Topic 10: Agency

Eddie Shore
Corporate Finance – ECONS4280

August 3, 2022

In this set of notes, we introduce and develop the notion of agency problems. We start with a brief discussion of economic models: why we use them, and how we should interpret them. We then turn to outlining the intuition behind a simple Principal-Agent model. With this in place, we discuss the principle[1] insights that agency theory provides: that contract *shape* matters, that steeper returns to effort create stronger incentives, and that steeper slopes are not *always* better. We then turn to a slightly more involved model that relaxes the assumption that *output* is observable. Instead, the principal has access to *performance metrics.* We will show that both the *scale* and the *alignment* of the performance metrics are important for determining the optimal contract. We then show how failure to recognise the potential mismatch between performance and profitability metrics have created major problems for companies in real world settings. Okay, let's go!

1 Economic Models

What is an economic model? They are much maligned! Listen to some commentators and you would imagine that economists build models out of some evangelical, feverish desire to crowbar mathematics into inappropriate settings simply for the pleasure of excluding the uninitiated. This is sometimes correct!! A favorite meme of mine outlines a satire of the types of paper economists tend to write. I've repeated it here:

[1] Get it??

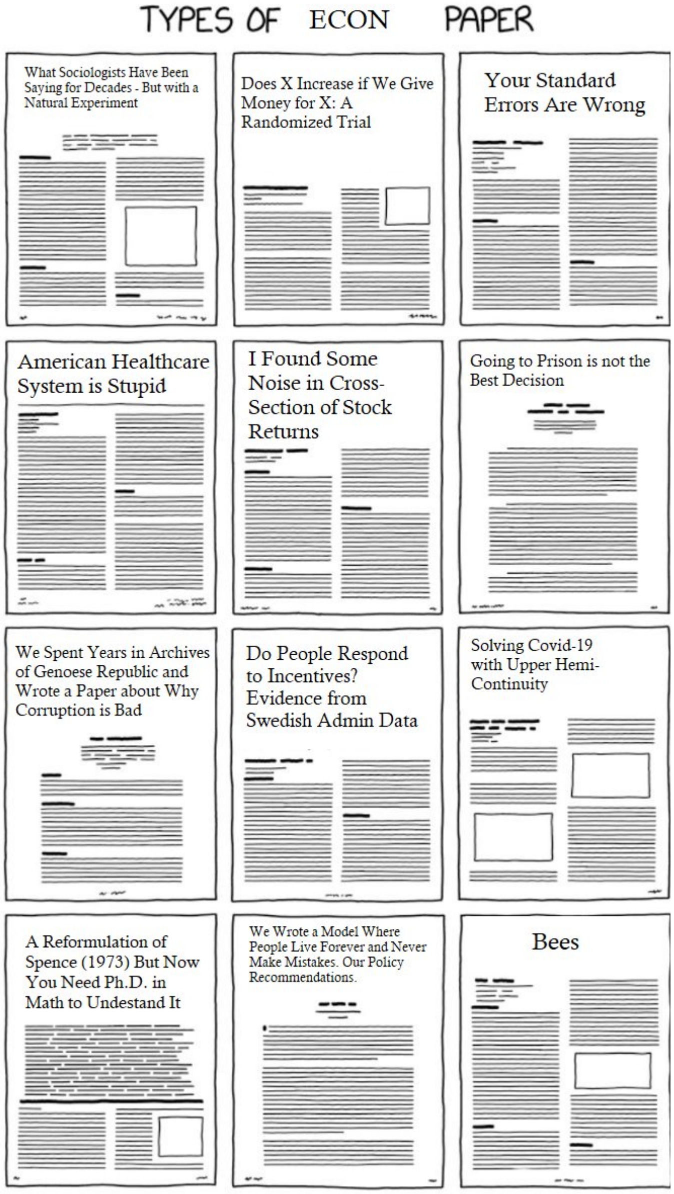
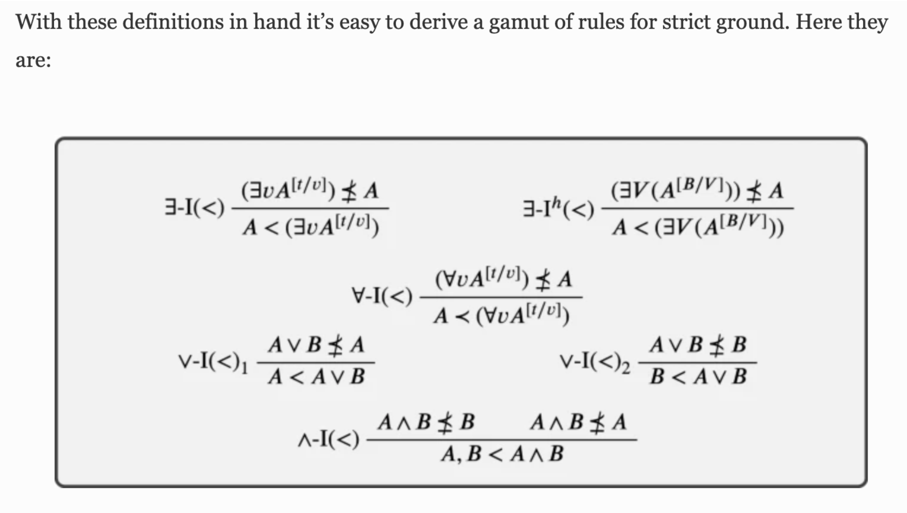

TYPES OF ECON PAPER
What Sociologists Have Been Saying for Decades - But with a Natural Experiment
Does X Increase if We Give Money for X: A Randomized Trial
Your Standard Errors Are Wrong
American Healthcare System is Stupid
I Found Some Noise in Cross-Section of Stock Returns
Going to Prison is not the Best Decision
We Spent Years in Archives of Genoese Republic and Wrote a Paper about Why Corruption is Bad
Do People Respond to Incentives? Evidence from Swedish Admin Data
Solving Covid-19 with Upper Hemi-Continuity
A Reformulation of Spence (1973) But Now You Need Ph.D. in Math to Undestand It
We Wrote a Model Where People Live Forever and Never Make Mistakes. Our Policy Recommendations.
Bees

Figure 1: To be an economist!

