\left(y_1, y_2\right) = s + b\left(y_1 + y_2\right).$$

Guide:

(a) Define w_{y_1,y_2} to be the wage conditional on y_1 in period 1 and y_2 in period 2. Then, using the (IC) constraints for period 2, show that

$$e^{rc}\left[1 + p_H\left\{\frac{\exp\{-rw_{0,1}\}}{\exp\{-rw_{0,0}\}} - 1\right\}\right] = 1 + p_L\left\{\frac{\exp\{-rw_{0,1}\}}{\exp\{-rw_{0,0}\}} - 1\right\}$$

and

$$e^{rc}\left[1 + p_H\left\{\frac{\exp\{-rw_{1,1}\}}{\exp\{-rw_{1,0}\}} - 1\right\}\right] = 1 + p_L\left\{\frac{\exp\{-rw_{1,1}\}}{\exp\{-rw_{1,0}\}} - 1\right\}.$$

This implies there exists

$$b = w_{0,1} - w_{0,0} = w_{1,1} - w_{1,0}.$$

Why did CARA utility matter for this argument?

(b) Now, using the (IC) constraint for period 1, show that we have

$$e^{rc}\left[1 + p_H\left\{\frac{u_1}{u_0} - 1\right\}\right] = 1 + p_L\left\{\frac{u_1}{u_0} - 1\right\},$$

where u_i is the expected utility conditional on success in the first period ($i = 1$) or failure ($i = 0$).

(c) Note that

$$\exp\{r(c - w_{y,1})\} = \exp\{-rb\}\exp\{r(c - w_{y,0})\}$$

for each $y \in \{0, 1\}$. Now show that

$$\frac{u_1}{u_0} = \exp\{-r(w_{1,0} - w_{0,0})\} = \exp\{-r(w_{1,1} - w_{0,1})\}.$$

Therefore, we must have $b = w_{0,1} - w_{0,0} = w_{1,0} - w_{0,0} = w_{1,1} - w_{0,1}$.

3 The First-Order Approach

Last time, we imposed a lot of structure on the Principal-Agent problem and solved for optimal affine contracts. One of the problems we identified with that approach was that there was not a particularly compelling reason for restricting attention to affine contracts. Moreover, in that particular setting, if we allowed the contracts to take more general functional forms, there in fact was no optimal contract.

Today, we will return to a slightly modified version of the more general setup of the problem and consider an alternative approach to characterizing optimal contracts without imposing any assumptions on the functional forms they might take. One change we will be making is that the Agent's preferences are now given by

$$U(w, e) = \int_{y \in \mathcal{Y}} [u(w(y)) - c(e)]\, dF(y|\, e) = E_y[u(w)|\, e] - c(e),$$

where u is strictly increasing and strictly concave, and the utility the Agent receives from money is additively separable from his effort costs.

Recall from last time that the Principal's problem is to choose an output-contingent contract $w \in \mathcal{W} \subset \{w : \mathcal{Y} \to \mathbb{R}\}$ and to "propose" an effort level e to solve:

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e)$$

subject to an incentive-compatibility constraint

$$e \in \underset{\hat{e} \in \mathcal{E}}{\operatorname{argmax}} \int_{y \in \mathcal{Y}} u\left(w\left(y\right)\right) dF\left(y \middle| \hat{e}\right) - c\left(\hat{e}\right)$$

and an individual-rationality constraint

$$\int_{y \in \mathcal{Y}} u\left(w\left(y\right)\right) dF\left(y \middle| e\right) - c\left(e\right) \geq \bar{u}.$$

One of the problems with solving this problem at this level of generality is that the incentive-compatibility constraint is quite a complicated set of conditions. The contract has to ensure that, of all the effort levels the Agent could potentially choose, he prefers to choose e. In other words, the contract has to deter the Agent from choosing any other effort level $\hat{e}$: for all $\hat{e} \in \mathcal{E}$, we must have

$$\int_{y \in \mathcal{Y}} \left[u\left(w\left(y\right)\right) - c\left(e\right)\right] dF\left(y \middle| e\right) \geq \int_{y \in \mathcal{Y}} \left[u\left(w\left(y\right)\right) - c\left(\hat{e}\right)\right] dF\left(y \middle| \hat{e}\right).$$

When effort is continuous, the incentive-compatibility constraint is actually a continuum of constraints of this form. It seems like it should be the case that if we impose more structure on the problem, we can safely ignore most of these constraints. This turns out to be true. If we impose some relatively stringent but somewhat sensible assumptions on the problem, then if it is the case that the Agent does not want to deviate *locally* to another $\hat{e}$, then he also does not want to deviate to an $\hat{e}$ that is farther away. When local constraints are sufficient, we will in fact be able to replace the Agent's incentive-compatibility constraint with the first-order condition to his problem.

Throughout, we will be focusing on models that satisfy the following assumptions.

Assumption A1 (Continuous Effort and Continuous Output). Effort is continuous and satisfies $\mathcal{E} = \mathbb{R}_+$. Output is continuous, with $\mathcal{Y} = \mathbb{R}$, and for each $e \in \mathcal{E}$, $F\left(\cdot \middle| e\right)$ has support $\left[\underline{y}, \bar{y}\right]$ and has density $f\left(\cdot \middle| e\right)$, where $f\left(\cdot \middle| e\right)$ is differentiable in e.

Assumption A2 (First-Order Stochastic Dominance—FOSD). The output distribution function satisfies $F_e(y|e) \leq 0$ for all $e \in \mathcal{E}$ and all y with strict inequality for some y for each e.

Assumption $(A2)$ roughly says that higher effort levels make lower output realizations less likely and higher output realizations more likely. This assumption provides sufficient conditions under which higher effort increases total expected surplus, ignoring effort costs.

We will first explore the implications of being able to replace the incentive-compatibility constraint with the Agent's first-order condition, and then we will provide some sufficient conditions under which doing so is without loss of generality. Under Assumption $(A1)$, if we replace the Agent's incentive-compatibility constraint with his first-order condition, the Principal's problem becomes:

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{\underline{y}}^{\bar{y}} (py - w(y)) f(y|e) \, dy$$

subject to the local incentive-compatibility constraint

$$c'(e) = \int_{\underline{y}}^{\bar{y}} u(w(y)) f_e(y|e) \, dy$$

and the individual-rationality constraint

$$\int_{\underline{y}}^{\bar{y}} u(w(y)) f(y|e) \, dy - c(e) \geq \bar{u}.$$

This problem is referred to as the **first-order approach** to characterizing second-best incentive contracts. It is now just a constrained-optimization problem with an equality constraint and an inequality constraint. We can therefore write the Lagrangian for this

problem as

$$\begin{aligned}\mathcal{L} &= \int_{\underline{y}}^{\bar{y}} (py - w(y)) f(y|e)\, dy + \lambda \left(\int_{\underline{y}}^{\bar{y}} u(w(y)) f(y|e)\, dy - c(e) - \bar{u} \right) \\ &\quad + \mu \left(\int_{\underline{y}}^{\bar{y}} u(w(y)) f_e(y|e)\, dy - c'(e) \right),\end{aligned}$$

where λ is the Lagrange multiplier on the individual-rationality constraint, and μ is the Lagrange multiplier on the local incentive-compatibility constraint. We can derive the conditions for the optimal contract $w^*(y)$ inducing optimal effort e^* by taking first-order conditions, point-by-point, with respect to $w(y)$. These conditions are:

$$\frac{1}{u'(w^*(y))} = \lambda + \mu \frac{f_e(y|e^*)}{f(y|e^*)}.$$

Contracts satisfying these conditions are referred to as Holmström-Mirrlees contracts (or (λ, μ) contracts as one of my colleagues calls them). There are several points to notice here. First, the left-hand side is increasing in $w(y)$, since u is concave. Second, if $\mu = 0$, then this condition would correspond to the conditions for an optimal risk-sharing rule between the Principal and the Agent. Under a Pareto-optimal risk allocation, the **Borch Rule** states that the ratio of the Principal's marginal utility to the Agent's marginal utility is equalized across states. In this case, the Principal's marginal utility is one. Any optimal-risk sharing rule will equalize the Agent's marginal utility of income across states and therefore give the Agent a constant wage.

Third, Holmström (1979) shows that under Assumption $(A2)$, $\mu > 0$, so that the right-hand side of this equation is increasing in $f_e(y|e^*)/f(y|e^*)$. You might remember from econometrics that this ratio is called the **score**—it tells us how an increase in e changes the log likelihood of e given output realization y. To prevent the Agent from choosing effort level e instead of e^*, the contract has to pay the Agent more for outputs that are more likely under e^* than under e. Since by assumption, we are looking at only local incentive constraints, the

contract will pay the Agent more for outputs that are more likely under e^* than under effort levels arbitrarily close to e^*.

Together, these observations imply that the optimal contract $w^*(y)$ is increasing in the score. Just because an optimal contract is increasing in the score does not mean that it is increasing in output. The following assumption guarantees that the score is increasing in y, and therefore optimal contracts are increasing in output.

Assumption A3 (Monotone Likelihood Ratio Property—MLRP). Given any two effort levels $e, e' \in \mathcal{E}$ with $e > e'$, the ratio $f(y|\, e)/f(y|\, e')$ is increasing in y.

MLRP guarantees, roughly speaking, that higher levels of output are more indicative of higher effort levels.[1] Under Assumption $(A1)$, MLRP is equivalent to the condition that $f_e(y|\, e)/f(y|\, e)$ is increasing in y. We can therefore interpret the optimality condition as telling us that the optimal contract is increasing in output precisely when higher output levels are more indicative of higher effort levels. Put differently, the optimal contract "wants" to reward *informative* output, not necessarily *high* output.

The two statistical properties, FOSD and MLRP, that we have assumed come up a lot in different settings, and it is easy to lose track of what they each imply. To recap, the FOSD property tells us that higher effort makes higher output more likely, and it guarantees that there is always a benefit of higher effort levels, gross of effort costs. The MLRP property tells us that higher output is more indicative of higher effort, and it guarantees that optimal contracts are increasing in output. These two properties are related: MLRP implies FOSD, but not the reverse.

[1] The property can also be interpreted in terms of statistical hypothesis testing. Suppose the null hypothesis is that the Agent chose effort level e', and the alternative hypothesis is that the Agent chose effort level $e > e'$. If, given output realization y, a likelihood ratio test would reject the null hypothesis of lower effort, the same test would also reject the null hypothesis for any higher output realization.

3.1 Informativeness Principle

Before we provide conditions under which the first-order approach is valid, we will go over what I view as the most important result to come out of this model. Suppose there is another contractible performance measure $m \in \mathcal{M}$, where y and m have joint density function $f(y, m|\, e)$, and the contracting space is $\mathcal{W} = \{w : \mathcal{Y} \times \mathcal{M} \to \mathbb{R}\}$. Under what conditions will an optimal contract $w(y, m)$ depend nontrivially on m? The answer is: whenever m provides additional information about e. To make this argument precise, we will introduce the following definition.

Definition 1. Given two random variables Y and M, Y is **sufficient for** (Y, M) **with respect to** $e \in \mathcal{E}$ if and only if the joint density function $f(y, m|\, e)$ is multiplicatively separable in m and e:

$$f(y, m|\, e) = g(m|\, e)\, h(y, m).$$

We will say that M is **informative about** $e \in \mathcal{E}$ if Y is not sufficient for (Y, M) with respect to $e \in \mathcal{E}$.

We argued above that optimal contracts pay the Agent more for outputs that are more indicative of high effort. This same argument also extends to other performance measures, as long as they are informative about effort. This result is known as the *informativeness principle* and was first established by Holmström (1979) and Shavell (1979).

Theorem 1 (Informativeness Principle). Assume the first-order approach is valid. Let $w(y)$ be the optimal contract when m is noncontractible. If m is contractible, there exist a contract $w(y, m)$ that Pareto dominates $w(y)$ if and only if m is informative about $e \in \mathcal{E}$.

Proof. In both cases, the optimal contract gives the Agent $\bar{u}$, so we just need to show that the Principal can be made strictly better off if m is contractible.

If the first-order approach is valid, the optimality conditions for the Principal's problem

when both y and m are contractible are given by

$$\frac{1}{u'\left(w^*\left(y,m\right)\right)} = \lambda + \mu \frac{f_e\left(y,m\middle|\, e^*\right)}{f\left(y,m\middle|\, e^*\right)}.$$

The optimal contract $w^*\left(y,m\right)$ is independent of m if and only if y is sufficient for (y,m) with respect to e^*.

This result seems like it should be obvious: optimal contracts clearly should make use of all available information. But it is not ex ante obvious this would be the case. In particular, one could easily have imagined that optimal contracts should only depend on performance measures that are "sufficiently" informative about effort—after all, basing a contract on another performance measure could introduce additional noise as well. Or one could have imagined that optimal contracts should only depend on performance measures that are directly affected by the Agent's effort choice. The informativeness principle says that optimal contracts should depend on every performance measure that is even slightly informative.

This result has both positive and negative implications. On the positive and practical side, it says that optimal contracts should make use of benchmarks: a fund manager should be evaluated for her performance relative to a market index, CEOs should be rewarded for firm performance relative to other firms in their industry, and employees should be evaluated relative to their peers. On the negative side, the result shows that optimal contracts are highly sensitive to the fine details of the environment. This implication is, in a real sense, a weakness of the theory: it is the reason why the theory often predicts contracts that bear little resemblance to what we actually see in practice.

The informativeness principle was derived under the assumption that the first-order approach was valid. When the first-order approach is not valid, the informativeness principle does not necessarily hold. The reason for this is that when the first-order approach does not hold, there may be multiple binding incentive-compatibility constraints at the optimum,

and just because an informative performance measure helps relax one of those constraints, if it does not help relax the other binding constraints, it need not strictly increase the firm's profits. Chaigneau, Edmans, and Gottlieb (2014) generalizes the informativeness principle to settings in which the first-order approach is not valid.

3.2 Validity of the First-Order Approach

Finally, we will briefly talk about some sufficient conditions ensuring the first-order approach is valid. Assumption $(A4)$, along with the following assumption, are sufficient.

Assumption A4 (Convexity of the Distribution Function Condition—CDFC). $F(\cdot|\, e)$ is twice differentiable, and $F_{ee}(\cdot|\, e) \geq 0$ for all e.

CDFC is a strong assumption. There is a fairly standard class of distributions that are often used in contract theory that satisfy it, but it is not satisfied by other well-known families of distributions. Let $F_H(y)$ and $F_L(y)$ be two distribution functions that have density functions $f_H(y)$ and $f_L(y)$ for which $f_H(y)/f_L(y)$ is increasing in y, and suppose

$$F(y|\, e) = eF_H(y) + (1-e)F_L(y).$$

Then $F(y|\, e)$ satisfies both MLRP and CDFC. In other words, MLRP and CDFC are satisfied if output is drawn from a mixture of a "high" and a "low" distribution, and higher effort increases the probability that output is drawn from the high distribution.

Theorem 2. Suppose $(A1)-(A4)$ are satisfied. If the local incentive-compatibility constraint is satisfied, the incentive-compatibility constraint is satisfied.

Proof sketch. The high-level idea of the proof is to show that MLRP and CDFC imply that the Agent's effort-choice problem is globally concave for any contract the Principal offers

him. Using integration by parts, we can rewrite the Agent's expected utility as follows.

$$\begin{aligned}\int_{\underline{y}}^{\bar{y}} u\left(w\left(y\right)\right) f\left(y|\, e\right) dy - c\left(e\right) &= u\left(w\left(y\right)\right) F\left(y|\, e\right)|_{\underline{y}}^{\bar{y}} \\ &\quad - \int_{\underline{y}}^{\bar{y}} u'\left(w\left(y\right)\right) \frac{dw\left(y\right)}{dy} F\left(y|\, e\right) dy - c\left(e\right) \\ &= u\left(w\left(\bar{y}\right)\right) - \int_{\underline{y}}^{\bar{y}} u'\left(w\left(y\right)\right) \frac{dw\left(y\right)}{dy} F\left(y|\, e\right) dy - c\left(e\right).\end{aligned}$$

Now, suppose $w\left(y\right)$ is increasing and differentiable. Differentiating the expression above with respect to e twice yields

$$- \int_{\underline{y}}^{\bar{y}} u'\left(w\left(y\right)\right) \frac{dw\left(y\right)}{dy} F_{ee}\left(y|\, e\right) dy - c''\left(e\right) < 0$$

for every $e \in \mathcal{E}$, since $F_{ee} > 0$. Thus, the Agent's second-order condition is globally satisfied, so if the local incentive constraint is satisfied, the incentive constraint is satisfied.■

I labeled this proof as a sketch, because while it follows Mirrlees's (1976) argument, the full proof (due to Rogerson (1985)) requires showing that $w\left(y\right)$ is in fact increasing and differentiable when MLRP is satisfied. We cannot use our argument above for why MLRP implies increasing contracts, because that argument presumed the first-order approach was valid, which is exactly what we are trying to prove here. The MLRP and CDFC conditions are known as the Mirrlees-Rogerson conditions.

There are other sufficient conditions for the first-order approach to be valid that do not require such strong distributional assumptions (see, for example, Jewitt (1988)). And there are other approaches to solving the moral hazard problem that do not rely on the first-order approach. These include Grossman and Hart (1983), which decomposes the Principal's problem into two steps: the first step solves for the cost-minimizing contract that implements a given effort level, and the second step solves for the optimal effort level. We will take this approach when we think about optimal contracts under limited liability in the next section.