My personal favorite is, ‘We wrote a model where people live forever and never make mistakes. Our policy recommendations’. Lol.

However, when done *correctly*, economic models can be both *useful* and *insightful*. Mathematics is a tool to formalise an intuition. It is a way of implementing techniques that create meaning from base principles. As you become more familiar with them, their appeal *does* grow stronger. Think of it as the economist’s version of formal logic. I have the great pleasure of being close friend’s with a philosopher, and I once asked him to send me one of his paper’s so I could read it. I was initially taken aback, perhaps even offended, when he said, ‘Sure, but you probably will not understand it’. Hmph!! We’ll see about that!! Well, he was right... Here is a screenshot from one of his papers:



With these definitions in hand it’s easy to derive a gamut of rules for strict ground. Here they are:

$$\exists\text{-I}(<)\ \frac{(\exists v A^{[t/v]}) \not\leq A}{A < (\exists v A^{[t/v]})} \qquad \exists\text{-I}^{h}(<)\ \frac{(\exists V(A^{[B/V]})) \not\leq A}{A < (\exists V(A^{[B/V]}))}$$

$$\forall\text{-I}(<)\ \frac{(\forall v A^{[t/v]}) \not\leq A}{A \prec (\forall v A^{[t/v]})}$$

$$\vee\text{-I}(<)_1\ \frac{A \vee B \not\leq A}{A < A \vee B} \qquad \vee\text{-I}(<)_2\ \frac{A \vee B \not\leq B}{B < A \vee B}$$

$$\wedge\text{-I}(<)\ \frac{A \wedge B \not\leq B \qquad A \wedge B \not\leq A}{A, B < A \wedge B}$$

Figure 2: Ahh, yes! It all slots into place now!

No one denies the benefit of thinking logically. Mathematics is simply a way of *enforcing* a standard of logic that is *universal* in its meaning. There can be (mostly) no bullshitting when it comes to mathematics: what you say is explicit and interpretable to anyone familiar with the techniques.

So what is the goal in establishing this logic? A model can have *qualitative* or *quantitative* predictions. A qualitative prediction is that ‘when x goes up, y goes down’. A quantitative prediction is that $y = \frac{1}{x}$. These predictions are useful when they *pair well with what we see in the data*. This is the classic method by which models are tested.

For qualitative predictions, all we need to show is that when x goes up, y does indeed go down. A classic example of this would be the finding from our final class on corporate financing: companies with higher risk should have lower debt-to-equity ratios. Qualitative results are typically resistant to small changes in the model's specific assumptions, but more importantly *may be preserved as we add richness* to the model.

Contrast this with *quantitative* predictions. A good example of this from our class is CAPM: in that model, there is an *explicit* ratio between returns and a *specific quantitative* object: β. Models with quantitative predictions are only sensible if the assumptions we're making are a close match to the environment we're studying. It is a frustration with the obsessive pursuit of quantitative implications in macroeconomics that ultimately led me to leave the field: by construction the model is insufficiently close to, I don't know, the entire global economy, for quantitative predictions to be meaningful, *at least in my view!*

So what about the principal-agent model? In my view this is a model that delivers, and delivers big, on the *qualitative* set of predictions. It is also a *very* robust model, in the sense that its basic lessons are retained even as the model increases in complexity. As such, it is often seen as a *classic*, and a well-deserved one at that. So, what is it, and how does it work?

2 A simple principal-agent model

Principal-agent frameworks relate to situations where someone (the *principal*) wishes to incentivise someone else (the *agent*) into performing a certain task. If the agent's actions are visible, this is easy: I will pay you if you do x, and if you don't I'll give you diddly squat. If the agent's actions are entirely invisible, then it's hard to imagine how we can incentivise them at all! However, in most settings, actions are *neither* fully observable, *nor* perfectly unobservable. Even direct observation of agents actual behavior can be misleading, as this sensational quote from the extraordinary book, 'One day in the life of Ivan Denisovich' by Aleksandr Solzhenitsyn, shows:

> "Work was like a stick. It had two ends. When you worked for the knowing you gave them quality; when you worked for a fool you simply gave him eyewash."

So, the challenge is to understand how to incentivise an agent when you only have access to a *limited* impression of their behavior. There are four elements to every

Principal-Agent framework:

- The technology of production (*how does effort translate into output*)
- The set of feasible contracts (*what can the Principal offer?*
- The payoffs to parties (*what's in it for me, and you?*)
- The timing of events

Let's consider each in turn.

2.1 The Technology of Production

In the simplest framework, we assume the production process is summarised by three variables:

- The agent's contribution to firm value, y
- The action the agent takes, a
- The events that are beyond the agent's control, ϵ

The simplest way of expressing these things together is just as a simple linear sum:

$$y = a + \epsilon$$

So what is y, the contribution of the worker to firm value? Sometimes this is easy to conceptualise: a sharecropper, for instance, will produce a certain amount of verifiable output each year. A CEO by contrast delivers value in a far less transparent fashion: perhaps we consider their output as the *change in the wealth of shareholders* through appreciation of the stock price? Even more oblique is the contribution of general firm employees: for those of you who have worked, have you ever felt some degree of uncertainty as to whether you're adding value, or how you might be doing so?

The action that the agent takes, a, can be understood most straightforwardly as *effort*. However, this is not the only way we can interpret this notion: for a CEO, the action may have less to do with their 'effort level' than their 'focus'. All CEOs work very, very, very hard (just ask one!), but the actions they take will have varying levels of impact on shareholders. CEOs may prioritise projects that fulfil their own goals, be that ‘empire-building’, or ‘insider-led corporate philanthropy’. In any case, the key idea is that the action a is one that the agent *would not take* without being incentivised to do so, otherwise there is no tension. It is also reasonable in the context of a strategic interaction over rewards: as Heath Ledger’s Joker tells us, ‘if you’re good at something, never do it for free’.

To capture the idea that effort is only *partially* observable, we introduce the idea that some events are beyond the agent’s control. By adding ϵ, we cannot simply observe output and back out what effort must have been. Perhaps this year’s low stock price is simply an expression of ‘animal spirits’ within the market place? We cannot know *definitively* what the action/effort was by simply looking at output. In general, we assume that $\mathbb{E}[\epsilon] = 0$.

2.2 The set of feasible contracts

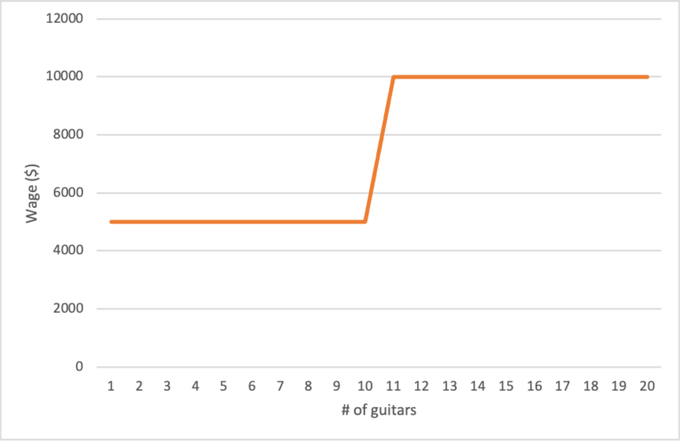
Note that there are many ways to skin a cat, and at least as many and more to write a contract. For example, suppose you are a piece-rate worker hand-making guitars for me in my boutique California factory: I could offer you a contract that would pay you a rate of $w = 0.12418g^4 + 6.43g^3 + 0.914g$ for each guitar, g, that you make. Or, I could agree to pay you double every time there is a lunar eclipse. Or I could pay you more whenever you wear yellow to work. All of these are potential contracts that, with the exception I would think of the last one, are feasible.

But we have a problem: finding the ‘right’ contract out of this enormous set of possible options is too hard! That’s why we always *restrict* the set of feasible contracts so we can *focus* on a manageable set of possibilities. The simplest contract, and the one we’ll work with, is a linear contract, i.e.

$$w = s + by$$

Here, the s can be thought of as a salary, and the b can be thought of as the ‘bonus’ rate. As well as being simple, a linear contract benefits from *always delivering the same return* to effort. To illustrate this with an example, suppose we had a *non*-linear contract of the following form: I will pay you \$5,000 if you make less than or equal to 10 guitars, and \$10,000 if you make more than 10 guitars. In terms of the number of guitars I make, the payoffs look like this:

What would you do in this scenario? Would you ever make more than 11 guitars? The *returns* to extra effort are not fixed across *how many guitars you’ve made so*



far. In a linear contract, the return to one more guitar *is always the same*.

2.3 Payoffs

In the simplest model of Principal-Agent, the Principal receives the Agent's contribution, y, minus the wage, w, and we call this the *profit*, π. So:

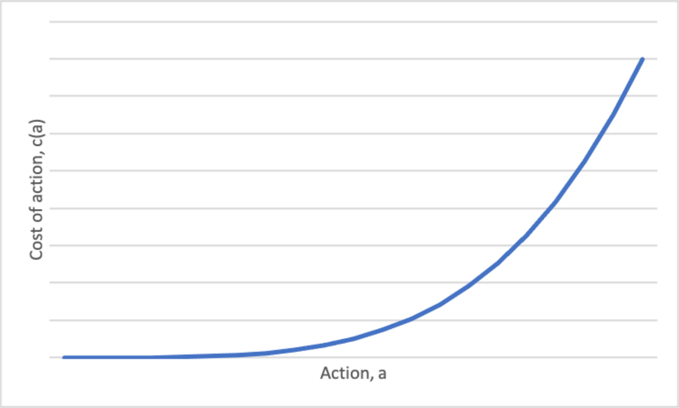
$$\pi = y - w$$

We usually assume that the principal is *risk-neutral*. What that means is that *they only care about the size of this expected payoff*. They do not care about the *variance*, because they are *indifferent* to risk.

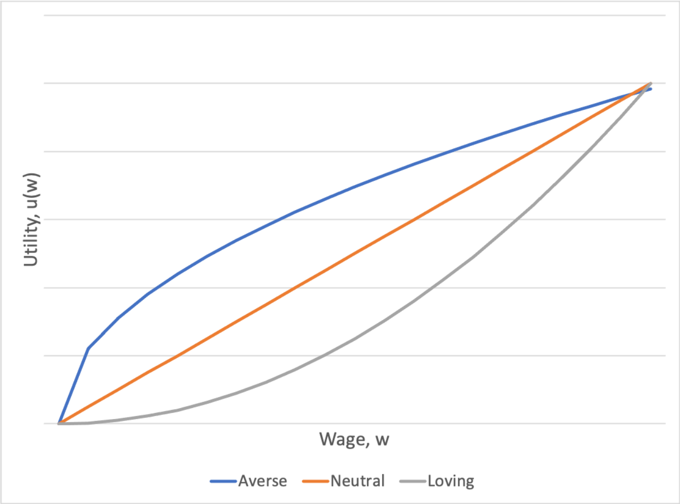
The agent of course receives the wage, w, but also pays a *cost* for performing a. Again, recall that for this problem to be *interesting*, the action needs to be something that the agent *would not do ordinarily*. We capture this cost using a *cost function*, $c(a)$, where $c(0) = 0$, and $\frac{dc}{da} > 0$. Another assumption we typically make is that $\frac{d^2c}{da^2} > 0$. What does this mean? It means that as we increase a, it gets progressively more and more costly to exert that action. Here's a graph to illustrate:

Why do we make this assumption? Well, because it is *reasonable*. Going from working 1 hours a day to working 4 hours a day is not so bad, but going from working 9 hours a day to 12 hours a day would suck! At the extreme, imagine going from working 21 hours a day to 24: this would kill us!