4 Limited Liability and the Motivation–Rent Extraction Trade-Off

We saw in the previous model that the optimal contract sometimes involved up-front payments from the Agent to the Principal. To the extent that the Agent is unable to afford such payments (or legal restrictions such as minimum wage laws prohibit such payments), the Principal will not be able to extract all the surplus that the Agent creates. Further, in order to extract surplus from the Agent, the Principal may have to put in place contracts that reduce the total surplus created. In equilibrium, the Principal may therefore offer a contract that induces effort below the first-best.

Description Again, there is a risk-neutral Principal (P). There is also a **risk-neutral** Agent (A). The Agent chooses an effort level $e \in \mathcal{E} \subset \mathbb{R}_+$ at a cost of $c(e)$, where $c : \mathbb{R}_+ \to \mathbb{R}_+$, with $c'', c' > 0$, and this effort level affects the distribution over outputs $y \in \mathcal{Y}$, with y distributed according to CDF $F(\cdot|\, e)$. These outputs can be sold on the product market for price p. The Principal can write a contract $w \in \mathcal{W} \subset \{w : \mathcal{Y} \to \mathbb{R}, w(y) \geq \underline{w} \text{ for all } y\}$ that determines a transfer $w(y)$ that she is compelled to pay the Agent if output y is realized. The Agent has an outside option that provides utility $\bar{u}$ to the Agent and $\bar{\pi}$ to the Principal. If the outside option is not exercised, the Principal's and Agent's preferences are, respectively,

$$\begin{aligned} \Pi(w, e) &= \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e) = E_y[py - w|\, e] \\ U(w, e) &= \int_{y \in \mathcal{Y}} (w(y) - c(e))\, dF(y|\, e) = E_y[w - c(e)|\, e]. \end{aligned}$$

There are two differences between this model and the previous model. The first difference is that the Agent is risk-neutral (so that absent any other changes, the equilibrium contract would induce first-best effort). The second difference is that the wage payment from the Principal to the Agent has to exceed, for each realization of output, a value $\underline{w}$. Depending

on the setting, this constraint is described as a liquidity constraint or a limited-liability constraint. In repeated settings, it is more naturally thought of as the latter—due to legal restrictions, the Agent cannot be legally compelled to make a transfer (larger than $-\underline{w}$) to the Principal. In static settings, either interpretation may be sensible depending on the particular application—if the Agent is a fruit picker, for instance, he may not have much liquid wealth that he can use to pay the Principal.

Timing The timing of the game is exactly the same as before.

1. P offers A a contract $w(y)$, which is commonly observed.
2. A accepts the contract ($d = 1$) or rejects it ($d = 0$) and receives $\bar{u}$, and the game ends. This decision is commonly observed.
3. If A accepts the contract, A chooses effort level e and incurs cost $c(e)$. e is only observed by A.
4. Output y is drawn from distribution with cdf $F(\cdot|\, e)$. y is commonly observed.
5. P pays A an amount $w(y)$. This payment is commonly observed.

Equilibrium The solution concept is the same as before. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0, 1\}$, and an effort choice $e^* : \mathcal{W} \times \{0, 1\} \to \mathbb{R}_+$ such that given the contract w^*, the Agent optimally chooses d^* and e^*, and given d^* and e^*, the Principal optimally offers contract w^*. We will say that the optimal contract induces effort e^*.

The Program The Principal offers a contract $w \in \mathcal{W}$ and proposes an effort level e in order to solve

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e)$$

subject to three constraints: the incentive-compatibility constraint

$$e \in \underset{\hat{e} \in \mathcal{E}}{\operatorname{argmax}} \int_{y \in \mathcal{Y}} \left(w\left(y\right) - c\left(\hat{e}\right) \right) dF\left(y \middle| \hat{e} \right),$$

the individual-rationality constraint

$$\int_{y \in \mathcal{Y}} \left(w\left(y\right) - c\left(e\right) \right) dF\left(y \middle| e \right) \geq \bar{u},$$

and the limited-liability constraint

$$w\left(y\right) \geq \underline{w} \text{ for all } y \in \mathcal{Y}.$$

Binary-Output Case We will impose much more structure on the problem to illustrate the main trade-off in this class of models. Innes (1990) and Jewitt, Kadan, and Swinkels (2008) explore a much more general analysis.

Assumption A1 (Binary Output). Output is $y \in \{0, 1\}$, and given effort e, its distribution satisfies $\Pr\left[y = 1 \middle| e \right] = e$.

Assumption A2 (Well-behaved Cost). The Agent's costs have a non-negative third derivative: $c''' \geq 0$, and they satisfy conditions that ensure an interior solution: $c'\left(0\right) = 0$ and $c'\left(1\right) = +\infty$. Or for comparison across models in this module, $c\left(e\right) = \frac{c}{2} e^2$, where $p \leq c$ to ensure that $e^{FB} < 1$.

Finally, we can restrict attention to affine, nondecreasing contracts

$$\begin{aligned} \mathcal{W} &= \left\{ w\left(y\right) = \left(1 - y\right) w_0 + y w_1, w_1 \geq w_0 \geq 0 \right\} \\ &= \left\{ w\left(y\right) = s + by, s \geq \underline{w}, b \geq 0 \right\}. \end{aligned}$$

When output is binary, this restriction to affine contracts is without loss of generality. Also, the restriction to nondecreasing contracts is not restrictive (i.e., any optimal contract of a

relaxed problem in which we do not impose that contracts are nondecreasing will also be the solution to the full problem). This result is something that needs to be shown and is not in general true, but in this case, it is straightforward.

As Grossman and Hart (1983) highlight, in Principal–Agent models, it is often useful to break the problem down into two steps. The first step takes a target effort level, e, as given and solves for the set of cost-minimizing contracts implementing effort level e. Any cost-minimizing contract implementing effort level e results in an expected cost of $C(e)$ to the principal. The second step takes the function $C(\cdot)$ as given and solves for the optimal effort choice.

In general, the cost-minimization problem tends to be a well-behaved convex-optimization problem, since (even if the agent is risk-averse) the objective function is weakly concave, and the constraint set is a convex set (since given an effort level e, the individual-rationality constraint and the limited-liability constraint define convex sets, and each incentive constraint ruling out effort level $\hat{e} \neq e$ also defines a convex set, and the intersection of convex sets is itself a convex set). The resulting cost function $C(\cdot)$ need not have nice properties, however, so the second step of the optimization problem is only well-behaved under restrictive assumptions. In the present case, Assumptions $(A1)$ and $(A2)$ ensure that the second step of the optimization problem is well-behaved.

Cost-Minimization Problem Given an effort level e, the cost-minimization problem is given by

$$C(e, \bar{u}, \underline{w}) = \min_{s,b} s + be$$

subject to the Agent's incentive-compatibility constraint

$$e \in \operatorname*{argmax}_{\hat{e}} \{s + b\hat{e} - c(\hat{e})\},$$

his individual-rationality constraint

$$s + be - c(e) \geq \bar{u},$$

and the limited-liability constraint

$$s \geq \underline{w}.$$

I will denote a **cost-minimizing contract implementing effort level** e by (s_e^*, b_e^*).

The first step in solving this problem is to notice that the Agent's incentive-compatibility constraint implies that any cost-minimizing contract implementing effort level e must have $b_e^* = c'(e)$.

If there were no limited-liability constraint, the Principal would choose s_e^* to extract the Agent's surplus. That is, given $b = b_e^*$, s would solve

$$s + b_e^* e = \bar{u} + c(e).$$

That is, s would ensure that the Agent's expected compensation exactly equals his expected effort costs plus his opportunity cost. The resulting s, however, may not satisfy the limited-liability constraint. The question then is: given $\bar{u}$ and $\underline{w}$, for what effort levels e is the Principal able to extract all the agent's surplus (i.e., for what effort levels does the limited-liability constraint not bind at the cost-minimizing contract?), and for what effort levels is she unable to do so? Figure 1 below shows cost-minimizing contracts for effort levels e_1 and e_2. Any contract can be represented as a line in this figure, where the line represents the expected pay the Agent will receive given an effort level e. The cost-minimizing contract for effort level e_1 is tangent to the $\bar{u} + c(e)$ curve at e_1 and its intercept is $s_{e_1}^*$. Similarly for e_2. Both $s_{e_1}^*$ and $s_{e_2}^*$ are greater than $\underline{w}$, which implies that for such effort levels, the

limited-liability constraint is not binding.

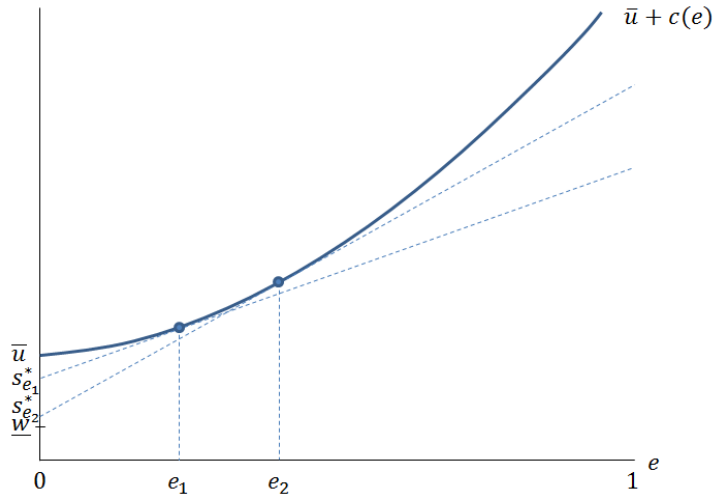

Figure 1: Cost-minimizing contracts

For effort sufficiently high, the limited-liability constraint will be binding in a cost-minimizing contract, and it will be binding for all higher effort levels. Define the threshold $\bar{e}(\bar{u}, \underline{w})$ to be the effort level such that for all $e \geq \bar{e}(\bar{u}, \underline{w})$, $s^*_e = \underline{w}$. Figure 2 illustrates that $\bar{e}(\bar{u}, \underline{w})$ is the effort level at which the contract tangent to the $\bar{u} + c(e)$ curve at $\bar{e}(\bar{u}, \underline{w})$ intersects the vertical axis at exactly $\underline{w}$. That is, $\bar{e}(\bar{u}, \underline{w})$ solves

$$c'\left(\bar{e}\left(\bar{u}, \underline{w}\right)\right) = \frac{\bar{u} + c\left(\bar{e}\left(\bar{u}, \underline{w}\right)\right) - \underline{w}}{\bar{e}\left(\bar{u}, \underline{w}\right)}.$$

Figure 2 also illustrates that for all effort levels $e > \bar{e}(\bar{u}, \underline{w})$, the cost-minimizing contract involves giving the Agent strictly positive surplus. That is, the cost to the Principal of getting the agent to choose effort $e > \bar{e}(\bar{u}, \underline{w})$ is equal to the Agent's opportunity costs $\bar{u}$

plus his effort costs $c(e)$ plus **incentive costs** $IC(e, \bar{u}, \underline{w})$.

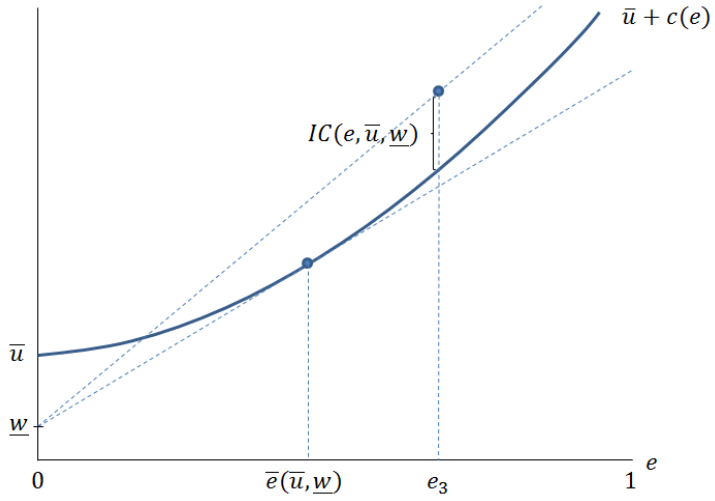

Figure 2: Incentive Costs for High Effort Levels

The incentive costs $IC(e, \bar{u}, \underline{w})$ are equal to the Agent's expected compensation given effort choice e and cost-minimizing contract (s_e^*, b_e^*) minus his costs:

$$\begin{aligned} IC(e, \bar{u}, \underline{w}) &= \begin{cases} 0 & e \leq \bar{e}(\bar{u}, \underline{w}) \\ \underline{w} + c'(e)\, e - c(e) - \bar{u} & e \geq \bar{e}(\bar{u}, \underline{w}) \end{cases} \\ &= \max\{0, \underline{w} + c'(e)\, e - c(e) - \bar{u}\} \end{aligned}$$

where I used the fact that for $e \geq \bar{e}(\bar{u}, \underline{w})$, $s_e^* = \underline{w}$ and $b_e^* = c'(e)$. This incentive-cost function $IC(\cdot, \bar{u}, \underline{w})$ is the key object that captures the main contracting friction in this model. I will sometimes refer to $IC(e, \bar{u}, \underline{w})$ as the **incentive rents** required to get the Agent to choose effort level e. Putting these results together, we see that

$$C(e, \bar{u}, \underline{w}) = \bar{u} + c(e) + IC(e, \bar{u}, \underline{w}).$$

That is, the Principal's total costs of implementing effort level e are the sum of the Agent's

costs plus the incentive rents required to get the Agent to choose effort level e.

Since $IC(e, \bar{u}, \underline{w})$ is the main object of interest in this model, I will describe some of its properties. First, it is continuous in e (including, in particular, at $e = \bar{e}(\bar{u}, \underline{w})$). Next, $\bar{e}(\bar{u}, \underline{w})$ and $IC(e, \bar{u}, \underline{w})$ depend on $(\bar{u}, \underline{w})$ only inasmuch as $(\bar{u}, \underline{w})$ determines $\bar{u} - \underline{w}$, so I will abuse notation and write these expressions as $\bar{e}(\bar{u} - \underline{w})$ and $IC(e, \bar{u} - \underline{w})$. Also, given that $c'' > 0$, IC is increasing in e (since $\underline{w} + c'(e)\, e - c(e) - \underline{u}$ is strictly increasing in e, and IC is just the max of this expression and zero). Further, given that $c''' \geq 0$, IC is convex in e. For $e \geq \bar{e}(\bar{u} - \underline{w})$, this property follows, because

$$\frac{\partial^2}{\partial e^2} IC = c''(e) + c'''(e)\, e \geq 0.$$

And again, since IC is the max of two convex functions, it is also a convex function. Finally, since $IC(\cdot, \bar{u} - \underline{w})$ is flat when $e \leq \bar{e}(\bar{u} - \underline{w})$ and it is strictly increasing (with slope independent of $\bar{u} - \underline{w}$) when $e \geq \bar{e}(\bar{u} - \underline{w})$, the slope of IC with respect to e is (weakly) decreasing in $\bar{u} - \underline{w}$, since $\bar{e}(\bar{u} - \underline{w})$ is increasing in $\bar{u} - \underline{w}$. That is, $IC(e, \bar{u} - \underline{w})$ satisfies decreasing differences in $(e, \bar{u} - \underline{w})$.

Motivation-Rent Extraction Trade-off The second step of the optimization problem takes as given the function

$$C(e, \bar{u} - \underline{w}) = \bar{u} + c(e) + IC(e, \bar{u} - \underline{w})$$

and solves the Principal's problem for the optimal effort level:

$$\begin{aligned} & \max_e pe - C(e, \bar{u} - \underline{w}) \\ = \; & \max_e pe - \bar{u} - c(e) - IC(e, \bar{u} - \underline{w}). \end{aligned}$$

Note that total surplus is given by $pe - \bar{u} - c(e)$, which is therefore maximized at $e = e^{FB}$ (which, if $c(e) = ce^2/2$, then $e^{FB} = p/c$). Figure 3 below depicts the Principal's expected benefit line pe, and her expected costs of implementing effort e at minimum cost, $C(e, \bar{u} - \underline{w})$. The first-best effort level, e^{FB} maximizes the difference between pe and $\bar{u} + c(e)$, while the equilibrium effort level e^* maximizes the difference between pe and $C(e, \bar{u} - \underline{w})$.

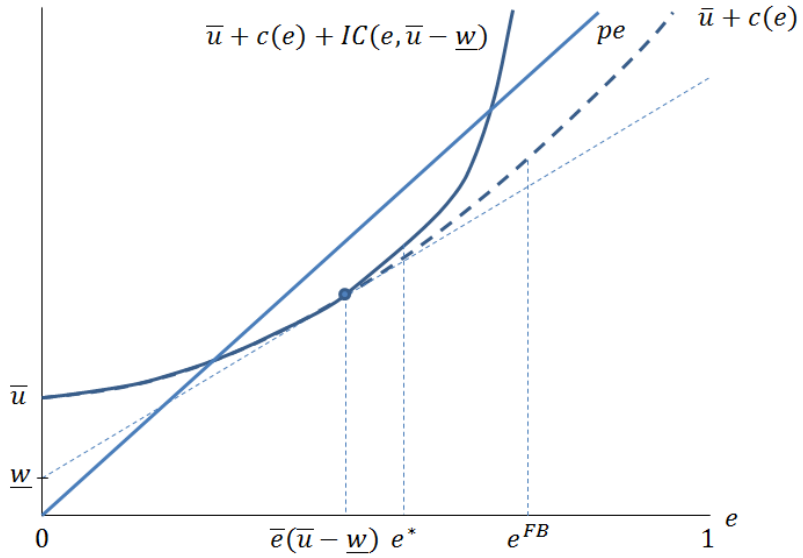

Figure 3: Optimal Effort Choice

If $c(e) = ce^2/2$, we can solve explicitly for $\bar{e}(\bar{u} - \underline{w})$ and for $IC(e, \bar{u} - \underline{w})$ when $e > \bar{e}(\bar{u} - \underline{w})$. In particular,

$$\bar{e}(\bar{u} - \underline{w}) = \left(\frac{2(\bar{u} - \underline{w})}{c}\right)^{1/2}$$

and when $e > \bar{e}(\bar{u} - \underline{w})$,

$$IC(e, \bar{u} - \underline{w}) = \underline{w} + \frac{1}{2}ce^2 - \bar{u}.$$

If $\underline{w} < 0$ and p is sufficiently small, we can have $e^* = e^{FB}$ (i.e., these are the conditions required to ensure that the limited-liability constraint is not binding for the cost-minimizing contract implementing $e = e^{FB}$). If p is sufficiently large relative to $\bar{u} - \underline{w}$, we will have $e^* = \frac{1}{2}\frac{p}{c} = \frac{1}{2}e^{FB}$. For p somewhere in between, we will have $e^* = \bar{e}(\bar{u} - \underline{w}) < e^{FB}$. In

particular, $C(e, \bar{u} - \underline{w})$ is kinked at this point.