The payoff to the agent is then taken as the difference between the benefit of w, and the cost of action a. Note that in a general form, the benefit of w need not necessarily equal w! We could have $u(w)$, as is common in micro. The shape of



$u(w)$ determines whether or not the agent is risk *averse*, risk *neutral*, or risk *loving*. Can you guess what shapes correspond to which?



If the utility of the wage is *linear* in the wage, then I am happy to take some random mix of payoffs versus a certain one *only if their expected value is at least as large*. Similarly, if I am risk-*averse*, then I would prefer a *lower* expected value if it compensates me for risk! We're going to assume that the agent is *risk-neutral*, so that payoffs are just *linearly* increasing in the wage. So, the agent's payoff, U, is just:

$$U = \mathbb{E}[w] - c(a)$$

2.4 Timing

Here's how we put everything in order:

1. The principal and agent sign a compensation contract, $w = s + by$.
2. The agent chooses an action, a, *but the principal cannot observe this choice.*
3. Events beyond the agent's control occur, ϵ.
4. Together, the action and the noise determine the output, y.
5. Output is observed by the principal *and* the agent.
6. The agent receives the compensation specified by the contract.

And that's it!

2.5 Bringing it all together—The Solution

How do we 'solve' this problem? Where do we start? We need to understand how two people will behave: the agent and the principal. The agent has no choice over the *terms* of the contract: they can just accept or reject, and then act. The principal has no choice over *actions*, they can only affect the *terms* of the contract. For the principal to understand which terms will be most advantageous to them though, they *have to know what the agent will likely do for a given contract.* So, from the principal's perspective, they need to understand the agent's problem *first*, before turning to their own.

This approach is known as *backwards induction.* We start from the problem that the agent will face for a given contract. Then, we can see how when we change the *contract terms*, what is the likely behavior it will create! Then the principal, with this understanding in mind, creates the contract that will deliver the best returns to them. Then the agent reacts. This is how we solve the model! So, how do we do that exactly? Let's start by thinking about the agent's problem.

2.5.1 The agent's problem

We will solve for a *generic* linear contract. A risk neutral agent wishes to choose the action that maximises the expected value of the payoff, $U = \mathbb{E}[w] - c(a)$. Since the wage is given by a *linear* contract, $w = s + by$, we can write the agent's payoff in the following way:

$$\begin{aligned}U &= \mathbb{E}[w] - c(a)\\ &= \mathbb{E}[s + by] - c(a)\\ &= \mathbb{E}[s] + \mathbb{E}[by] - c(a)\end{aligned}$$

Note that we are assuming that the agent *knows* the details of the contract once they make their choice! See the section above on timing. So, s and b are known.

$$U = s + b\mathbb{E}[y] - c(a)$$

Recall the definition of y? $y = a + \epsilon$. Thus, the agent's *belief* over y is just:

$$\begin{aligned}\mathbb{E}[y] &= \mathbb{E}[a + \epsilon]\\ &= \mathbb{E}[a] + \mathbb{E}[\epsilon]\\ &= a\end{aligned}$$

Here, we can safely assume that the agent *knows* their own action, so $\mathbb{E}[a] = a$, and we know that the expected value of the error, ϵ, is just 0. So, putting it all together, the expected payoff to the agent of a given linear contract, $w = s + by$, is just:

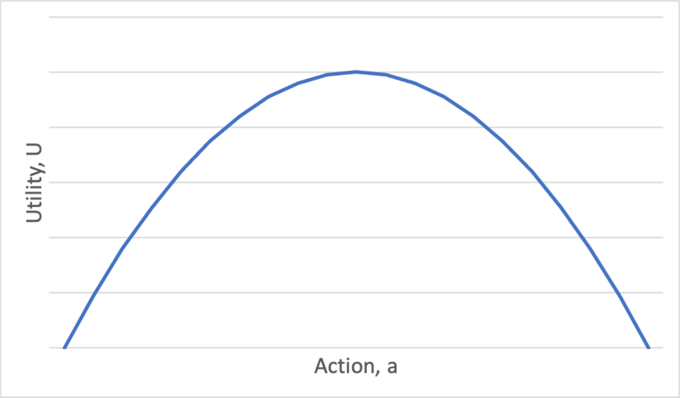
$$U = s + ba - c(a)$$

How do we find their best response? It will be the choice over a that maximises the utility! Suppose that we let the cost function be:

$$c(a) = \frac{1}{2}a^2$$

Then the utility for a given choice of a is shown in the following graph, where I am simply drawing the line, $U = s + ba - \frac{1}{2}a^2$:

How do we find the choice of a that maximises this utility? Well, the best choice is just at the *summit* of the hill created by this total payoff function! How do we find that summit? Well, it's just the value of a where the slope goes from positive to negative, i.e. where $\frac{dU}{da} = 0$. Let's call that 'best' value, a^*. Let's find that now:



$$
\begin{aligned}
U &= s + ba - \frac{1}{2}a^2 \\
\therefore \frac{dU}{da} &= b - a \\
\frac{dU}{da} = 0 &\implies a^* = b
\end{aligned}
$$

So, whatever the principal gives as a bonus rate, *that will determine how much action the agent takes*: $a^* = b$. The surplus, s, *has nothing to do with it.*

2.5.2 Back to the Principal

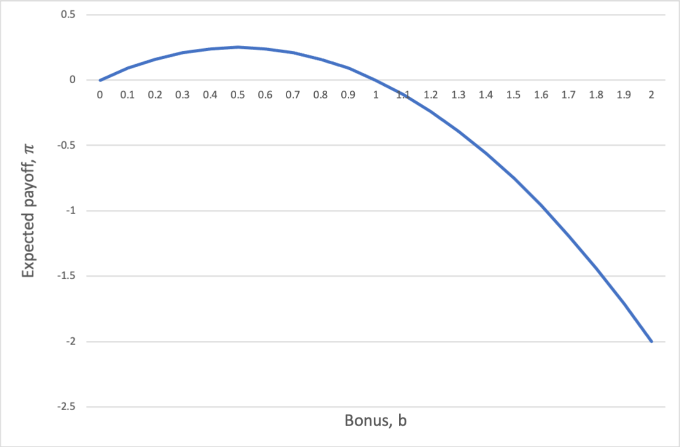
So the principal knows that whatever they offer for b is what they will get for a. How do they maximise their expected payoff? Recall:

$$
\begin{aligned}
\pi &= y - w \\
&= a + \epsilon - (s + by) \\
&= a + \epsilon - s - b(a + \epsilon) \\
\therefore \mathbb{E}[\pi] &= \mathbb{E}[a] + \mathbb{E}[\epsilon] - \mathbb{E}[s] - \mathbb{E}[ba] - \mathbb{E}[b\epsilon] \\
&= \mathbb{E}[a] - s - b\mathbb{E}[a]
\end{aligned}
$$