As in the risk–incentives model, we can illustrate through a partial characterization why (and when) effort is less-than first-best. Since we know that e^{FB} maximizes $pe - \bar{u} - c(e)$, we therefore have that

$$\frac{d}{de}\left[pe - \bar{u} - c(e) - IC(e, \bar{u} - \underline{w})\right]_{e=e^{FB}} = -\frac{\partial}{\partial e} IC\left(e^{FB}, \bar{u} - \underline{w}\right) \leq 0,$$

with strict inequality if the limited-liability constraint binds at the cost-minimizing contract implementing e^{FB}. This means that, even though e^{FB} maximizes total surplus, if the Principal has to provide the agent with rents at the margin, she may choose to implement a lower effort level. Reducing the effort level away from e^{FB} leads to second-order losses in terms of total surplus, but it leads to first-order gains in profits for the Principal. In this model, there is a tension between total-surplus creation and rent extraction, which yields less-than-first-best effort in equilibrium.

In my view, liquidity constraints are extremely important and are probably one of the main reasons for why many jobs do not involve first-best incentives. The logic that first-best efforts can be implemented if the firm transfers the entire profit stream to each of its members in exchange for a large up-front payment seems simultaneously compelling, trivial, and obviously impracticable. In for-profit firms, in order to make it worthwhile to transfer a large enough share of the profit stream to an individual worker to significantly affect his incentives, the firm would require a large up-front transfer that most workers cannot afford to pay. It is therefore not surprising that we do not see most workers' compensation tied directly to the firm's overall profits in a meaningful way. One implication of this logic is that firms have to find alternative instruments to use as performance measures, which we will turn to next. In principle, models in which firms do not motivate their workers by writing contracts directly on profits should include assumptions under which the firm optimally chooses not to write contracts directly on profits, but they almost never do.

Exercise 21. This exercise goes through a version of Diamond's (1998) and Barron, Georgiadis, and Swinkels's (2018) argument for why linear contracts are optimal when the Agent is able to "take on risk." Suppose the Principal and the Agent are both risk neutral, and let $\mathcal{Y} = [0, \bar{y}]$ and $\mathcal{E} = \mathbb{R}_+$. There is a limited-liability constraint, and the contracting space is $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}_+\}$. After the Agent chooses an effort level e, he can then choose any distribution function $F(y)$ over output that satisfies $e = \int_0^{\bar{y}} y dF(y)$. In other words, his effort level determines his *average* output, but he can then add mean-preserving noise to his output. Given a contract w, effort e, and distribution F, the Agent's expected utility is

$$\int_0^{\bar{y}} w(y) \, dF(y) - c(e),$$

where c is strictly increasing and strictly convex. The Principal's expected profits are $\int_0^{\bar{y}} (y - w(y)) \, dF(y)$. The Agent's outside option gives both parties a payoff of zero.

(a) Show that a linear contract of the form $w(y) = by$ maximizes the Principal's expected profits. To do so, you will want to argue that given any contract $w(y)$ that implements effort level e, there is a linear contract that also implements effort level e but at a weakly lower cost to the Principal. [Hint: instead of thinking about all the possible distribution functions the Agent can choose among, it may be useful to just look at distributions that put weight on two levels of output, $0 \leq y_L < y_H \leq \bar{y}$ satisfying $e = (1 - q) y_L + q y_H$.]

(b) Are there other contracts that maximize the Principal's expected profits? If so, how are they related to the optimal linear contract? If not, provide an intuition for why linear contracts are uniquely optimal.

5 Misaligned Performance Measures

In the previous two models, the Principal cared about output, and output, though a noisy measure of effort, was perfectly measurable. This assumption seems sensible when we think about overall firm profits (ignoring basically everything that accountants think about every day), but as we alluded to in the previous discussion, overall firm profits are too blunt of an instrument to use to motivate individual workers within the firm if they are liquidity-constrained. As a result, firms often try to motivate workers using more specific performance measures, but while these performance measures are informative about what actions workers are taking, they may be less useful as a description of how the workers' actions affect the objectives the firm cares about. And paying workers for what is measured may not get them to take actions that the firm cares about. This observation underpins the title of the famous

1975 paper by Steve Kerr called "On the Folly of Rewarding A, While Hoping for B."

As an example, think of a retail firm that hires an employee both to make sales and to provide customer service. It can be difficult to measure the quality of customer service that a particular employee provides, but it is easy to measure that employee's sales. Writing a contract that provides the employee with high-powered incentives directly on sales will get him to put a lot of effort into sales and very little effort into customer service. And in fact, he might only be able to put a lot of effort into sales by intentionally neglecting customer service. If the firm cares equally about both dimensions, it might be optimal not to offer high-powered incentives to begin with. This is what Holmström and Milgrom (1991) refers to as the "multitask problem." We will look at a model that captures some of this intuition, although not as directly as Holmström and Milgrom's model.

Description Again, there is a risk-neutral Principal (P) and a risk-neutral Agent (A). The Agent chooses an effort vector $e = (e_1, e_2) \in \mathcal{E} \subset \mathbb{R}^2_+$ at a cost of $\frac{c}{2}(e_1^2 + e_2^2)$. This effort vector affects the distribution of output $y \in \mathcal{Y} = \{0, 1\}$ and a performance measure $m \in \mathcal{M} = \{0, 1\}$ as follows:

$$\begin{aligned} \Pr[y = 1 | e] &= f_1 e_1 + f_2 e_2 \\ \Pr[m = 1 | e] &= g_1 e_1 + g_2 e_2, \end{aligned}$$

where it may be the case that $f = (f_1, f_2) \neq (g_1, g_2) = g$. Assume that $f_1^2 + f_2^2 = g_1^2 + g_2^2 = 1$ (i.e., the norms of the f and g vectors are unity). The output can be sold on the product market for price p. Output is noncontractible, but the performance measure is contractible. The Principal can write a contract $w \in \mathcal{W} \subset \{w : \mathcal{M} \to \mathbb{R}\}$ that determines a transfer $w(m)$ that she is compelled to pay the Agent if performance measure m is realized. Since the performance measure is binary, contracts take the form $w = s + bm$. The Agent has an outside option that provides utility $\bar{u}$ to the Agent and $\bar{\pi}$ to the Principal. If the outside

option is not exercised, the Principal's and Agent's preferences are, respectively,

$$\begin{aligned}\Pi(w,e) &= f_1e_1 + f_2e_2 - s - b(g_1e_1 + g_2e_2)\\ U(w,e) &= s + b(g_1e_1 + g_2e_2) - \frac{c}{2}\left(e_1^2 + e_2^2\right).\end{aligned}$$

Timing The timing of the game is exactly the same as before.

1. P offers A a contract w, which is commonly observed.
2. A accepts the contract ($d = 1$) or rejects it ($d = 0$) and receives $\bar{u}$ and the game ends. This decision is commonly observed.
3. If A accepts the contract, A chooses effort vector e. e is only observed by A.
4. Performance measure m and output y are drawn from the distributions described above. m is commonly observed.
5. P pays A an amount $w(m)$. This payment is commonly observed.

Equilibrium The solution concept is the same as before. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0,1\}$, and an effort choice $e^* : \mathcal{W} \times \{0,1\} \to \mathbb{R}^2_+$ such that given the contract w^*, the Agent optimally chooses d^* and e^*, and given d^* and e^*, the Principal optimally offers contract w^*. We will say that the optimal contract induces effort e^*.

The Program The principal offers a contract w and proposes an effort level e to solve

$$\max_{s,b,e} p(f_1e_1 + f_2e_2) - (s + b(g_1e_1 + g_2e_2))$$

subject to the incentive-compatibility constraint

$$e \in \operatorname*{argmax}_{\hat{e}\in\mathbb{R}^2_+} s + b(g_1\hat{e}_1 + g_2\hat{e}_2) - \frac{c}{2}\left(\hat{e}_1^2 + \hat{e}_2^2\right)$$

and the individual-rationality constraint

$$s + b\left(g_1 e_1 + g_2 e_2\right) - \frac{c}{2}\left(e_1^2 + e_2^2\right) \geq \bar{u}.$$

Equilibrium Contracts and Effort Given a contract $s + bm$, the Agent will choose

$$e_1^*(b) = \frac{b}{c} g_1;\ e_2^*(b) = \frac{b}{c} g_2.$$

The Principal will choose s so that the individual-rationality constraint holds with equality

$$s + b\left(g_1 e_1^*(b) + g_2 e_2^*(b)\right) = \bar{u} + \frac{c}{2}\left(e_1^*(b)^2 + e_2^*(b)^2\right).$$

Since contracts send the Agent off in the "wrong direction" relative to what maximizes total surplus, providing the Agent with higher-powered incentives by increasing b sends the agent farther off in the wrong direction. This is costly for the Principal because in order to get the Agent to accept the contract, she has to compensate him for his effort costs, even if they are in the wrong direction.

The Principal's unconstrained problem is therefore

$$\max_b p\left(f_1 e_1^*(b) + f_2 e_2^*(b)\right) - \frac{c}{2}\left(e_1^*(b)^2 + e_2^*(b)^2\right) - \bar{u}.$$

Taking first-order conditions,

$$p f_1 \underbrace{\frac{\partial e_1^*}{\partial b}}_{g_1/c} + p f_2 \underbrace{\frac{\partial e_2^*}{\partial b}}_{g_2/c} = c\underbrace{e_1^*(b^*)}_{b^* g_1/c}\underbrace{\frac{\partial e_1^*}{\partial b}}_{g_1/c} + c\underbrace{e_2^*(b^*)}_{b^* g_2/c}\underbrace{\frac{\partial e_2^*}{\partial b}}_{g_2/c},$$

or

$$b^* = p\frac{f_1 g_1 + f_2 g_2}{g_1^2 + g_2^2} = p\frac{f \cdot g}{g \cdot g} = p\frac{||f||}{||g||}\cos\theta = p\cos\theta,$$

where $\cos\theta$ is the angle between the vectors f and g. That is, the optimal incentive slope

depends on the relative magnitudes of the f and g vectors (which in this model were assumed to be the same, but in a richer model this need not be the case) as well as how well-aligned they are. If m is a perfect measure of what the firm cares about, then g is a linear transformation of f and therefore the angle between f and g would be zero, so that $\cos\theta = 1$. If m is completely uninformative about what the firm cares about, then f and g are orthogonal, and therefore $\cos\theta = 0$. As a result, this model is often referred to as the **"cosine of theta model."**

It can be useful to view this problem geometrically. Since formal contracts allow for unrestricted lump-sum transfers between the Principal and the Agent, the Principal would optimally like efforts to be chosen in such a way that they maximize total surplus:

$$\max_e p\left(f_1 e_1 + f_2 e_2\right) - \frac{c}{2}\left(e_1^2 + e_2^2\right),$$

which has the same solution as

$$\max_e -\left(e_1 - \frac{p}{c} f_1\right)^2 - \left(e_2 - \frac{p}{c} f_2\right)^2.$$

That is, the Principal would like to choose an effort vector that is collinear with the vector f:

$$\left(e_1^{FB}, e_2^{FB}\right) = \frac{p}{c} \cdot \left(f_1, f_2\right).$$

This effort vector would coincide with the first-best effort vector, since it maximizes total surplus, and the players have quasilinear preferences.

Since contracts can only depend on m and not directly on y, the Principal has only limited control over the actions that the Agent chooses. That is, given a contract specifying incentive slope b, the Agent chooses $e_1^*(b) = \frac{b}{c} g_1$ and $e_2^*(b) = \frac{b}{c} g_2$. Therefore, the Principal

can only indirectly "choose" an effort vector that is collinear with the vector g:

$$\left(e_1^*(b), e_2^*(b)\right) = \frac{b}{c} \cdot (g_1, g_2).$$

The question is then: which such vector maximizes total surplus, which the Principal will extract with an ex ante lump-sum transfer? That is, which point along the $k \cdot (g_1, g_2)$ ray minimizes the mean-squared error distance to $\frac{p}{c} \cdot (f_1, f_2)$?

The following figure illustrates the first-best effort vector e^{FB} and the equilibrium effort vector e^*. The concentric rings around e^{FB} are the Principal's iso-profit curves. The rings that are closer to e^{FB} represent higher profit levels. The optimal contract induces effort vector e^*, which also coincides with the orthogonal projection of e^{FB} onto the ray $k \cdot (g_1, g_2)$.

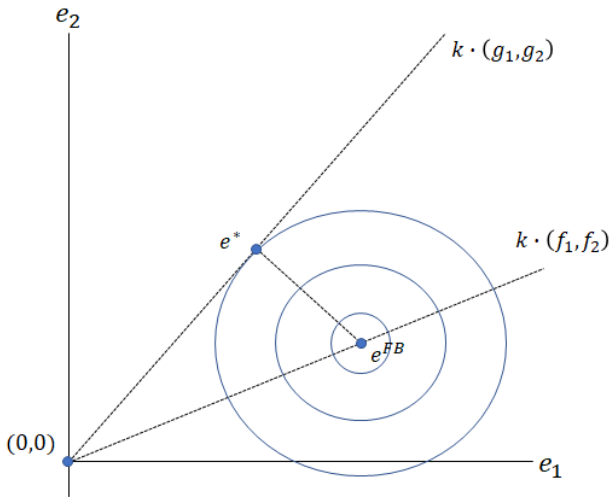

Figure 4: Optimal Effort Vector

This is a more explicit "incomplete contracts" model of motivation. That is, we are explicitly restricting the set of contracts that the Principal can offer the Agent in a way that directly determines a subset of the effort space that the Principal can induce the Agent to choose among. And it is founded not on the idea that certain measures (in particular, y) are

unobservable, but rather that they simply cannot be contracted upon.

One observation that is immediate is that it may sometimes be optimal to offer incentive contracts that provide no incentives for the Agent to choose positive effort levels (i.e., $b^* = 0$). This was essentially never the case in the model in which the Agent chose only a one-dimensional effort level, yet we often see that many employees are on contracts that look like they offer no performance-based payments. As this model highlights, this may be optimal precisely when the set of available performance measures are quite bad. As an example, suppose

$$\Pr[y = 1 | e] = \alpha + f_1 e_1 + f_2 e_2,$$

where $\alpha > 0$ and $f_2 < 0$, so that higher choices of e_2 reduce the probability of high output. And suppose the performance measure is again satisfies

$$\Pr[m = 1 | e] = g_1 e_1 + g_2 e_2,$$

with $g_1, g_2 > 0$.

We can think of $y = 1$ as representing whether a particular customer buys something that he does not later return, which depends on how well he was treated when he went to the store. We can think of $m = 1$ as representing whether the Agent made a sale but not whether the item was later returned. In order to increase the probability of making a sale, the Agent can exert "earnest" sales effort e_1 and "shady" sales effort e_2. Both are good for sales, but the latter increases the probability the item is returned. If the vectors f and g are sufficiently poorly aligned (i.e., if it is really easy to make sales by being shady), it may be

better for the firm to offer a contract with $b^* = 0$, as the following figure illustrates.

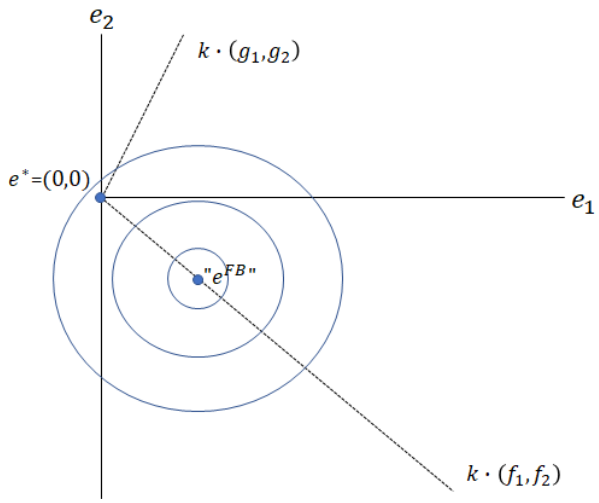

Figure 5: Sometimes Zero Effort is Optimal

This example illustrates that paying the Agent for sales can be a bad idea when what the Principal wants is *sales that are not returned*. The Kerr (1975) article is filled with many colorful examples of this problem. One such example concerns the incentives offered to the managers of orphanages. Their budgets and prestige were determined largely by the number of children they enrolled and not by whether they managed to place their children with suitable families. The claim made in the article is that the managers often denied adoption applications for inappropriate reasons: they were being rewarded for large orphanages, while the state hoped for good placements.