Here we were able to simplify because the principal knows the details of the contract, $\{s, b\}$, and the expected value of the error is 0. Now we also know that $\mathbb{E}[a] = b$, so:

$$\begin{aligned}\mathbb{E}[\pi] &= \mathbb{E}[a] - s - b\mathbb{E}[a] \\ &= b - s - b^2\end{aligned}$$

What does this look like? Suppose that the salary, s is just 0. Then the expected payoff to the principal is just given by the function $b - b^2$. What does this look like?



Okay, so *again*, there is a special level of b that maximises the payoff. Again, we can find that by finding where the slope goes flat:

$$\begin{aligned}\mathbb{E}[\pi] &= b - s - b^2 \\ \therefore \frac{d\pi}{db} &= 1 - 2b \\ \frac{d\pi}{db} = 0 &\implies b^* = \frac{1}{2}\end{aligned}$$

Note that *again* the optimal choice of b, b^* does not depend on s. We should *always* pick the same value of b. So, if $b = 0.5$, what does that mean for the solution?

2.5.3 Solution

If the value of b is 0.5, then the agent puts in $a = 0.5$. Thus, the expected value of output is 0.5. The utility that the agent receives is:

$$\begin{aligned} U &= w - c(a) \\ &= s + 0.5a - \frac{1}{2}a^2 \\ &= s + 0.5(0.5) - frac12(0.5)^2 \\ &= s + 0.125 \end{aligned}$$

And the expected payoff to the principal is:

$$\begin{aligned} \mathbb{E}[\pi] &= \mathbb{E}[a] - s - b\mathbb{E}[a] \\ &= b - s - b^2 \\ &= 0.5 - s - (0.5)^2 \\ &= 0.25 - s \end{aligned}$$

All that is left is to identify s. Remember, the *principal* writes the contract. What value of s should they choose? In order to think about this, we have to introduce the idea of an *outside option.* There has to be some alternative that the worker could choose, even if that alternative is simply to walk away and receive 0. So if the *outside option* of the agent is to receive 0, then the principal needs to ensure that the contract they offer is *better than 0.* However, note that *only just* better than 0 is still better than 0! So, if the principal chooses $s = -0.124999999999999999999999999999999999999999999$, then the agent will be *better off* taking the contract than walking away, will receive a utility of $U = 0.000000000000000000000000000000000000000000001$, and the principal takes hope almost the entire ‘surplus’ of the interaction.

2.6 What was the point of that?

What did we learn from that exercise? At least three lessons emerged:

- Contract shape matters
  - Linear vs. non-linear contracts have very different consequences for be-

havior.

- Steeper slopes create *stronger* incentives
  - The greater was b, the greater was the 'effort'.
- Steeper slopes are *not always better*
  - There was a sweet spot with respect to effort! $b^* \neq 1$

On top of this, we found a way of expressing *and analysing* abstract concepts like reward, effort, and incentives, in rigorous, formal ways that allowed us to investigate carefully the causal chain.

3 Getting what you pay for...?

Note that we made a strong assumption in the previous simple example; namely that we can *observe* the output of a worker. In practice, this is rarely true. Instead, we usually have access to some combination of performance *indicators* that we hope correlate with worker output. What does this do to our analysis? How much of a problem is this?

Note: This next section is a little tough, but don't worry, you won't be examined on any of this. The mathematics is a little challenging, but if you get the rough gist of what we're doing, and can interpret the end result, then I see that as a big win!

3.1 Setting things up

We will extend the basic setup from before in the following way. First, we will introduce a *performance* measurement, p. This measurement is visible to principal and agent, and will form the basis for our *linear* contracts. This measure will be a *non-deterministic* function of the agent's actions: that is to say, there is a random component to the measure that means we cannot infer actions from observing p.

We will now assume that output, y, is *uncontractable*; that is to say, we can't write contracts based on y, because we cannot observe it. Finally, we will assume that the agent has the choice over *two* types of action, a_1 and a_2, that both are relevant for output, and that *both* contribute towards the performance indicator, *but potentially in ways distinct from their effect on output.* Let's put this all together:

- Technology of production:

$$y = f_1 a_1 + f_2 a_2 + \epsilon, \epsilon \sim N(0,1)$$

- Performance measure:

$$p = g_1 a_1 + g_2 a_2 + \phi, \phi \sim N(0,1)$$

- Set of feasible contracts:

$$w = s + bp$$

- Payoffs:

$$\pi = y - w$$

$$
\begin{aligned}
U &= y - c(a_1, a-2) \\
c(a_1, a_2) &= \frac{1}{2}a_1^2 + \frac{1}{2}a_2^2
\end{aligned}
$$

Note that we have two shocks now, ϵ and ϕ. For now, let's suppose these are *independent shocks*, which is to say that their covariance is 0. They do not move together, essentially. The timing will be essentially the same as before:

1. The principal and agent sign a compensation contract, $w = s + by$.
2. The agent chooses actions, $\{a_1, a_2\}$, *but the principal cannot observe this choice.*
3. Events beyond the agent's control occur, $\{\epsilon, \phi\}$.
4. Together, the actions and the noise terms determine the output, y, and the performance measure, p.
5. Measured performance is observed by the principal *and* the agent.
6. The agent receives the compensation specified by the contract as a function of p.