5.1 Limits on Activities

Firms have many instruments to help address the problems that arise in multitasking situations. We will describe two of them here in a small extension to the model. Suppose now that the Principal can put some restrictions on the types of actions the Agent is able to undertake. In particular, in addition to writing a contract on the performance measure m,

she can write a contract on the dummy variables $1_{e_1>0}$ and $1_{e_2>0}$. In other words, while she cannot directly contract upon, say, e_2, she can write a contract that heavily penalizes any positive level of it. The first question we will ask here is: when does the Principal want to exclude the Agent from engaging in task 2?

We can answer this question using the graphical intuition we just developed above. The following figure illustrates this intuition. If the Principal does not exclude task 2, then she can induce the Agent to choose any effort vector of the form $k \cdot (g_1, g_2)$. If she does exclude task 2, then she can induce the Agent to choose any effort vector of the form $k \cdot (g_1, 0)$. In the former case, the equilibrium effort vector will be e^*, which corresponds to the orthogonal projection of e^{FB} onto the ray $k \cdot (g_1, g_2)$. In the latter case, the equilibrium effort will be e^{**}, which corresponds to the orthogonal projection of e^{FB} onto the ray $k \cdot (g_1, 0)$.

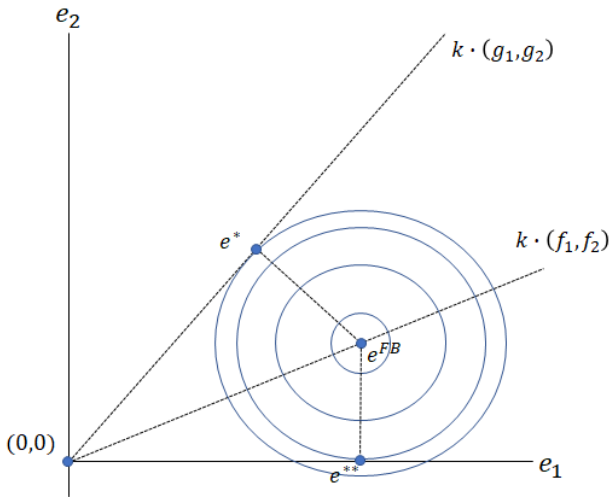

Figure 6: Excluding Task 2

This figure shows that for the particular vectors f and g it illustrates, it will be optimal for the Principal to exclude e_2: e^{**} lies on a higher iso-profit curve than e^* does. This will in fact be the case whenever the angle between vector f and g is larger than the angle between f and $(g_1, 0)$—if by excluding task 2, the performance measure m acts as if it is more closely

aligned with f, then task 2 should be excluded.

5.2 Job Design

Finally, we will briefly touch upon what is referred to as job design. Suppose f and g are such that it is not optimal to exclude either task on its own. The firm may nevertheless want to hire *two* Agents who each specialize in a single task. For the first Agent, the Principal could exclude task 2, and for the second Agent, the Principal could exclude task 1. The Principal could then offer a contract that gets the first Agent to choose $\left(e_1^{FB}, 0\right)$ and the second agent to choose $\left(0, e_2^{FB}\right)$. The following figure illustrates this possibility.

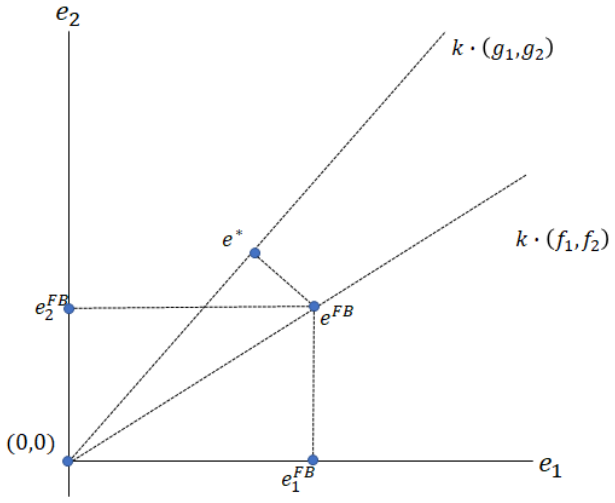

Figure 7: Job Design

When is it optimal for the firm to hire two Agents who each specialize in a single task? It depends on the Agents' opportunity cost. Total surplus under a single Agent under the optimal contract will be

$$pf \cdot e^* - \frac{c}{2} e^* \cdot e^* - \bar{u},$$

and total surplus with two specialized agents under optimal contracts will be

$$pf \cdot e^{FB} - \frac{c}{2} e^{FB} \cdot e^{FB} - 2\bar{u}.$$

Adding an additional Agent in this case is tantamount to adding an additional performance measure, which allows the Principal to choose induce any $e \in \mathbb{R}^2_+$, including the first-best effort vector. She gains from being able to do this, but to do so, she has to cover the additional Agent's opportunity cost $\bar{u}$.

6 Indistinguishable Individual Contributions

So far, we have discussed three contracting frictions that give rise to equilibrium contracts that induce effort that is not first-best. We will now discuss a final contracting friction that arises when multiple individuals contribute to a single project, and while team output is contractible, individual contributions to the team output are not. This indistinguishability gives rise to Holmström's (1982) classic "moral hazard in teams" problem.

The Model There are $I \geq 2$ risk-neutral Agents $i \in \mathcal{I} = \{1, \ldots, I\}$ who each choose efforts $e_i \in \mathcal{E}_i = \mathbb{R}_+$ at cost $c_i(e_i)$, which is increasing, convex, differentiable, and satisfies $c_i'(0) = 0$. The vector of efforts $e = (e_1, \ldots, e_I)$ determine team output $y \in \mathcal{Y} = \mathbb{R}_+$ according to a function $y(e)$ which is increasing in each e_i, concave in e, differentiable, and satisfies $\lim_{e_i \to 0} \partial y / \partial e_i = \infty$. Note that output is not stochastic, although the model can be easily extended to allow for stochastic output. Output is contractible, and each Agent i is subject to a contract $w_i \in \mathcal{W} = \{w_i : \mathcal{Y} \to \mathbb{R}\}$. We will say that the vector of contracts $w = (w_1, \ldots, w_I)$ is a **sharing rule** if

$$\sum_{i \in \mathcal{I}} w_i(y) = y$$

for each output level y. Each Agent i's preferences are given by

$$U_i(w, e) = w_i(y(e)) - c_i(e_i).$$

Each Agent i takes the contracts as given and chooses an effort level. Output is realized and each agent receives payment $w_i(y)$. The solution concept is Nash equilibrium, and we will say that w **induces** e^* if e^* is a Nash equilibrium effort profile given the vector of contracts w.

Sharing Rules and the Impossibility of First-Best Effort Since the Agents have quasilinear preferences, any Pareto-optimal outcome under a sharing rule w will involve an effort level that maximizes total surplus, so that

$$e^{FB} \in \underset{e \in \mathbb{R}^I_+}{\operatorname{argmax}} y(e) - \sum_{i \in \mathcal{I}} c_i(e_i).$$

Under our assumptions, there is a unique first-best effort vector, and it satisfies

$$\frac{\partial y\left(e^{FB}\right)}{\partial e_i} = c_i'\left(e_i^{FB}\right) \text{ for all } i \in \mathcal{I}.$$

First-best effort equates the social marginal benefit of each agent's effort level with its social marginal cost. We will denote the **first-best output level** $y\left(e^{FB}\right)$ by y^{FB}.

We will give an informal argument for why no sharing rule w induces e^{FB}, and then we will make that argument more precise. Suppose w is a sharing rule for which $w_i(y)$ is weakly concave and differentiable in y for all $i \in \mathcal{I}$. For any Nash equilibrium effort vector e^*, it must be the case that

$$w_i'(y) \cdot \frac{\partial y(e^*)}{\partial e_i} = c_i'(e_i^*) \text{ for all } i \in \mathcal{I}.$$

In order for e^* to be equal to e^{FB}, it has to be the case that these equilibrium conditions coincide with the Pareto-optimality conditions. This is only possible if $w_i'(y) = 1$ for all i,

but because w is a sharing rule, we must have that

$$\sum_{i \in \mathcal{I}} w_i'(y) = 1 \text{ for all } y.$$

Equilibrium effort e^* therefore cannot be first-best. This argument highlights the idea that getting each Agent to choose first-best effort requires that he be given the entire social marginal benefit of his effort, but it is not possible (at least under a sharing rule) for *all* the Agents simultaneously to receive the entire social marginal benefit of their efforts.

This argument is not a full argument for the impossibility of attaining first-best effort under sharing rules because it does not rule out the possibility of non-differentiable sharing rules inducing first-best effort. It turns out that there is no sharing rule, even a non-differentiable one, that induces first-best effort.

Theorem 3 (Moral Hazard in Teams). If w is a sharing rule, w does not induce e^{FB}.

Proof. This proof is due to Stole (2001). Take an arbitrary sharing rule w, and suppose e^* is an equilibrium effort profile under w. For any $i, j \in \mathcal{I}$, define $e_j(e_i)$ by the relation $y\left(e^*_{-j}, e_j(e_i)\right) = y\left(e^*_{-i}, e_i\right)$. Since y is continuous and increasing, a unique value of $e_j(e_i)$ exists for e_i sufficiently close to e_i^*. Take such an e_i. For e^* to be a Nash equilibrium, it must be the case that

$$w_j(y(e^*)) - c_j\left(e_j^*\right) \geq w_j\left(y\left(e^*_{-j}, e_j(e_i)\right)\right) - c_j(e_j(e_j)),$$

since this inequality has to hold for all $e_j \neq e_j^*$. Rewriting this inequality, and summing up over $j \in \mathcal{I}$, we have

$$\sum_{j \in \mathcal{I}} \left(w_j(y(e^*)) - w_j\left(y\left(e^*_{-i}, e_i\right)\right)\right) \geq \sum_{j \in \mathcal{I}} \left(c_j\left(e_j^*\right) - c_j(e_j(e_i))\right).$$

Since w is a sharing rule, the left-hand side of this expression is just $y(e^*) - y\left(e^*_{-i}, e_i\right)$, so

this inequality can be written

$$y(e^*) - y\left(e^*_{-i}, e_i\right) \geq \sum_{j \in \mathcal{I}} c_j\left(e^*_j\right) - c_j\left(e_j\left(e_i\right)\right).$$

Since this must hold for all e_i close to e^*_i, we can divide by $e^*_i - e_i$ and take the limit as $e_i \to e^*_i$ to obtain

$$\frac{\partial y(e^*)}{\partial e_i} \geq \sum_{j \in \mathcal{J}} c'_j\left(e^*_j\right) \frac{\partial y(e^*)/\partial e_i}{\partial y(e^*)/\partial e_j}.$$

Now suppose that $e^* = e^{FB}$. Then $c'_j\left(e^*_j\right) = \partial y(e^*)/\partial e_j$, so this inequality becomes

$$\frac{\partial y(e^*)}{\partial e_i} \geq I \frac{\partial y(e^*)}{\partial e_i},$$

which is a contradiction because y is increasing in e_i.■

Joint Punishments and Budget Breakers Under a sharing rule, first-best effort cannot be implemented because in order to deter an Agent from choosing some $e_i < e^{FB}_i$, it is necessary to punish him. But because contracts can only be written on team output, the only way to deter each agent from choosing $e_i < e^{FB}_i$ is to simultaneously punish *all* the Agents when output is less than $y\left(e^{FB}\right)$. But punishing all the Agents simultaneously requires that they throw output away, which is impossible under a sharing rule. It turns out, though, that if we allow for contracts w that allow for **money burning**, in the sense that it allows for

$$\sum_{i \in \mathcal{I}} w_i(y) < y$$

for some output levels $y \in \mathcal{Y}$, first-best effort can in fact be implemented, and it can be implemented with a contract that does not actually burn money in equilibrium.

Proposition 1. There exist a vector of contracts w that induces e^{FB} for which $\sum_{i \in \mathcal{I}} w_i\left(y^{FB}\right) = y^{FB}$.

Proof. For all i, set $w_i(y) = 0$ for all $y \neq y^{FB}$, and let $w_i\left(y^{FB}\right) > c_i\left(e^{FB}_i\right)$ for all i so that

$\sum_{i\in\mathcal{I}} w_i\left(y^{FB}\right) = y^{FB}$. Such a vector of contracts is feasible, because $y^{FB} > \sum_{i\in\mathcal{I}} c_i\left(e_i^{FB}\right)$. Finally, under w, e^{FB} is a Nash equilibrium effort profile because if all other Agents choose e_{-i}^{FB}, then if Agent i chooses $e_i \neq e_i^{FB}$, he receives $-c_i\left(e_i\right)$, if he chooses $e_i = e_i^{FB}$, he receives $w_i\left(y^{FB}\right) - c_i\left(e_i^{FB}\right) > 0$.■

Proposition 1 shows that in order to induce first-best effort, the Agents have to subject themselves to costly joint punishments in the event that one of them deviates and chooses $e_i \neq e_i^{FB}$. A concern with such contracts is that in the event that the Agents are required by the contract to burn money, they could all be made better off by renegotiating their contract and not burning money. If we insist, therefore, that w is *renegotiation-proof*, then w must be a sharing rule and therefore cannot induce e^{FB}.

This is no longer the case if we introduce an additional party, which we will call a Principal, who does not take any actions that affect output. In particular, if we denote the Principal as Agent 0, then the following sharing rule induces e^{FB}:

$$\begin{aligned} w_i\left(y\right) &= y - k \text{ for all } i = 1,\ldots,I \\ w_0\left(y\right) &= Ik - \left(I-1\right)y, \end{aligned}$$

where k satisfies

$$k = \frac{I-1}{I} y^{FB}.$$

This vector of contracts is a sharing rule, since for all $y \in \mathcal{Y}$,

$$\sum_{i=0}^{I} w_i\left(y\right) = Iy - \left(I-1\right)y = y.$$

This vector of contracts induces e^{FB} because it satisfies $\partial w_i\left(y^{FB}\right)/\partial e_i = 1$ for all $i = 1,\ldots,I$, and if we imagine the Principal having an outside option of 0, this choice of k ensures that in equilibrium, she will in fact receive 0. In this case, the Principal's role is to serve as a **budget breaker**. Her presence allows the Agents to "break the margins budget,"

allowing for $\sum_{i=1}^{I} w_i'(y) = I > 1$, while still allowing for renegotiation-proof contracts.

Under these contracts, the Principal essentially "sells the firm" to *each* agent for an amount k. Then, since each Agent earns the firm's entire output at the margin, each Agent's interests are aligned with society's interest. One limitation of this approach is that while each Agent earns the entire marginal benefit of his efforts, the Principal *loses* $I-1$ times the marginal benefit of each Agent's efforts. The Principal has strong incentives to collude with one of the Agents—while the players are jointly better off if Agent i chooses e_i^{FB} than any $e_i < e_i^{FB}$, Agent i and the Principal together are jointly better off if Agent i chose $e_i = 0$.

7 Introduction to the Theory of the Firm

The central question in this part of the literature goes back to Ronald Coase (1937): if markets are so great at coordinating productive activity, why is productive activity carried out within firms rather than by self-employed individuals who transact on a spot market? And indeed it is, as Herbert Simon (1991) vividly illustrated:

> A mythical visitor from Mars... approaches Earth from space, equipped with a telescope that reveals social structures. The firms reveal themselves, say, as solid green areas with faint interior contours marking out divisions and departments. Market transactions show as red lines connecting firms, forming a network in the spaces between them. Within firms (and perhaps even between them) the approaching visitor also sees pale blue lines, the lines of authority connecting bosses with various levels of workers... No matter whether our visitor approached the United States or the Soviet Union, urban China or the European Community, the greater part of the space below it would be within the green areas, for almost all inhabitants would be employees, hence inside the firm boundaries. Organizations would be the dominant feature of the landscape. A message sent back home, describing the scene, would speak of "large green areas interconnected by

> red lines." It would not likely speak of "a network of red lines connecting green spots." ...When our visitor came to know that the green masses were organizations and the red lines connecting them were market transactions, it might be surprised to hear the structure called a market economy. "Wouldn't 'organizational economy' be the more appropriate term?" it might ask. (pp. 27-28)

It is obviously difficult to put actual numbers on the relative importance of trade within and between firms, since, I would venture to say, most transactions within firms are not recorded. From dropping by a colleague's office to ask for help finding a reference, transferring a shaped piece of glass down the assembly line for installation into a mirror, getting an order of fries from the fry cook to deliver to the customer, most economic transactions are difficult even to define as such, let alone track. But we do have some numbers. The first sentence of Antràs (2003) provides a lower bound: "Roughly one-third of world trade is intrafirm trade."

Of course, it could conceivably be the case that boundaries don't really matter—that the nature of a particular transaction and the overall volume of transactions is the same whether boundaries are in place or not. And indeed, this would exactly be the case if there were no costs of carrying out transactions: Coase's (1960) eponymous theorem suggests, roughly, that in such a situation, outcomes would be the same no matter how transactions were organized. But clearly this is not the case—in 1997, to pick a random year, the volume of corporate mergers and acquisitions was $1.7 trillion dollars (Holmström and Roberts, 1998). It is implausible that this would be the case if boundaries were irrelevant, as even the associated legal fees have to ring up in the billions of dollars.

And so, in a sense, the premise of the Coase Theorem's contrapositive is clearly true. Therefore, there must be transaction costs. And understanding the nature of these transaction costs will hopefully shed some light on the patterns we see. Moreover, as D.H. Robertson vividly illustrated, there are indeed patterns to what we see. Firms are "islands of conscious power in this ocean of unconscious co-operation like lumps of butter coagulating in a pail of buttermilk." So the question becomes: what transaction costs are important, and how are

they important? How, in a sense, can they help make sense out of the pattern of butter and buttermilk?

The field was basically dormant for the next forty years until the early 1970s, largely because "transaction costs" came to represent essentially "a name for the residual"—any pattern in the data could trivially be attributed to some story about transaction costs. The empirical content of the theory was therefore essentially zero.

Williamson put structure on the theory by identifying specific factors that composed these transaction costs. And importantly, the specific factors he identified had implications about economic objects that at least could, in principle, be contained in a data set. Therefore his causal claims could be, and were, tested. (As a conceptual matter, it is important to note that even if Williamson's causal claims were refuted, this would not invalidate the underlying claim that "transaction costs are important," since as discussed earlier, this more general claim is essentially untestable, because it is impossible to measure, or even conceive of, *all* transaction costs associated with *all* different forms of organization.)

The gist of Williamson's Transaction Cost Economics (TCE) theory is that when contracts are incomplete, and parties have disagreements, they may waste resources "haggling" over the appropriate course of action if they transact in a market, whereas if they transact within a firm, these disagreements can be settled by authority or by "fiat." Integration is therefore more appealing than the market when haggling costs are higher, which is the case in situations in which contracts are relatively more incomplete and parties disagree more.

As a classic example (due to Joskow (1985)), think about the relationship between an underground coal mine and a coal fired power plant. It is much more efficient for the power plant to be located close to the coal mine, but the power plant is unlikely to do so absent contractual safeguards. Maybe the parties then end up signing a 20-year contract detailing the type of coal that the mine will send to the power plant and at what price. But after a few years, there may be a regulatory change preventing the use of that particular type of coal. Since such a change is difficult to foresee, the parties may not have specified what to do in this

event, and they will have to renegotiate the contract, and these renegotiations may be costly. One way to avoid the problems associated with such renegotiations is vertical integration: the electricity company could buy the coal mine instead of entering into a contract with it. And in the event of a regulatory change, the electricity company just orders the coal mine to produce a different type of coal.

But there was a sense in which TCE theory (and the related work by Klein, Crawford, and Alchian (1978)) was silent on many foundational questions. After all, why does moving the transaction from the market into the firm imply that parties no longer haggle—that is, what is integration? Further, if settling transactions by fiat is more efficient than by haggling, why aren't all transactions carried out within a single firm? Williamson's and others' response was that there are bureaucratic costs ("accounting contrivances," "weakened incentives," and others) associated with putting more transactions within the firm. But surely those costs are also higher when contracts are more incomplete and when there is more disagreement between parties. Put differently, Williamson identified particular costs associated with transacting in the market and other costs associated with transacting within the firm and made assertions about the rates at which these costs vary with the underlying environment. The resulting empirical implications were consistent with evidence, but the theory still lacked convincing foundations, because it treated these latter costs as essentially exogenous and orthogonal.

The Property Rights Theory (PRT), initiated by Grossman and Hart (1986) and expanded upon in Hart and Moore (1990), proposed a theory which (a) explicitly answered the question of "what is integration?" and (b) treated the costs and benefits of integration symmetrically. Related to the first point is an observation by Alchian and Demsetz that

> It is common to see the firm characterized by the power to settle issues by fiat, by authority, or by disciplinary action superior to that available in the conventional market. This is delusion. The firm does not own all its inputs. It has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. I can "punish" you

> only by withholding future business or by seeking redress in the courts for any failure to honor our exchange agreement. This is exactly all that any employer can do. He can fire or sue, just as I can fire my grocer by stopping purchases from him or sue him for delivering faulty products. (1972, p. 777)

What, then, is the difference between me "telling my grocer what to do" and me "telling my employee what to do?" In either case, refusal would potentially cause the relationship to break down. The key difference, according to Grossman and Hart's theory, is in what happens after the relationship breaks down. If I stop buying goods from my grocer, I no longer have access to his store and all its associated benefits. He simply loses access to a particular customer. If I stop employing a worker, on the other hand, the worker loses access to all the assets associated with my firm. I simply lose access to that particular worker.

Grossman and Hart's (1986) key insight is that property rights determine who can do what in the event that a relationship breaks down—property rights determine what they refer to as the residual rights of control. And allocating these property rights to one party or another may change their incentives to take actions that affect the value of this particular relationship. This logic leads to what is often interpreted as Grossman and Hart's main result: property rights (which define whether a particular transaction is carried out "within" a firm or "between" firms) should be allocated to whichever party is responsible for making more important investments in the relationship.

From a theoretical foundations perspective, Grossman and Hart (1986) was a huge step forward—the theory treats the costs of integration and the costs of non-integration symmetrically and systematically analyzes how different factors drive these two costs in a single unified framework. From a conceptual perspective, however, all the action in the theory is related to how organization affects parties' incentives to make relationship-specific investments. As we will see, the theory assumes that conditional on relationship-specific investments, transactions are always carried out efficiently. A manager never wastes time and resources arguing with an employee. An employee never wastes time and resources trying to convince the boss

to let him do a different, more desirable task.

Even the Property Rights Theory does not stand on fully firm theoretical grounds, since the theory considers only a limited set of institutions the players can put in place to manage their relationship. That is, PRT focuses only on the allocation of control, ignoring the possibility that individuals may write contracts or put in place other types of mechanisms that could potentially do better. In particular, it rules out revelation mechanisms that, in principle, should induce first-best investment. We will briefly talk about this after we talk about the model.

8 Property Rights Theory

Essentially the main result of TCE is the observation that when haggling costs are high under non-integration, then integration is optimal. This result is unsatisfying in at least two senses. First, TCE does not tell us what exactly is the mechanism through which haggling costs are reduced under integration, and second, it does not tell us what the associated costs of integration are, and it therefore does not tell us when we would expect such costs to be high. In principle, in environments in which haggling costs are high under non-integration, then the within-firm equivalent of haggling costs should also be high.

Grossman and Hart (1986) and Hart and Moore (1990) set aside the "make or buy" question and instead begin with the more fundamental question, "What is a firm?" In some sense, nothing short of an answer to *this* question will consistently provide an answer to the questions that TCE leaves unanswered. Framing the question slightly differently, what do I get if I buy a firm from someone else? The answer is typically that I become the owner of the firm's non-human assets.

Why, though, does it matter who owns non-human assets? If contracts are complete, it does not matter. The parties to a transaction will, ex ante, specify a detailed action plan. One such action plan will be optimal. That action plan will be optimal regardless of who owns

the assets that support the transaction, and it will be feasible regardless of who owns the assets. If contracts are incomplete, however, not all contingencies will be specified. The key insight of the PRT is that ownership endows the asset's owner with the right to decide what to do with the assets in these contingencies. That is, ownership confers **residual control rights**. When unprogrammed adaptations become necessary, the party with residual control rights has **power** in the relationship and is protected from expropriation by the other party. That is, control over non-human assets leads to control over human assets, since they provide leverage over the person who lacks the assets. Since she cannot be expropriated, she therefore has incentives to make investments that are specific to the relationship.

Firm boundaries are tantamount to asset ownership, so detailing the costs and benefits of different ownership arrangements provides a complete account of the costs and benefits of different firm-boundary arrangements. Asset ownership, and therefore firm boundaries, determine who possesses power in a relationship, and power determines investment incentives. Under integration, I have all the residual control rights over non-human assets and therefore possess strong investment incentives. Non-integration splits apart residual control rights, and therefore provides me with weaker investment incentives and you with stronger investment incentives. If I own an asset, you do not. Power is scarce and therefore should be allocated optimally.

Methodologically, PRT makes significant advances over the preceding theory. PRT's conceptual exercise is to hold technology, preferences, information, and the legal environment constant across prospective governance structures and ask, for a given transaction with given characteristics, whether the transaction is best carried out within a firm or between firms. That is, prior theories associated "make" with some vector $(\alpha_1, \alpha_2, \dots)$ of characteristics and "buy" with some other vector $(\beta_1, \beta_2, \dots)$ of characteristics. "Make" is preferred to "buy" if the vector $(\alpha_1, \alpha_2, \dots)$ is preferred to the vector $(\beta_1, \beta_2, \dots)$. In contrast, PRT focuses on a single aspect: α_1 versus β_1. Further differences may arise between "make" and "buy," but to the extent that they are also choice variables, they will arise optimally rather than

by assumption.

The Model There is a risk-neutral upstream manager U, a risk-neutral downstream manager D, and two assets a_1 and a_2. Managers U and D make investments $e_U \in \mathcal{E}_U = \mathbb{R}_+$ and $e_D \in \mathcal{E}_D = \mathbb{R}_+$ at private cost $c_U(e_U)$ and $c_D(e_D)$. These investments determine the value that each manager receives if trade occurs, $V_U(e_U, e_D)$ and $V_D(e_U, e_D)$. There is a state of the world, $s \in \mathcal{S} = \mathcal{S}_C \cup \mathcal{S}_{NC}$, with $\mathcal{S}_C \cap \mathcal{S}_{NC} = \emptyset$ and $\Pr[s \in \mathcal{S}_{NC}] = \mu$. In state s, the identity of the ideal good to be traded is s—if the managers trade good s, they receive $V_U(e_U, e_D)$ and $V_D(e_U, e_D)$. If the managers trade good $s' \neq s$, they both receive $-\infty$. The managers choose an asset allocation, denoted by g, from a set $\mathcal{G} = \{UI, DI, NI, RNI\}$. Under $g = UI$, U owns both assets. Under $g = DI$, D owns both assets. Under $g = NI$, U owns asset a_1 and D owns asset a_2. Under $g = RNI$, D owns asset a_1, and U owns asset a_2. In addition to determining an asset allocation, manager U also offers an incomplete contract $w \in \mathcal{W} = \{w : S_C \to \mathbb{R}\}$ to D. The contract specifies a transfer $w(s)$ to be paid from D to U if they trade good $s \in \mathcal{S}_C$. If the players want to trade a good $s \in \mathcal{S}_{NC}$, they do so in the following way. With probability $\frac{1}{2}$, U makes a take-it-or-leave-it offer $w_U(s)$ to D, specifying trade and a price. With probability $\frac{1}{2}$, D makes a take-it-or-leave-it offer $w_D(s)$ to U specifying trade and a price. If trade does not occur, then manager U receives payoff $v_U(e_U, e_D; g)$ and manager D receives payoff $v_D(e_U, e_D; g)$, which depends on the asset allocation.

Timing There are five periods:

1. U offers D an asset allocation $g \in \mathcal{G}$ and a contract $w \in \mathcal{W}$. Both g and w are commonly observed.

2. U and D simultaneously choose investment levels e_U and e_D at private cost $c(e_U)$ and $c(e_D)$. These investment levels are commonly observed by e_U and e_D.

3. The state of the world, $s \in \mathcal{S}$ is realized.

4. If $s \in \mathcal{S}_C$, D buys good s at price specified by w. If $s \in \mathcal{S}_{NC}$, U and D engage in 50-50 take-it-or-leave-it bargaining.

5. Payoffs are realized.

Equilibrium A **subgame-perfect equilibrium** is an asset allocation g^*, a contract w^*, investment strategies $e_U^* : \mathcal{G} \times \mathcal{W} \to \mathbb{R}_+$ and $e_D^* : \mathcal{G} \times \mathcal{W} \to \mathbb{R}_+$, and a pair of offer rules $w_U^* : \mathcal{E}_D \times \mathcal{E}_U \times \mathcal{S}_{NC} \to \mathbb{R}$ and $w_D^* : \mathcal{E}_D \times \mathcal{E}_U \times \mathcal{S}_{NC} \to \mathbb{R}$ such that given $e_U^*(g^*, w^*)$ and $e_D^*(g^*, w^*)$, the managers optimally make offers $w_U^*(e_U^*, e_D^*)$ and $w_D^*(e_U^*, e_D^*)$ in states $s \in \mathcal{S}_{NC}$; given g^* and w^*, managers optimally choose $e_U^*(g^*, w^*)$ and $e_D^*(g^*, w^*)$; and U optimally offers asset allocation g^* and contract w^*.

Assumptions We will assume $c_U(e_U) = \frac{1}{2}e_U^2$ and $c_D(e_D) = \frac{1}{2}e_D^2$. We will also assume that $\mu = 1$, so that the probability that an ex ante specifiable good is optimal to trade ex post is zero. Let

$$\begin{aligned} V_U(e_U, e_D) &= f_{UU}e_U + f_{UD}e_D \\ V_D(e_U, e_D) &= f_{DU}e_U + f_{DD}e_D \\ v_U(e_U, e_D; g) &= h_{UU}^g e_U + h_{UD}^g e_D \\ v_D(e_U, e_D; g) &= h_{DU}^g e_U + h_{DD}^g e_D, \end{aligned}$$

and define $F_U = f_{UU} + f_{DU}$ and $F_D = f_{UD} + f_{DD}$. Finally, outside options are more sensitive to one's own investments the more assets one owns:

$$\begin{aligned} h_{UU}^{UI} &\geq h_{UU}^{NI} \geq h_{UU}^{DI}, h_{UU}^{UI} \geq h_{UU}^{RNI} \geq h_{UU}^{DI} \\ h_{DD}^{DI} &\geq h_{DD}^{NI} \geq h_{DD}^{UI}, h_{DD}^{DI} \geq h_{DD}^{RNI} \geq h_{DD}^{UI}. \end{aligned}$$

The Program We solve backwards. For all $s \in \mathcal{S}_{NC}$, with probability $\frac{1}{2}$, U will offer price $w_U(e_U, e_D)$. D will accept this offer as long as $V_D(e_U, e_D) - w_U(e_U, e_D) \geq v_D(e_U, e_D; g)$.

U's offer will ensure that this holds with equality (or else U could increase w_U a bit and increase his profits while still having his offer accepted), so that $\pi_U = V_U + V_D - v_D$ and $\pi_D = v_D$.

Similarly, with probability $\frac{1}{2}$, D will offer price $w_D(e_U, e_D)$. U will accept this offer as long as $V_U(e_U, e_D) + w_D(e_U, e_D) \geq v_U(e_U, e_D; g)$. D's offer will ensure that this holds with equality (or else D could decrease w_D a bit and increase her profits while still having her offer accepted), so that $\pi_U = v_U$ and $\pi_D = V_U + V_D - v_U$.

In period 2, manager U will conjecture e_D and solve

$$\max_{\hat{e}_U} \frac{1}{2}\left(V_U(\hat{e}_U, e_D) + V_D(\hat{e}_U, e_D) - v_D(\hat{e}_U, e_D; g)\right) + \frac{1}{2} v_U(\hat{e}_U, e_D; g) - c(\hat{e}_U)$$

and manager D will conjecture e_U and solve

$$\max_{\hat{e}_D} \frac{1}{2} v_D(e_U, \hat{e}_D; g) + \frac{1}{2}\left(V_U(e_U, \hat{e}_D) + V_D(e_U, \hat{e}_D) - v_U(e_U, \hat{e}_D; g)\right) - c(\hat{e}_D).$$

Given our functional form assumptions, these are well-behaved objective functions, and in each one, there are no interactions between the managers' investment levels, so each manager has a dominant strategy. We can therefore solve for the associated equilibrium investment levels by taking first-order conditions:

$$\begin{aligned} e_U^{*g} &= \frac{1}{2} F_U + \frac{1}{2}\left(h_{UU}^g - h_{DU}^g\right) \\ e_D^{*g} &= \frac{1}{2} F_D + \frac{1}{2}\left(h_{DD}^g - h_{UD}^g\right) \end{aligned}$$

Each manager's incentives to invest are derived from two sources: (1) the marginal impact of investment on total surplus and (2) the marginal impact of investment on the "threat-point differential." The latter point is worth expanding on. If U increases his investment, his outside option goes up by h_{UU}^g, which increases the price that D will have to offer him when she makes her take-it-or-leave-it offer, which increases U's ex-post payoff if $h_{UU}^g > 0$.

Further, D's outside option goes up by h^g_{DU}, which increases the price that U has to offer D when he makes his take-it-or-leave-it-offer, which decreases U's ex-post payoff if $h^g_{DU} > 0$.

Contrasting these equilibrium conditions with the conditions satisfied by first-best effort levels is informative. First-best effort levels satisfy $e^{FB}_U = F_U$ and $e^{FB}_D = F_D$. In contrast, when parties can use renegotiation opportunities to their own advantage, (1) they have weaker incentives to make value-increasing investments that are specific to the relationship, and (2) they may have excessive incentives to make strategic investments in their own outside options or in reducing the outside option of the other party.