3.2 The Agent's problem

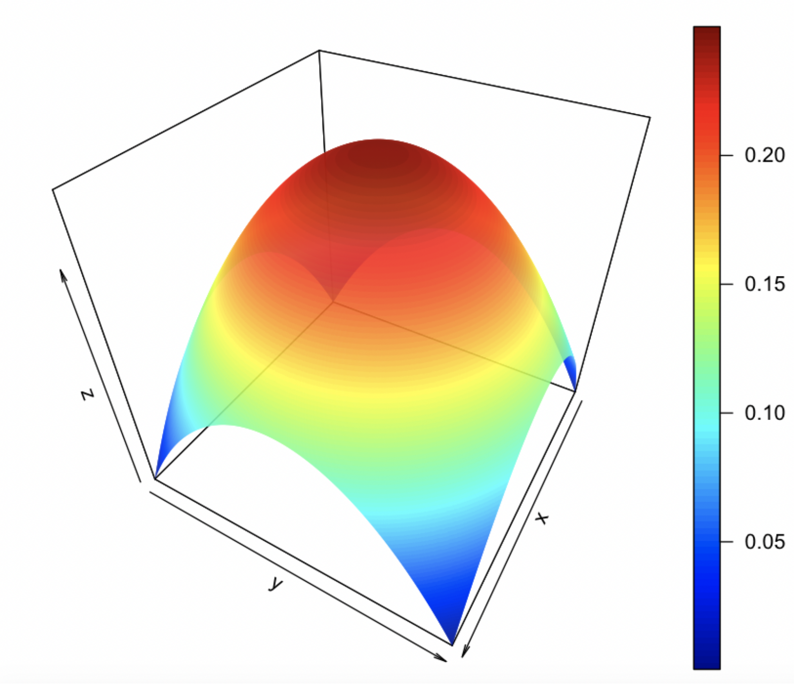
Once again, we solve by *backward induction.* The agent's problem is:

$$
\max_{a_1,a_2} s + b(g_1a_1 + g_2a_2) - \frac{1}{2}a_1^2 - \frac{1}{2}a_2^2
$$

Note that we took a few simplifying steps here! We know that the agent maximises the *expected* payoff, which is a function of p. Note that the expectation of p is just:

$$
\begin{aligned}
p &= g_1a_1 + g_2a_2 + \phi \\
\therefore \mathbb{E}[p] &= \mathbb{E}[g_1a_1 + g_2a_2 + \phi] \\
&= \mathbb{E}[g_1a_1] + \mathbb{E}[g_2a_2] + \mathbb{E}[\phi] \\
&= g_1a_1 + g_2a_2
\end{aligned}
$$

Okay, so for a given $\{s, b\}$ contract, what does the payoff look like?



Funky!! So how do we find that red peak? The combination of $\{a_1, a_2\}$ that makes them the best off they can be *in expectation*? The answer is to take *partial* derivatives with respect to a_1 *and* a_2, and set these two to zero. The thinking is that we're trying to find the point where *both* the slope going towards more a_1 *and* going towards a_2 are zero. Let's think about this a little more carefully.

3.3 Partial Derivatives

What is a partial derivative? Very simply, a *partial* derivative is when we have a function that has *more than one* variable, and we take the derivative with respect to *one* of those variables, *treating the other one as though it were fixed.* So, for example, if we had:

$$f(x, y) = x^2 + y^2 + xy$$

Then the partial derivative *with respect to x* would be:

$$2x + y$$

Notationally, we differentiate[2] between *total* derivatives and *partial* derivatives by calling a total derivative $\frac{dy}{dx}$, and a partial derivative $\frac{\partial y}{\partial x}$.

But what does that mean?? Let's go back to our 3D image above. When I take the partial derivative with respect to x, what am I doing? Well, for a *fixed* y, it allows me to understand how the slope of the function $f(x, y)$ changes *as I increase* x *keeping* y *fixed.* Think of it as taking a slice out of that big hump, where the slice is taken at a given y. We can then think about how that *slice* behaves as we move x around. Cool right?

3.4 Back to the problem!

So why will partial derivatives help? Well, we want to find the top of that hill. So, we want *two* things. For a given a_1, the slope of the function for a_2 is zero, *and* for a given a_2, the slope of the function for a_1 is zero! This is just when the partial derivatives of *both* a_1 *and* a_2 are equal to zero. Let's do that now:

$$
\begin{aligned}
U &= s + b(g_1 a_1 + g_2 a_2) - \frac{1}{2}a_1^2 - \frac{1}{2}a_2^2 \\
\therefore \frac{\partial U}{\partial a_1} &= bg_1 - a_1 \\
\frac{\partial U}{\partial a_2} &= bg_2 - a_2
\end{aligned}
$$

So, setting *both* of these to zero gives us that:

$$a_1^* = bg_1$$

$$a_2^* = bg_2$$

Note the difference from the previous case? Now what matters for the optimal action is the bonus rate *and* the degree to which the action affects the *performance measure.*

[2] GET IT???

3.5 The Principal's Problem

So now the principal knows how the agent will respond to a given contract, we can work backwards and determine how they will behave. Note that expected payoff for the principal is:

$$\begin{aligned}
\pi &= y - w \\
\therefore \mathbb{E}[\pi] &= \mathbb{E}[y] - \mathbb{E}[w] \\
&= \mathbb{E}[f_1 a_1 + f_2 a_2 + \epsilon] - \mathbb{E}[s - b g_1 a_1 + b g_2 a_2 + \phi] \\
&= f_1 \mathbb{E}[a_1] + f_2 \mathbb{E}[a_2] + \mathbb{E}[\epsilon] - s - b g_1 \mathbb{E}[a_1] + b g_2 \mathbb{E}[a_2] + \mathbb{E}[\phi] \\
&= f_1 \mathbb{E}[a_1] + f_2 \mathbb{E}[a_2] - s - b g_1 \mathbb{E}[a_1] - b g_2 \mathbb{E}[a_2] \\
&= \mathbb{E}[a_1](f_1 - b g_1) + \mathbb{E}[a_2](f_2 - b g_2) - s
\end{aligned}$$

And we know what the principal should expect of the agent's actions: we just found that!