Ex ante, players' equilibrium payoffs are:

$$\begin{aligned}
\Pi^{*g}_U &= \frac{1}{2}\left(F_U e^{*g}_U + F_D e^{*g}_D\right) + \frac{1}{2}\left(\left(h^g_{UU} - h^g_{DU}\right) e^{*g}_U + \left(h^g_{UD} - h^g_{DD}\right) e^{*g}_D\right) - \frac{1}{2}\left(e^{*g}_U\right)^2 \\
\Pi^{*g}_D &= \frac{1}{2}\left(F_U e^{*g}_U + F_D e^{*g}_D\right) + \frac{1}{2}\left(\left(h^g_{DU} - h^g_{UU}\right) e^{*g}_U + \left(h^g_{DD} - h^g_{UD}\right) e^{*g}_D\right) - \frac{1}{2}\left(e^{*g}_D\right)^2.
\end{aligned}$$

If we let $\theta = \left(f_{UU}, f_{UD}, f_{DU}, f_{DD}, \{h^g_{UU}, h^g_{UD}, h^g_{DU}, h^g_{DD}\}_{g\in G}\right)$ denote the parameters of the model, the Coasian objective for **governance structure** g is:

$$W^g(\theta) = \Pi^{*g}_U + \Pi^{*g}_D = F_U e^{*g}_U + F_D e^{*g}_D - \frac{1}{2}\left(e^{*g}_U\right)^2 - \frac{1}{2}\left(e^{*g}_D\right)^2.$$

The **Coasian Problem** that describes the optimal governance structure is then:

$$W^*(\theta) = \max_{g\in\mathcal{G}} W^g(\theta).$$

At this level of generality, the model is too rich to provide straightforward insights. In order to make progress, we will introduce the following definitions. If $f_{ij} = h^g_{ij} = 0$ for $i \neq j$, we say that investments are **self-investments**. If $f_{ii} = h^g_{ii} = 0$, we say that investments are **cross-investments**. When investments are self-investments, the following definitions are useful. Assets A_1 and A_2 are **independent** if $h^{UI}_{UU} = h^{NI}_{UU} = h^{RNI}_{UU}$ and $h^{DI}_{DD} = h^{NI}_{DD} = h^{RNI}_{DD}$ (i.e., if owning the second asset does not increase one's marginal

incentives to invest beyond the incentives provided by owning a single asset). Assets A_1 and A_2 are **strictly complementary** if either $h^{NI}_{UU} = h^{RNI}_{UU} = h^{DI}_{UU}$ or $h^{NI}_{DD} = h^{RNI}_{DD} = h^{UI}_{DD}$ (i.e., if for one player, owning one asset provides the same incentives to invest as owning no assets). U's **human capital is essential** if $h^{DI}_{DD} = h^{UI}_{DD}$, and D's human capital is essential if $h^{UI}_{UU} = h^{DI}_{UU}$.

With these definitions in hand, we can get a sense for what features of the model drive the optimal governance-structure choice.

Theorem 4. If A_1 and A_2 are independent, then NI or RNI is optimal. If A_1 and A_2 are strictly complementary, then DI or UI is optimal. If U's human capital is essential, UI is optimal. If D's human capital is essential, DI is optimal. If both U's and D's human capital is essential, all governance structures are equally good.

These results are straightforward to prove. If A_1 and A_2 are independent, then there is no additional benefit of allocating a second asset to a single party. Dividing up the assets therefore strengthens one party's investment incentives without affecting the other's. If A_1 and A_2 are strictly complementary, then relative to integration, dividing up the assets necessarily weakens one party's investment incentives without increasing the other's, so one form of integration clearly dominates. If U's human capital is essential, then D's investment incentives are independent of which assets he owns, so UI is at least weakly optimal.

The more general results of this framework are that (a) allocating an asset to an individual strengthens that party's incentives to invest, since it increases his bargaining position when unprogrammed adaptation is required, (b) allocating an asset to one individual has an opportunity cost, since it means that it cannot be allocated to the other party. Since we have assumed that investment is always socially valuable, this implies that assets should always be allocated to exactly one party (if joint ownership means that both parties have a veto right). Further, allocating an asset to a particular party is more desirable the more important that party's investment is for joint welfare and the more sensitive his/her investment is to asset ownership. Finally, assets should be co-owned when there are complementarities between

them.

While the actual results of the PRT model are sensible and intuitive, there are many limitations of the analysis. First, as Holmström (1999) points out, "The problem is that the theory, as presented, really is a theory about asset ownership by individuals rather than by firms, at least if one interprets it literally. Assets are like bargaining chips in an entirely autocratic market... Individual ownership of assets does not offer a theory of organizational identities unless one associates individuals with firms." Holmström concludes that, "... the boundary question is in my view fundamentally about the distribution of activities: What do firms do rather than what do they own? Understanding asset configurations should not become an end in itself, but rather a means toward understanding activity configurations." That is, by taking payoff functions V_U and V_D as exogenous, the theory is abstracting from what Holmström views as the key issue of what a firm really is.

Second, after assets have been allocated and investments made, adaptation is made efficiently. The managers always reach an ex post efficient arrangement in an efficient manner, and all inefficiencies arise ex ante through inadequate incentives to make relationship-specific investments. Williamson (2000) argues that, "The most consequential difference between the TCE and [PRT] setups is that the former holds that maladaptation in the contract execution interval is the principal source of inefficiency, whereas [PRT] vaporize ex post maladaptation by their assumptions of common knowledge and ex post bargaining." That is, Williamson believes that ex post inefficiencies are the primary sources of inefficiencies that have to be managed by adjusting firm boundaries, while the PRT model focuses solely on ex ante inefficiencies. The two approaches are obviously complementary, but there is an entire dimension of the problem that is being left untouched under this approach.

Finally, in the Coasian Problem of the PRT model, the parties are unable to write formal contracts (in the above version of the model, this is true only when $\mu = 1$) and therefore the only instrument they have to motivate relationship-specific investments is the allocation of assets. The implicit assumption underlying the focus on asset ownership is that

the characteristics defining what should be traded in which state of the world are difficult to write into a formal contract in a way that a third-party enforcer can unambiguously enforce. State-contingent trade is therefore unverifiable, so contracts written directly or indirectly on relationship-specific investments are infeasible. However, PRT assumes that relationship-specific investments, and therefore the value of different ex post trades, are commonly observable to U and D. Further, U and D can correctly anticipate the payoff consequences of different asset allocations and different levels of investment. Under the assumptions that relationship-specific investments are commonly observable and that players can foresee the payoff consequences of their actions, Maskin and Tirole (1999) shows that the players should always be able to construct a mechanism in which they truthfully reveal the payoffs they would receive to a third-party enforcer. If the parties are able to write a contract on these announcements, then they should indirectly be able to write a contract on ex ante investments. This debate over the "foundations of incomplete contracting" mostly played out over the mid-to-late 1990s, but it has attracted some recent attention.

Exercise 22 (Adapted from Bolton and Dewtripont, Question 42). Consider the following vertical integration problem: there are two risk-neutral managers, each running an asset a_i, where $i = 1, 2$. Both managers make ex ante investments. Only ex post spot contracts regulating trade are feasible. Ex post trade at price P results in the following payoffs: $R(e_D) - P$ for the downstream manager D and $P - C(e_U)$ for the upstream manager U, where the e_i's denote ex ante investment levels. Investing e_U costs the upstream manager e_U, and investing e_D costs the downstream manager e_D.
If the two managers do not trade with each other, their respective payoffs are

$$r(e_D, \mathcal{A}_D) - P_m \text{ and } P_m - c(e_U, \mathcal{A}_U),$$

where P_m is a market price, and $\mathcal{A}_i$ denotes the collection of assets owned by manager i. In this problem, $\mathcal{A}_i = \emptyset$ under j-integration, $\mathcal{A}_i = \{a_1, a_2\}$ under i-integration, and $\mathcal{A}_i = \{a_i\}$ under nonintegration.
As in the Grossman-Hart-Moore setting, it is assumed that

$$R(e_D) - C(e_U) > r(e_D, \mathcal{A}_1) - c(e_2, \mathcal{A}_2)$$

for all $(e_D, e_U) \in [0, \bar{e}]^2$ and all $\mathcal{A}_i$,

$$R'(e_D) > r'(e_D, \{a_1, a_2\}) \geq r'(e_D, \{a_i\}) \geq r'(e_D, \emptyset) \geq 0,$$

and

$$-C'(e_U) > -c'(e_U, \{a_1, a_2\}) \geq -c'(e_U, \{a_i\}) \geq -c'(e_U, \emptyset) \geq 0.$$

(a) Characterize the first-best allocation of assets and investment levels.

(b) Assuming that the managers split the ex post gains from trade in half, identify conditions on $r'(e_D, \mathcal{A}_i)$ and $c'(e_D, \mathcal{A}_i)$ such that nonintegration is optimal.

Exercise 23. Suppose a downstream buyer D and an upstream seller U meet at date $t = 1$ and trade a widget at date $t = 3$. The value of the widget to the buyer is e_D, and the seller's cost of production is 0. Here, e_D represents an (unverifiable) investment made by the buyer at date $t = 2$. The cost of investment, which is borne entirely by the buyer, is $ce_D^2/2$. No long-term contracts can be written, and there is no discounting.

(a) What is the first-best investment level e_D^{FB}?

(b) Suppose there is a single asset. If the buyer owns it, he has an outside option of λe_D, where $\lambda \in (0, 1)$. If the seller owns it, she has an outside option of v, which is independent of and smaller than e_D. (Imagine that the seller can sell the asset for v in the outside market, and the minimal investment e_D is bigger than v.) Assume that the buyer and seller divide the ex post gains from trade $50 : 50$ (Nash bargaining).

Compute the buyer's investment for the case where the buyer owns the asset and for the case where the seller owns the asset.

(c) Now assume a different bargaining game at date $t = 3$. If both parties have outside options that are valued below $e_D/2$, the parties split the surplus, giving $e_D/2$ to each party. If one of the parties has an outside option that gives $r > e_D/2$, then the party gets r and the other party gets the remainder $e_D - r$. Supposing that $\lambda > 1/2$, compute the buyer's investment when the buyer owns the asset. Compare this with the outcome when the seller owns the asset, distinguishing between the situations where v is high and v is low. Note: for this part, assume that, under S-ownership, B's outside option is $\bar{w} < -v$, making it irrelevant.

Long Hint: this part is a bit complicated due to the non-standard bargaining game, but it is illustrative of how the bargaining structure affects investment incentives (and it makes Nash bargaining look very nice in comparison). This hint is meant to guide you through the problem.

- Under seller ownership, the bargaining game is such that the buyer chooses e_D to

$$\max_{e_D} \left\{ \min \left\{ e_D - v, \frac{e_D}{2} \right\} - \frac{c}{2} e_D^2 \right\}.$$

- Break it up into cases:
  - If $e_D - v < e_D/2$, then what is the buyer's optimal choice of e_D? Plug back in to check that the condition holds.
  - If $e_D - v > e_D/2$, then what is the buyer's optimal choice of e_D? Plug back in to check that the condition holds—what happens if it does not?
- Write the buyer's optimal choice of e_D as a step function with arguments v and c.

9 Foundations of Incomplete Contracts

Property rights have value when contracts are incomplete because they determine who has residual rights of control, which in turn protects that party (and its relationship-specific investments) from expropriation by its trading partners. We will now discuss some of the commonly given reasons for why contracts might be incomplete, and in particular, we will focus on whether it makes sense to apply these reasons as justifications for incomplete contracts in the Property Rights Theory.

Contracts may not be as complete as parties would like for one of three reasons. First, parties might have private information. This is the typical reason given for why, in our discussion of moral hazard models, contracts could only depend on output or a misaligned performance measure rather than directly on the agent's effort. But in such models, contracts specified in advance are likely to be just as incomplete as contracts that are filled in at a later date. We typically do not refer to such models as models of incomplete contracting models, and we reserve the term "incomplete" to refer to a contract that simply does not lay out all the future contingencies.

One often-given justification for incomplete contracts (in this more precise sense) is that it may just be costly to write a complicated state-contingent decision rule into a contract that is enforceable by a third party. This is surely important, and several authors have modeled this idea explicitly (Dye, 1985; Bajari and Tadelis, 2001; and Battigalli and Maggi, 2002) and drawn out some of its implications. Nevertheless, I will focus instead on the final reason.

The final reason often given is that parties may like to specify what to do in each state of the world in advance, but some of these states of the world are either unforeseen or indescribable by these parties. As a result, parties may leave the contract incomplete and "fill in the details" once more information has arrived. Decisions may be ex ante non-contractible but ex post contractible (and importantly for applied purposes, tractably derived by the economist as the solution to an efficient bargaining protocol), as in the Property Rights

Theory.

I will focus on the third justification, providing some of the arguments given in a sequence of papers (Maskin and Tirole, 1999; Maskin and Moore, 1999; Maskin, 2002) about why this justification alone is insufficient if parties can foresee the payoff consequences of their actions, which they must if they are to accurately assess the payoff consequences of different allocations of property rights. In particular, these papers point out that there exists auxiliary mechanisms that are capable of ensuring truthful revelation of mutually known, payoff-relevant information as part of the unique subgame-perfect equilibrium. Therefore, even though payoff-relevant information may not be directly observable by a third-party enforcer, truthful revelation via the mechanism allows for indirect verification, which implies that any outcome attainable with ex ante describable states of the world is also attainable with ex ante indescribable states of the world.

This result is troubling in its implications for the Property Rights Theory. Comparing the effectiveness of second-best institutional arrangements (e.g., property-rights allocations) under incomplete contracts is moot when a mechanism exists that is capable of achieving, in this setting, first best outcomes. Here, I will provide an example of the types of mechanisms that have been proposed in the literature, and I will point out a couple of recent criticisms of these mechanisms.

9.1 An Example of a Subgame-Perfect Implementation Mechanism

I will first sketch an elemental hold-up model, and then I will show that it can be augmented with a subgame-perfect implementation mechanism that induces first-best outcomes.

Hold-Up Problem There is a Buyer (B) and a Seller (S). S can choose an effort level $e \in \{0, 1\}$ at cost ce, which determines how much B values the good that S produces. B values this good at $v = v_L + e\,(v_H - v_L)$. There are no outside sellers who can produce this

good, and there is no external market on which the seller could sell his good if he produces it. Assume $(v_H - v_L)/2 < c < (v_H - v_L)$.

There are three periods:

1. S chooses e. e is commonly observed but unverifiable by a third party.

2. v is realized. v is commonly observed but unverifiable by a third party.

3. With probability $1/2$, B makes a take-it-or-leave-it offer to S, and with probability $1/2$, S makes a take-it-or-leave-it offer to B.

This game has a unique subgame-perfect equilibrium. At $t = 3$, if B gets to make the offer, B asks for S to sell him the good at price $p = 0$. If S gets to make the offer, S demands $p = v$ for the good. From period 1's perspective, the expected price that S will receive is $E[p] = v/2$, so S's effort-choice problem is

$$\max_{e \in \{0,1\}} \frac{1}{2} v_L + \frac{1}{2} e (v_H - v_L) - ce.$$

Since $(v_H - v_L)/2 < c$, S optimally chooses $e^* = 0$. In this model, ex ante effort incentives arise as a by-product of ex post bargaining, and as a result, the trade price may be insufficiently sensitive to S's effort choice to induce him to choose $e^* = 1$. This is the standard hold-up problem. Note that the assumption that v is commonly observed is largely important, because it simplifies the ex post bargaining problem.

Subgame-Perfect Implementation Mechanism While effort is not verifiable by a third-party court, public announcements can potentially be used in legal proceedings. Thus, the two parties can in principle write a contract that specifies trade as a function of announcements $\hat{v}$ made by B. If B always tells the truth, then his announcements can be used to set prices that induce S to choose $e = 1$. One way of doing this is to implement a mechanism that allows announcements to be challenged by S and to punish B any time he

is challenged. If S challenges only when B has told a lie, then the threat of punishment will ensure truth telling.

The crux of the implementation problem, then, is to give S the power to challenge announcements, but to prevent "he said, she said" scenarios wherein S challenges B's announcements when he has in fact told the truth. The key insight of SPI mechanisms is to combine S's challenge with a test that B will pass if and only if he in fact told the truth.

To see how these mechanisms work, and to see how they could in principle solve the hold-up problem, let us suppose the players agree ex-ante to subject themselves to the following multi-stage mechanism.

1. B and S write a contract in which trade occurs at price $p(\hat{v})$. $p(\cdot)$ is commonly observed and verifiable by a third party.

2. S chooses e. e is commonly observed but unverifiable by a third party.

3. v is realized. v is commonly observed but unverifiable by a third party.

4. B announces $\hat{v} \in \{v_L, v_H\}$. $\hat{v}$ is commonly observed and verifiable by a third party.

5. S can challenge B's announcement or not. The challenge decision is commonly observed and verifiable by a third party. If S does not challenge the announcement, trade occurs at price $p(\hat{v})$. Otherwise, play proceeds to the next stage.

6. B pays a fine F to a third-party enforcer and is presented with a counter offer in which he can purchase the good at price $\hat{p}(\hat{v}) = \hat{v} + \varepsilon$. B's decision to accept or reject the counter off is commonly observed and verifiable by a third party.

7. If B accepts the counter offer, then S receives F from the third-party enforcer. If B does not, then S also has to pay F to the third-party enforcer.

The game induced by this mechanism seems slightly complicated, but we can sketch out

the game tree in a relatively straightforward manner.

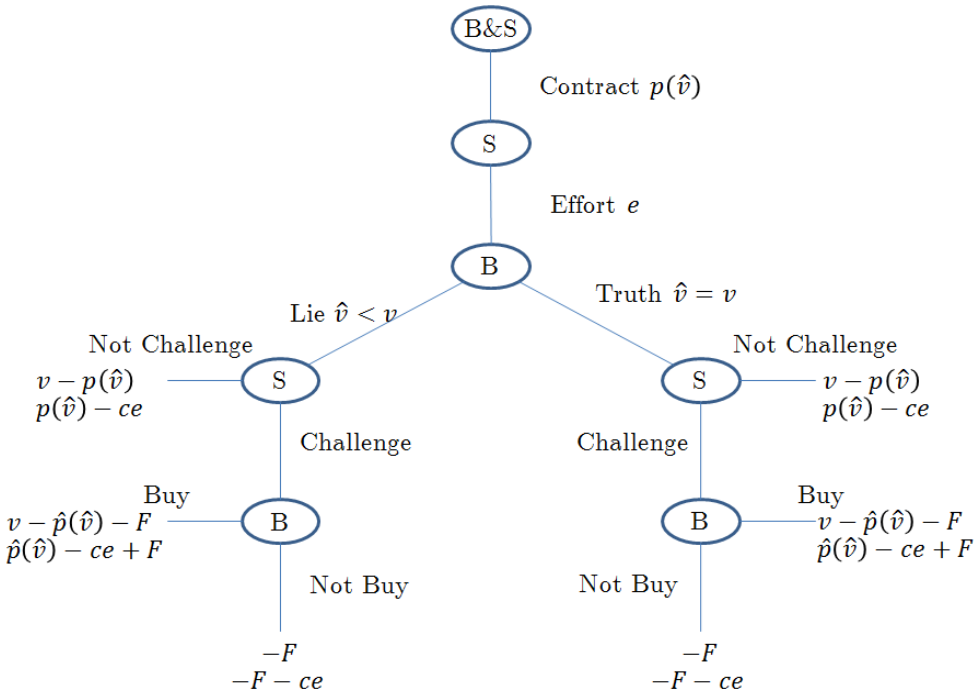

Figure 8: Maskin and Tirole mechanism

If the fine F is large enough, the unique SPNE of this game involves the following strategies. If B is challenged, he accepts the counter offer and buys the good at the counter-offer price if $\hat{v} < v$ and he rejects it if $\hat{v} \geq v$. S challenges B's announcement if and only if $\hat{v} < v$, and B announces $\hat{v} = v$. Therefore, B and S can, in the first stage, write a contract of the form $p(\hat{v}) = \hat{v} + k$, and as a result, S will choose $e^* = 1$.

To fix terminology, the mechanism starting from stage 4, after v has been realized, is a special case of the mechanisms introduced by Moore and Repullo (1988), so I will refer to that mechanism as the Moore and Repullo mechanism. The critique that messages arising from Moore and Repullo mechanisms can be used as a verifiable input into a contract to solve the hold-up problem (and indeed to implement a wide class of social choice functions) is known as the Maskin and Tirole (1999) critique. The main message of this criticism is that complete

information about payoff-relevant variables and common knowledge of rationality implies that verifiability is not an important constraint to (uniquely) implement most social choice functions, including those involving efficient investments in the Property Rights Theory model.

The existence of such mechanisms is troubling for the Property Rights Theory approach. However, the limited use of implementation mechanisms in real-world environments with observable but non-verifiable information has led several recent authors to question the Maskin and Tirole critique itself. As Maskin himself asks: "To the extent that [existing institutions] do not replicate the performance of [subgame-perfect implementation mechanisms], one must ask why the market for institutions has not stepped into the breach, an important unresolved question." (Maskin, 2002, p. 728)

Recent theoretical work by Aghion et al. (2012) demonstrates that the truth-telling equilibria in Moore and Repullo mechanisms are fragile. By perturbing the information structure slightly, they show that the Moore and Repullo mechanism does not yield even approximately truthful announcements for any setting in which multi-stage mechanisms are necessary to obtain truth-telling as a unique equilibrium of an indirect mechanism. Aghion et al. (2017) takes the Moore and Repullo mechanism into the laboratory and show that indeed, when they perturb the information structure away from common knowledge of payoff-relevant variables, subjects do not make truthful announcements.

Relatedly, Fehr et al. (2017) takes an example of the entire Maskin and Tirole critique into the lab and ensure that there is common knowledge of payoff-relevant variables. They show that in the game described above, there is a strong tendency for B's to reject counter offers after they have been challenged following small lies, S's are reluctant to challenge small lies, B's tend to make announcements with $\hat{v} < v$, and S's often choose low effort levels.

These deviations from SPNE predictions are internally consistent: if indeed B's reject counter offers after being challenged for telling a small lie, then it makes sense for S to be reluctant to challenge small lies. And if S often does not challenge small lies, then it makes

sense for B to lie about the value of the good. And if B is not telling the truth about the value of the good, then a contract that conditions on B's announcement may not vary sufficiently with S's effort choice to induce S to choose high effort.

The question then becomes: why do B's reject counter offers after being challenged for telling small lies if it is in their material interests to accept such counter offers? One possible explanation, which is consistent with the findings of many laboratory experiments, is that players have preferences for negative reciprocity. In particular, after B has been challenged, B must immediately pay a fine of F that he cannot recoup no matter what he does going forward. He is then asked to either accept the counter offer, in which case S is rewarded for appropriately challenging his announcement; or he can reject the counter offer (at a small, but positive, personal cost), in which case S is punished for inappropriately challenging his announcement.

The failure of subjects to play the unique SPNE of the mechanism suggests that at least one of the assumptions of Maskin and Tirole's critique is not satisfied in the lab. Since Fehr et al. (2017) is able to design the experiment to ensure common knowledge of payoff-relevant information, it must be the case that players lack common knowledge of preferences and rationality, which is also an important set of implicit assumptions that are part of Maskin and Tirole's critique. Indeed, Fehr et al. (2017) provides suggestive evidence that preferences for reciprocity are responsible for their finding that B's often reject counter offers.

The findings of Aghion et al. (2017) and Fehr et al. (2017) do not necessarily imply that it is impossible to find mechanisms in which in the unique equilibrium of the mechanisms, the hold-up problem can be effectively solved. What they do suggest, however, is that if subgame-perfect implementation mechanisms are to be more than a theoretical curiosity, they must incorporate relevant details of the environment in which they might be used. If people have preferences for reciprocity, then the mechanism should account for this. If people are concerned about whether their trading partner is rational, then the mechanism should account for this. If people are concerned that uncertainty about what their trading partner

is going to do means that the mechanism imposes undue risk on them, then the mechanism should account for this.

10 Financial Contracting

The last topic that we will cover in this class applies the tools we have developed over the last couple weeks in order to think about corporate governance, which Shleifer and Vishny (1997) define as "ways in which the suppliers of finance to corporations assure themselves of getting a return on their investment." We will think about a setting in which a capital-constrained Entrepreneur needs capital from capital-rich potential Investor to undertake a project that yields a positive return. We will look at the different instruments the Entrepreneur has to credibly commit herself to return funds to such an Investor in order to attract financing from them.

In a world of complete contracts and complete financial markets, how a project is financed—whether through debt or equity or some other, more complicated arrangement—is irrelevant for the total value of the project, and every positive net-present value project will be funded. The irrelevance result is known as the Modigliani-Miller theorem (Modigliani and Miller, 1958) and it is not so different from versions of the Coase theorem that we have mentioned in passing a few times. (Very) roughly speaking, we can think of the expected discounted revenues from the project as some value V. If undertaking the project requires K dollars worth of capital, then the Investor has to get at least K dollars back. One way he could get K dollars back is if he gets a share of the future revenues for which the expected present discounted value is K. Or the Entrepreneur could write a debt contract for which the expected present discounted value of payments is K. Either way, the Entrepreneur will receive $V - K$ and will undertake the project if $V > K$.

The Modigliani-Miller theorem served as a benchmark and spawned a literature providing explanations for when and why debt has advantages over equity based on two classes

of explanations: differences in tax treatment and incentive problems. Our focus will be on the latter and in particular on how different arrangements lead the Entrepreneur to make different decisions that in turn affect the value of the project. Without appropriate contractual safeguards, the Investor might worry that the Entrepreneur will make decisions that are privately beneficial to the Entrepreneur but harmful to the Investor. The moral hazard problems that arise in these settings may include insufficient effort on the part of the Entrepreneur, although this may not take the form of the Entrepreneur working too few hours, but rather that she might avoid unpleasant tasks like firing people or a taking a tough stance in negotiations with suppliers. The problem may take the form of unnecessary or extravagant investments aimed at growing the Entrepreneur's "empire" at the expense of the Investor's returns. Or it may take the form of self dealing and excessive perk consumption: buying costly private jets, expensive art for the corporate headquarters, or hiring friends and family members.

When actions like the ones described above are not contractible, credit may be *rationed* in the sense that the Entrepreneur may be unable to "obtain the loan [she] wants even though [she] is willing to pay the interest that the lenders are asking, and perhaps even a higher interest rate." (Tirole, 2005, p. 113) Positive net-present value projects may therefore not be undertaken. We will begin with a workhorse model that builds off our analysis of limited liability constraints to provide a reason why credit may be rationed. As in our earlier discussion of such models, the Entrepreneur must be given a rent in order to provide her with incentives to take the right action. The total returns from the Entrepreneur's project net of the incentive rents the Entrepreneur must receive is what we will refer to as her *pledgeable income*. Even if the overall income from the project would be high enough to cover the Investor's capital costs, if the Entrepreneur's pledgeable income is not, she will be unable to attract funding from the Investor.

The form of the optimal contract in this model can, depending on how you look at it, be interpreted either as a debt contract or as a contract involving outside equity. But it lacks

the richness of form that real financing arrangements take. In particular, when we think of equity, we typically think of a contract in which an outside Investor owns some share of a firm's profits and is also able to exercise some limited control over some of the firm's decisions. When we think of debt, we think of contracts in which the Investor is guaranteed some payments, and if the Entrepreneur does not repay the Investor, the Investor gains control over the associated assets and can then make decisions about how they are used. The model above has no notion of control rights, so it is unable to provide a compelling argument for why such contracts might move around control rights in a contingent way. We will therefore take an incomplete contracts view to think about how contingent control rights might be used in an optimal arrangement.

11 Pledgeable Income and Credit Rationing

There is a risk-neutral Entrepreneur (E) and a risk-neutral Investor (I). The Investor has capital but no project, and the Entrepreneur has a project but no capital. In order to pursue the project, the Entrepreneur needs K units of capital. Once the project has been pursued, the project yields revenues py, where $y \in \{0,1\}$ is the project's output, and p is the market price for that output. The Entrepreneur chooses an action $e \in [0,1]$ that determines the probability of a successful project, $\Pr[y=1|\,e] = e$, as well as a private benefit $b(e)$ that accrues to the Entrepreneur, where b is strictly decreasing and concave in e and satisfies $b'(0) = 0$ and $\lim_{e\to 1} b'(e) = -\infty$.

The Entrepreneur can write a contract $w \in \mathcal{W} = \{w : \{0,1\} \to \mathbb{R},\ 0 \leq w(y) \leq py\}$ that pays the Investor $w(y)$ if output is y and therefore shares the projects revenues with the Investor. If the Investor declines the contract, he keeps the K units of capital, and the Entrepreneur receives a payoff of 0. If the Investor accepts the contract, the Entrepreneur's

and Investor's preferences are

$$\begin{aligned} U_E(w,e) &= E[py - w(y)|\, e] + b(e) \\ U_I(w,e) &= E[w(y)|\, e]. \end{aligned}$$

There are strong parallels between this model and the limited-liability Principal-Agent model we studied earlier. We can think of the Entrepreneur as the Agent and the Investor as the Principal. There is one substantive difference and two cosmetic differences. The substantive difference is that the Entrepreneur is the one writing the contract, and while the contract must still satisfy the Entrepreneur's incentive-compatibility constraint, the individual rationality constraint it has to satisfy is the *Investor's*. The two cosmetic differences are: (1) the payments in the contract flow from the Entrepreneur to the Investor, and (2) instead of higher values of e costing the Entrepreneur $c(e)$, they reduce her private benefits $b(e)$.

Timing The timing of the game is as follows.

1. E offers I a contract $w(y)$, which is commonly observed.
2. I accepts the contract ($d = 1$) or rejects it ($d = 0$) and keeps K, and the game ends. This decision is commonly observed.
3. If I accepts the contract, E chooses action e and receives private benefit $b(e)$. e is only observed by E.
4. Output $y \in \{0, 1\}$ is drawn, with $\Pr[y = 1|\, e] = e$. y is commonly observed.
5. E pays I an amount $w(y)$. This payment is commonly observed.

Equilibrium The solution concept is the same as always. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0, 1\}$, an action choice $e^* : \mathcal{W} \times \{0, 1\} \to [0, 1]$ such that given contract w^*, the Investor optimally

chooses d^*, and the Entrepreneur optimally chooses e^*, and given d^*, the Investor optimally offers contract w^*. We will say that the optimal contract induces action e^*.

The Program The Entrepreneur offers a contract $w \in \mathcal{W}$, which specifies a payment $w(0) = 0$ and $0 \leq w(1) \leq p$ and proposes an action e to solve

$$\max_{w(1),e} (p - w(1)) e + b(e)$$

subject to the incentive-compatibility constraint

$$e \in \operatorname*{argmax}_{\hat{e} \in [0,1]} (p - w(1)) \hat{e} + b(\hat{e}),$$

the Investor's individual-rationality (or break-even) constraint

$$w(1) e \geq K.$$

Analysis We can decompose the problem into two steps. First, we can ask: for a given action e, how much rents must the Entrepreneur receive in order to choose action e, and therefore, what is the maximum amount that the Investor can be promised if the Entrepreneur chooses e? Second, we can ask: given that the Investor must receive K, what action e^* maximizes the Entrepreneur's expected payoff?

The following figure illustrates the problem using a graph similar to the one we looked at when we thought about limited liability constraints. The horizontal axis is the Entrepreneur's action e, and the segment pe is the expected revenues as a function of e. The dashed line $(p - w_{e_1}) e$ represents, for a contract that pays the Investor $w(1) = w_{e_1}$ if $y = 1$, the Entrepreneur's expected monetary payoff, and $-b(e)$ represents the Entrepreneur's cost of choosing different actions. As the figure illustrates, the contract that gets the Entrepreneur to choose action e_1 can pay the Investor at most $w_{e_1} e_1$ in expectation.

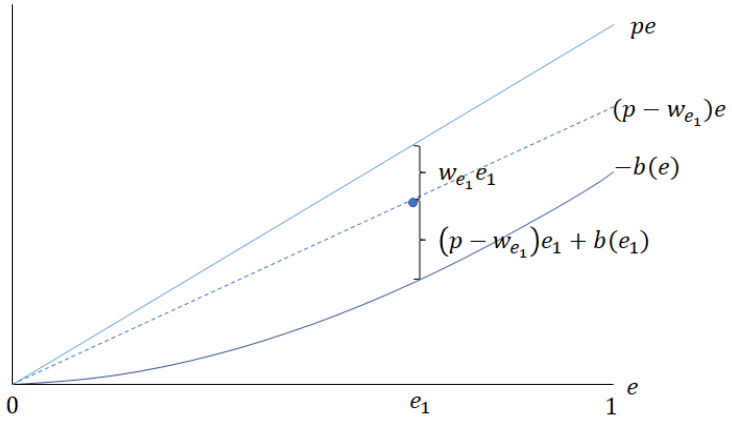

Figure 9: Entrepreneur Incentive Rents

The next figure illustrates, for different actions e, the rents $(p - w_e)\, e + b\,(e)$ that the Entrepreneur must receive for e to be incentive-compatible. Note that because $w_e \geq 0$, there is no incentive-compatible contract that gets the Entrepreneur to choose any action $e > e^{FB}$. The vertical distance between the expected revenue pe curve and the Entrepreneur rents curve is the Investor's expected payoff under the contract that gets the Entrepreneur to choose action e. For the Investor to be willing to sign such a contract, that vertical distance must be at least K, which is the amount of capital the Entrepreneur needs.

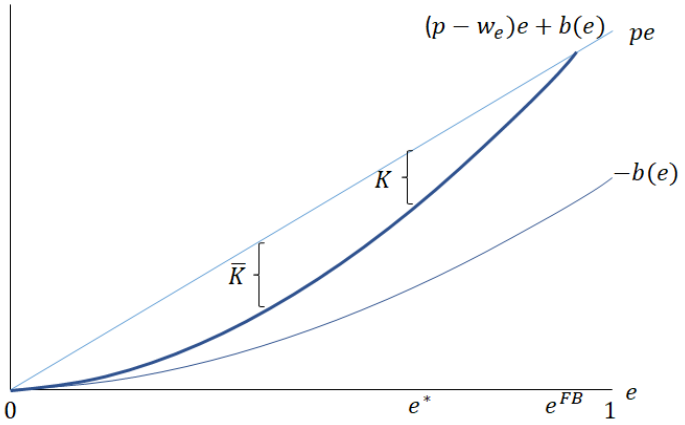

Figure 10: Equilibrium and Pledgeable Income

Two results emerge from this analysis. First, if $K > 0$, then in order to secure funding K, the Entrepreneur must share some of the project's earnings with the Investor, which means that the Entrepreneur does not receive all the returns from her actions and therefore will choose an action $e^* < e^{FB}$. Second, the value $\bar{K}$ represents the maximum expected payments the Entrepreneur can promise the Investor in any incentive-compatible contract. This value is referred to as the Entrepreneur's **pledgeable income**. If the project requires capital $K > \bar{K}$, then there is no contract the Entrepreneur can offer the Investor that the Investor will be willing to sign, even though the Entrepreneur would invest in the project if she had her own capital. When this is the case, we say that there is **credit rationing**.

As a final point about this model, with binary output, the optimal contract can be interpreted as either a debt contract or an equity contract. Under the debt contract interpretation, the Entrepreneur must reimburse w_{e^*} or else go bankrupt, and if the project is successful, she keeps the residual $p - w_{e^*}$. Under the equity contract interpretation, the Entrepreneur holds a share $(p - w_{e^*})/p$ of the project's equity, and the Investor holds a share w_{e^*}/p of the project's equity. That the optimal contract can be interpreted as either a debt contract or an equity contract highlights that if we want to actually understand the role of debt or equity contracts, we will need a richer model.

12 Control Rights and Financial Contracting

The previous model cannot explain the fact that equity has voting power while debt does not, except following default. Aghion and Bolton (1992) takes an incomplete contracting approach to thinking about financial contracting and brings control rights front and center. We will look at a simple version of the model that provides an explanation for debt contracts featuring *contingent* control. In this model, control rights matter because the parties disagree about important decisions that are ex ante noncontractible. The parties will renegotiate over these decisions ex post, but because the Entrepreneur is wealth-constrained, renegotiation

may not fully resolve the disagreement. Investor control will therefore lead to a smaller pie ex post, but the Investor will receive a larger share of that pie. As a result, even though Investor control destroys value, it may be the only way to get the Investor to be willing to invest to begin with.