$$\mathbb{E}[a_1] = b g_1$$

$$\mathbb{E}[a_2] = b g_2$$

So, putting it all together:

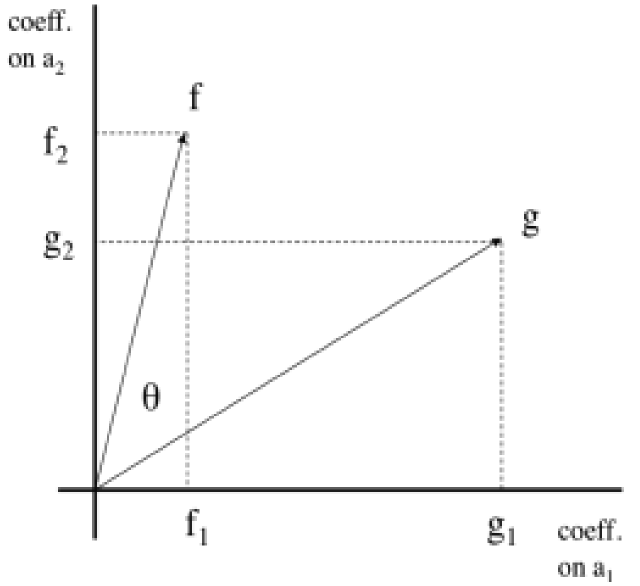
$$\mathbb{E}[\pi] = b g_1 f_1 - b^2 g_1^2 + b g_2 f_2 - b^2 g_2^2 - s$$

How should the principal pick b? Well, they should maximise their payoff, so that means finding the top of the slope!

$$\begin{aligned}
\frac{\partial \pi}{\partial b} &= g_1 f_1 - 2 b g_2 + g_2 f_2 - 2 b g_2^2 \\
\frac{\partial \pi}{\partial b} = 0 &\implies g_1 f_1 - 2 b g_2 + g_2 f_2 - 2 b g_2^2 = 0 \\
\therefore 2 b g_1^2 + 2 b g_2^2 &= g_1 f_1 + g_2 f_2 \\
\implies b(2 g_1^2 + 2 g_2^2) &= g_1 f_1 + g_2 f_2 \\
\therefore b^* &= \frac{g_1 f_1 + g_2 f_2}{2 g_1^2 + 2 g_2^2}
\end{aligned}$$

3.6 What the hell does that mean?

Before we unpack that result a little further, let's think a little bit more about the numbers $\{f_1, f_2\}$ and $\{g_1, g_2\}$. Consider the following figure:



Here we've mapped out two potential sets of these numbers. The θ is a measure of the angle between them. What can we say about these two pairs of numbers? Well, first off, the performance measure rewards a_1 more than it 'should'. There is a *misalignment* between the rewards for a_1 and the impact that a_1 has on output!

Similarly, if we were to imagine *scaling* the values of $\{g_1, g_2\}$, so that the line goes way out of the right of the page, then we might think that there is an *over* incentive to work *in general.* Because there are big returns to the performance measure of performing the action, perhaps the agent will perform the action *more* than is necessary to achieve good output.

These dual concerns are the central problems associated with pay for performance. If the *returns* to actions are *out of proportion*, or *misaligned*, then you can have agents doing things you would really prefer that they didn't! So, have we addressed these problems?

3.7 Back to our solution

Note that we just found that:

$$b^* = \frac{g_1 f_1 + g_2 f_2}{2g_1^2 + 2g_2^2}$$

It turns out that, with a little manipulation that I'm not going to go through here, we can turn this expression into the following:

$$b^* = \frac{1}{2}\frac{\sqrt{f_1^2 + f_2^2}}{\sqrt{g_1^2 + g_2^2}} cos(\theta)$$

Again: **what the hell is that?** Well, let's think about it. First off, note that the $\frac{1}{2}$ term denotes the optimal b *if we could observe output.* The case is *analogous* to the simple case we had before. Now look at the numerator and denominator of the fraction: what do these expressions tell us? Well, if we recall our Pythagoras, then the numerator is simply the *length* of that line we drew on the graph for $\{f_1, f_2\}$ above! Similarly, the denominator is the *length* of the line for $\{g_1, g_2\}$. So the fraction is the *ratio* of the their *scale.*

Suppose that actions were *over-incentivised*, i.e. we actions translate strongly into the performance metric. What our optimal contract result tells us is that the optimal *bonus rate* should be *scaled down* to account for this over-incentive. Similarly, if the returns on the performance measure are weakly related to actions, i.e. the length of the line of $\{g_1, g_2\}$ is short, then we should bump that b up to make sure they do what we want them to do!

Now let's turn to the inconceivable appearance of a cosine expression in a class about incentives. What is that measuring? You guessed it! It captures the *alignment* of the contract. Recall above that θ measured the angle between the line of $\{f_1, f_2\}$ and $\{g_1, g_2\}$? Suppose that there was *perfect* alignment. In that case the angle between the two would be zero. What is $cos(0)$? It's 1! In other words, there is no misalignment, so there is no need to scale the performance measure, b. Now suppose that the performance measure was really misaligned with the output measure. Suppose the angle between them was 90 degrees. What is $cos(90)$? That's right, *it's zero.* So when incentives are really misaligned, we *shouldn't reward the agent on the performance measure at all.* It will just cause them to do things we don't want them to do!

3.8 So, where are the cosines?

Have you ever received a work contract that had a cosine in it? Me neither! So, was this whole thing a stupid waste of mathematical time? No! Remember how we started: we said that models can have either *quantitative* or *qualitative* predictions.

Supposing that the optimal contract would contain a cosine, and anticipating this in the real world, would be a *quantitative* prediction. But the model is a long way of being realistic! There are only two tasks, only one performance measure, and two random, independent shocks that affect output and performance measures. This is not like any environment I've ever worked in. Instead, we should view this as a *qualitative* prediction. Namely, it provides structure to the claim that two things count: the *alignment* of performance measures with actual output, and the *scale* of how much actions increase performance *relative* to output.

3.9 What did we learn from this?

I think there are three main takeaways here.