The Model As in the previous model, there is a risk-neutral Entrepreneur (E) and a risk-neutral investor (I). The Investor has capital but no project, the Entrepreneur has a project but no capital, and the project costs K. The parties enter into an agreement, which specifies who will possess the right to make a decision $d \in \mathbb{R}_+$ once that decision needs to be made. After the state $\theta \in \mathbb{R}_+$, which is drawn according to density $f(\theta)$, is realized, the decision d is made. This decision determines verifiable profits $y(d)$, which we will assume accrue to the Investor.[2] It also determines nonverifiable private benefits $b(d)$ that accrue to the Entrepreneur.

The parties can contract upon a rule that specifies who will get to make the decision d in which state of the world: let $g : \mathbb{R}_+ \to \{E, I\}$ denote the **governance structure**, where $g(\theta) \in \{E, I\}$ says who gets to make the decision d in state θ. The decision d is itself not ex ante contractible, but it is ex post contractible, so that the parties can negotiate over it ex post. In particular, we will assume that the Entrepreneur has all the bargaining power, so that she will propose a take-it-or-leave-it offer specifying a decision d as well as a transfer $w \geq 0$ from the Investor to the Entrepreneur. Note that the transfer has to be nonnegative, because the Entrepreneur is cash-constrained.

Timing

1. E proposes a governance structure g. g is commonly observed.

[2] We could enrich the model to allow the parties to contract ex ante on the split of the verifiable profits that each party receives. Giving all the verifiable profits to the Investor maximizes the efficiency of the project because it maximizes the pledgeable income that he can receive without having to distort ex post decision making.

2. I chooses whether or not to go ahead with the investment. This decision is commonly observed.

3. The state θ is realized and is commonly observed.

4. E makes a take-it-or-leave-it offer of (d, w) to I, who either accepts or rejects it.

5. If I rejects the offer, party $g(\theta)$ chooses d.

Analysis As usual, let us start by describing the first-best decision that maximizes the sum of the profits and the private benefits:

$$d^{FB} \in \underset{d \in \mathbb{R}_+}{\operatorname{argmax}}\, y(d) + b(d).$$

Assume y and b are strictly concave and single-peaked, so that there is a unique first-best decision. Moreover, assume $y(d)$ is maximized at some decision d^I, and $b(d)$ is maximized at some other decision $d^E < d^I$. These assumptions imply that $d^E < d^{FB} < d^I$. Now, let us see what happens depending on who has control.

We will first look at what happens under Entrepreneur control. This corresponds to $g(\theta) = E$ for all θ. In this case, if the Investor rejects the Entrepreneur's offer in stage 4, the Entrepreneur will choose d to maximize her private benefit and will therefore choose d^E. Recall that the Entrepreneur does not care about the profits of the project because we have assumed that the profits accrue directly to the Investor. The decision d^E is therefore the Investor's outside option in stage 4. It will not be the decision that is actually made, however, because the Entrepreneur can offer to make a higher decision in exchange for some money. In particular, she will offer $\left(d^{FB}, w\right)$, where w is chosen to extract all the ex post surplus from the Investor:

$$y\left(d^{FB}\right) - w = y\left(d^E\right) \text{ or } w = y\left(d^{FB}\right) - y\left(d^E\right) > 0.$$

Under Entrepreneur control, the Entrepreneur's payoff will therefore be $b\left(d^{FB}\right)+y\left(d^{FB}\right)-y\left(d^{E}\right)>b\left(d^{E}\right)$, and the Investor's payoff will be $y\left(d^{E}\right)$, which is effectively the Entrepreneur's pledgeable income. If $y\left(d^{E}\right)>K$, then the Investor will make the investment, and the first-best decision will be made, but if $y\left(d^{E}\right)<K$, this arrangement will not get the Investor to make the investment.

Now let us look at what happens under Investor control, which corresponds to $g(\theta)=I$ for all θ. In this case, if the Investor rejects the Entrepreneur's offer at stage 4, the Investor will choose d to maximize profits and will therefore choose d^{I}. The decision d^{I} is therefore the Investor's outside option in stage 4. At stage 4, the Entrepreneur would like to get the Inventor to make a decision $d<d^{I}$, but in order to get him to do so, she would have to choose $w<0$, which is not feasible. As a result, d^{I} will in fact be the decision that is made. Under Investor control, the Entrepreneur's payoff will be $b\left(d^{I}\right)$, and the Investor's payoff will be $y\left(d^{I}\right)$, which again is effectively the Entrepreneur's pledgeable income. Conditional on the investment being made, total surplus under Investor control is lower than under Entrepreneur control, but the benefit of Investor control is that it ensures the Investor a payoff of $y\left(d^{I}\right)$, which may exceed K even if $y\left(d^{E}\right)$ does not.

As in the Property Rights Theory, decision rights determine parties' outside options in renegotiations, which determines their incentives to make investments that are specific to the relationship. In contrast to the PRT, however, ex post renegotiation does not always lead to a surplus-maximizing outcome because the Entrepreneur is wealth-constrained. As such, in order to provide the Investor with incentives to make the relationship-specific investment of investing in the project, we may have to give the Investor ex post control, even though he will use it in a way that destroys total surplus.

If $y\left(d^{I}\right)>K>y\left(d^{E}\right)$, then Investor control is better than Entrepreneur control because it ensures the Investor will invest, but in some sense, it involves throwing away more surplus than necessary. In particular, consider a governance structure $g(\cdot)$ under which the Entrepreneur has control with probability π (i.e., $\Pr[g(\theta)=E]=\pi$), and the Investor has

control with probability $1-\pi$ (i.e., $\Pr[g(\theta)=I]=1-\pi$). The Entrepreneur can get the Investor to invest if she chooses π to satisfy

$$\pi y\left(d^{E}\right)+(1-\pi)\, y\left(d^{I}\right)=K,$$

which will be optimal.

Now, stochastic control in this sense is a bit tricky to interpret, but with a slight elaboration of the model, it has a more natural interpretation. In particular, suppose that the state of the world, θ, determines how sensitive the project's profits are to the decision, so that

$$y(d,\theta)=\alpha(\theta)\, y(d)+\beta(\theta),$$

where $\alpha(\theta)>0$, and $\alpha'(\theta)<0$. In this case, the optimal governance structure would involve a cutoff θ^{*} so that $g(\theta)=E$ if $\theta>\theta^{*}$ and $g(\theta)=I$ if $\theta\leq\theta^{*}$, where this cutoff is chosen so that the Investor's expected payoffs would be K.

If $\alpha'(\theta)\, y(d)+\beta'(\theta)>0$ for all d, then high-θ states correspond to high-profit states, and this optimal arrangement looks somewhat like a debt contract that gives control to the creditor in bad states and gives control to the Entrepreneur in the good states. In this sense, the model captures an important aspect of debt contracts, namely that they involve contingent allocations of control. This theory of debt contracting is not entirely compelling, though, because the most basic feature of debt contracts is that the shift in control to the Investor occurs *only if the Entrepreneur does not make a repayment.* The last model we will look at will have this feature.

13 Cash Diversion and Liquidation

We will look at one final model that involves an important decision that is often specified in debt contracts: whether to liquidate an ongoing project. We will show that when the

firm's cash flows are noncontractible, giving the Investor the rights to the proceeds from a liquidation event can protect him from short-run expropriation from an Entrepreneur who may want to direct the project's cash flows toward her own interests.

The Model As before, there is a risk-neutral Entrepreneur (E) and a risk-neutral investor (I). The Investor has capital but no project, the Entrepreneur has a project but no capital, and the project costs K. If the project is funded, it yields income over two periods, which accrue to the Entrepreneur. In the first period, it produces output $y_1 \in \mathcal{Y}_1 \equiv \{0, 1\}$, where $\Pr[y_1 = 1] = q$, and that output generates a cash flow of $p_1 y_1$. After y_1 is realized, the Entrepreneur can make a cash payment $0 \leq \hat{w}_1 \leq p_1 y_1$ to the Investor. The project can then be terminated, yielding a liquidation value of L, where $0 \leq L \leq K$, which accrues to the Investor. Denote the probability the project is continued by $r \in [0, 1]$. If the project is continued, in the second period, it produces output $y_2 = 1$, and that output generates cash flow of p_2. At this point, the Entrepreneur can again make a cash payment $0 \leq \hat{w}_2 \leq p_2$ to the Investor.

The cash flows are noncontractible, so the parties are unable to write a contract that specifies output-contingent repayments from the Entrepreneur to the Investor, but they can write a contract that specifies probabilities $r : \mathbb{R}_+ \to [0, 1]$ that determine the probability $r(\hat{w}_1)$ the project is continued if the Entrepreneur pays the Investor $\hat{w}_1$. The contracting space is therefore $\mathcal{W} = \{r : \mathbb{R}_+ \to [0, 1]\}$. The players' payoffs, if the Investor invests K in the project are:

$$\begin{aligned} u_E(\ell, y_1, \hat{w}_1, \hat{w}_2) &= p_1 y_1 - \hat{w}_1 + r(\hat{w}_1)(p_2 - \hat{w}_2) \\ u_I(\ell, y_1, \hat{w}_1, \hat{w}_2) &= \hat{w}_1 + (1 - r(\hat{w}_1))L + r(\hat{w}_1)\hat{w}_2. \end{aligned}$$

Throughout, we will assume that $p_2 > L$, so that liquidation strictly reduces total surplus.

Timing The timing of the game is as follows.

1. E offers I a contract $r(\hat{w}_1)$, which is commonly observed.

2. I accepts the contract ($d = 1$) or rejects it ($d = 0$) and keeps K, and the game ends. This decision is commonly observed.

3. If I accepts the contract, output $y_1 \in \{0, 1\}$ is realized. y_1 is commonly observed.

4. E makes a payment $0 \leq \hat{w}_1 \leq p_1 y_1$ to I. $\hat{w}_1$ is commonly observed.

5. The project is liquidated with probability $1 - r(\hat{w}_1)$. The liquidation event is commonly observed.

6. If the project has not been liquidated, output $y_2 = 1$ is realized. y_2 is commonly observed.

7. E makes a payment $0 \leq \hat{w}_2 \leq y_2$ to I. $\hat{w}_2$ is commonly observed.

Equilibrium The solution concept is the same as always. A **pure-strategy subgame-perfect equilibrium** is a continuation function $r^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0, 1\}$, a first-period payment rule $w_1^* : \mathcal{W} \times \{0, 1\} \to \mathbb{R}_+$, and a second-period payment rule $w_2^* : \mathcal{W} \times \{0, 1\} \times \{0, 1\} \times \mathbb{R}_+ \to \mathbb{R}_+$ such that given continuation function r^* and payment rules w_1^* and w_2^*, the Investor optimally chooses d^*, and given d^*, the Entrepreneur optimally offers continuation function r^* and chooses payment rules w_1^* and w_2^*.

The Program Models such as this one, in which the Entrepreneur's repayment decisions are not contractible, are referred to as **cash diversion** models. The Entrepreneur's problem will be to write a contract that specifies continuation probabilities and repayment amounts so that given those repayment-contingent continuation probabilities, the Entrepreneur will actually follow through with those repayments, and the Investor will at least break even. In this setting, it is clear that in any subgame-perfect equilibrium, the Entrepreneur will not make any positive payment $\hat{w}_2 > 0$, since she receives nothing in return for doing

so. Moreover, it will be without loss of generality for the Entrepreneur to specify a single repayment amount $0 < w_1 \leq p_1$ to be repaid if $y_1 = 1$, and a pair of probabilities r_0 and r_1, where r_0 is the probability the project is continued (and not liquidated) if $\hat{w}_1 \neq w_1$, and r_1 is the probability the project is continued if $\hat{w}_1 = w_1$. The Entrepreneur's problem is therefore

$$\max_{r_0, r_1, w_1 \leq p_1} q\left(p_1 - w_1 + r_1 p_2\right) + \left(1 - q\right) r_0 p_2$$

subject to the Entrepreneur's incentive-compatibility constraint

$$p_1 - w_1 + r_1 p_2 \geq p_1 + r_0 p_2$$

and the Investor's break-even constraint

$$q\left(w_1 + \left(1 - r_1\right) L\right) + \left(1 - q\right)\left(1 - r_0\right) L \geq K.$$

It will be useful to rewrite the incentive-compatibility constraint as

$$\left(r_1 - r_0\right) p_2 \geq w_1,$$

which says that in order for repayment w_1 to be incentive-compatible, it has to be the case that by making the payment w_1 (instead of paying zero), the probability r_1 that the project is continued (and hence the Entrepreneur receives p_2) if she makes the payment is sufficiently high relative to the probability r_0 the project is continued when she does not make the payment.

Analysis In order to avoid multiple cases, we will assume that

$$p_1 > \frac{p_2}{q p_2 + \left(1 - p\right) L} K,$$

which will ensure that in the optimal contract, the Entrepreneur's first-period payment will satisfy $w_1^* < p_1$.

The Entrepreneur's problem is just a constrained maximization problem with a linear objective function and linear constraints, so it can in principle be easily solved using standard linear-programming techniques. We will instead solve the problem by thinking about a few perturbations that, at the optimum, must not be profitable. Taking this approach allows us to get some intuition for why the optimal contract will take the form it does.

First, we will observe that the Investor's break-even constraint must be binding in any optimal contract. To see why, notice that if the constraint were not binding, we could reduce the payment amount w_1 by a little bit and still maintain the break-even constraint. Reducing w_1 makes the incentive-compatibility constraint easier to satisfy, and it increases the Entrepreneur's objective function. This argument tells us that the Entrepreneur will receive all of the surplus the project generates, so her problem is to maximize that surplus.

The second observation is that in any optimal contract, the project is never liquidated following repayment. To see why, suppose $r_0 < r_1 < 1$ so that the project is continued with probability less than one following repayment. Consider an alternative contract in which r_1 is increased to $r_1 + \varepsilon$, for $\varepsilon > 0$ small. Since making this change alone will violate the Investor's breakeven constraint, let us also increase w_1 by εL so that

$$w_1 + \varepsilon L + (1 - r_1 - \varepsilon) L = w_1 + (1 - r_1) L.$$

Under this perturbation, the Investor's breakeven constraint is still satisfied, and the Entrepreneur's incentive-compatibility constraint is satisfied as long as

$$(r_1 + \varepsilon - r_0) p_2 \geq w_1 + \varepsilon L,$$

which is true because $(r_1 - r_0) p_1 \geq w_1$ (or else the original contract did not satisfy IC) and $\varepsilon (p_2 - L) > 0$ since continuing the project is optimal (i.e., $p_2 > L$). If the original contract

satisfied IC and IR, then so does this one, but this one also increases the Entrepreneur's objective by $q(-\varepsilon L + \varepsilon p_2)$, which again is strictly positive, since $p_2 > L$. This perturbation shows that increasing the probability of continuing the project following repayment is good for two reasons: it reduces the probability of inefficient liquidation, and it increases the Entrepreneur's incentives to repay.

Finally, the last step will be to show that the incentive constraint must bind at the optimum. It clearly must be the case that $r_0 < 1$, or else the incentive constraint would be violated. Again, suppose that the incentive constraint was not binding. Then consider a perturbation in which we raise r_0 to $r_0 + \varepsilon$, and to maintain the breakeven constraint, we increase w_1 to $w_1 + \varepsilon L(1-q)/q$. If the incentive constraint was not binding, then it will still be satisfied if r_0 is raised by a little bit. Lastly, this perturbation increases the Entrepreneur's payoff by

$$-q\left[\frac{\varepsilon L(1-q)}{q}\right] + (1-q)\varepsilon p_2 = (1-q)(p_2 - L)\varepsilon > 0.$$

In other words, if the incentive constraint is not binding, it is more efficient for the Entrepreneur to pay the Investor with cash than with an increased probability of liquidation, and since the Entrepreneur captures all the surplus, she will choose to pay in this more efficient way as much as she can.

To summarize, these three perturbations show that any optimal contract in this setting has to satisfy

$$(1 - r_0^*)\, p_2 = w_1^*$$

and

$$q w_1^* + (1-q)(1 - r_0^*)\, L = K.$$

This is just two equations in two unknowns, so we can solve for the probability that the

project is liquidated following nonpayment:

$$1 - r_0^* = \frac{K}{qp_2 + (1-q)L} > 0.$$

There is a complementarity between the repayment amount and the liquidation probability: if the project requires a lot of capital (i.e., K is large), then the Investor needs to be assured a bigger payment, and in order to assure that bigger payment, the project has to be liquidated with higher probability following nonpayment. If the project has high second-period cash flows (i.e., p_2 is high), then the Entrepreneur loses a lot following nonpayment, so the project does not need to be liquidated with as high of a probability to ensure repayment. Finally, if the liquidation value of the project is high, then the Investor earns more upon liquidation, so he can break even at a lower liquidation probability.

Under the first-best outcome, the project will never be liquidated, and the project will be undertaken as long as the expected cash flows exceed the required capital, or $qp_1 + p_2 > K$. The model features two sources of inefficiencies relative to the first-best outcome. First, in order to assure repayment, the Entrepreneur commits to a contract that with some probability inefficiently liquidates the project.

Second, there is credit rationing: the maximum amount the Entrepreneur can promise the Investor is p_2 in the event that output is high in the first period and L in the event that it is not, so if

$$qp_2 + (1-q)L < K < qp_1 + p_2,$$

the project will be one that should be undertaken but, in equilibrium, will not be undertaken. The liquidation value of the project is related to the collateral value of the assets underlying the project, and there is a literature beginning with Kiyotaki and Moore (1997) that endogenizes the market value of those assets and shows there can be important general equilibrium spillovers across firms.