- Objective performance measures typically cannot be used to create the ideal incentive structure.
- Efficient bonus rates depend on **scale** and **alignment**.
- Performance measures that correlate with performance are not *necessarily* good measures.

We've talked a little about the first two. In the first case, we noted that the optimal contract where we can observe output was $b = \frac{1}{2}$. If $f_1 = g_1$ and $f_2 = g_2$, then this is clearly the same as being able to observe output, and so indeed we back out the ideal incentive structure. However, whenever we *cannot* observe output, we have to scale this incentive structure. This takes us to the second point that the scaling of that incentive should be proportional to *both* the scale *and* the misalignment of the performance measure.

What about that third point? Note that above, we assumed that the two shocks, ϕ and ϵ were *independent.* Suppose we relax that assumption. In fact, we will suppose that they are *the same shock*, i.e. $\epsilon = \phi$. Can we come up with a performance measure that correlates with output, *but is a bad performance measure?* You betcha!

Suppose that $f_1 = 1, f_2 = 0, g_1 = 0, g_2 = 1$. What does this imply for output and the performance measure?

$$y = a_1 + \epsilon$$

$$p = a_2 + \epsilon$$

Note that output and the performance measure *will absolutely be correlated.* They both contain the shock ϵ! But is this a good measure of performance? No! It's dreadful! How should we reward this measure according to our solution? Note that the angle between the line $\{0, 1\}$ and $\{1, 0\}$ is 90 degrees. So, $cos(90) = 0$, which tells us that $b^* = 0$. We should not reward on the basis of this performance measure!

This is a really neat application of how mathematics can give us insight that would be hard to outline in a formal, logical way otherwise. I'm not saying it's impossible, but that is a bloody neat result!!

4 Paying for A and Hoping for B

So, in the real world, how well do we do at rewarding the *right* performance measures? The answer is... not very. Here are a few examples.

4.1 Politics

When parties arrive at their policy platforms, their positions can typically be categorised as either 'official' goals, or 'operative' goals. Official goals are vague and general. Here are some examples from the most recent Democratic Party Manifesto:

- 'We must steel and strengthen our democracy, not distort and debase it. Democrats believe there is nothing to fear from the voices and votes of the American people.'

- 'We must heal our nation's deepest wounds, not fan the flames of hate. Democrats will root out structural and systemic racism in our economy and our society, and reform our criminal justice system from top to bottom, because we believe Black lives matter.'

- ‘We must lead the world in taking on the climate crisis, not deny the science and accelerate the damage.’

None of these goals specify *how* they will be achieved, merely that they are ambitions of the party. They are designed to have high appeal, but with limited responsibilities associated with them. By contrast, *operative* goals specify actual policies. Here are some more examples, again from the Democratic Party Manifesto:

- ‘We will fully implement the Help America Vote Act and require that polling places and elections are accessible for people with disabilities, and work to ensure that returning citizens have their voting rights restored upon release from jail or prison without the additional hurdle of having to pay fines and fees in order to vote.’

- ‘Democrats are committed to restoring the full power of the Voting Rights Act and ensuring every citizen can access the ballot box. We will enforce and strengthen the Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act, and will end racial and religious profiling in law enforcement.’

- ‘We will rejoin the Paris Climate Agreement and, on day one, seek higher ambition from nations around the world, putting the United States back in the position of global leadership where we belong. We will restore protections for irreplaceable public lands and waters, from Bears Ears National Monument to the Arctic National Wildlife Refuge.’

These operative goals are higher in quality, but by being specific, are more susceptible to push back. Which do you feel are more common?

It is a *classic* criticism of politicians that they are vague, and refuse to say what it is they will do. In effect, we tend to hold the view that politicians spend too much time outlining *official* goals, and not enough on *operative* goals. Trawling through the Democrat manifesto I can tell you that the distribution is probably around 90/10, and I strongly suspect this would be true for the Republican party, the Labour and Conservative party of the UK, En Marche in France, or the LDP in Japan.

So why do they do it? I think there are at least two reasons. The first is that being vague prevents you from closing doors with future collaborators. This is a *good thing.* If you have the freedom to move around the policy environment, you are more likely to find allies, and hence success, in politics. Of course, this only works well for the citizenry if the politician is a broadly decent person!

Another much less advantageous reason for this tendency *is that it is rewarded at the poll booth.* By only making vague, and essentially pleasant statements about abstract values, *politicians avoid the problem of turning people off.* I think an example of a politician who has been very explicit about their policy platforms, *and suffered for it*, is Elizabeth Warren.

So, what does this boil down to? We *hope* that these abstract claims will translate into action, *and maybe we'll be right*, but we certainly do not *explicitly* reward politicians for stating operative policies. We hope that these are correlated, but, as we just saw, correlation of a performance metric and an output can be misleading indeed!

4.2 Universities

What is the role of a professor? Many people believe that a professor's role is to teach. If you were to ask 100 professors what their role was, I wouldn't be surprised if *none* of them believed that teaching was their principal role.

Why? Because professors *are not rewarded for teaching.* Advancement and prestige in academia is a function *exclusively* of your research[3]. It actually goes beyond that: if you are not producing good research in your early career, *you will be fired*! By contrast, teaching factors *not at all* into your career trajectory or job security. Even very, very poor teachers will face almost no constraint on their career path, so long as they show up and attend required office hours/lectures.

So, what do we do? We *hope* that academics will be good teachers. But we put no incentive structure in place to encourage that! We assume that research ability and teaching ability are correlated, which even if they were, would not mean that good researchers would be good teachers! As we just saw, correlation does not imply alignment in incentives.

The incentive problems do not stop there. One of the biggest frustrations as an *instructor* is the emphasis that students place on grades. We think of these problem sets and exams simply as ways to discipline your understanding, and to build knowledge. We hope that you will see that grades don't really matter for understanding the course, so you shouldn't focus on them too much and forget to

[3] With the exception of Liberal Arts Colleges and, to a lesser extent, Business and Law Schools

actually interact with the material.

But this is stupid too! Of course students care about grades!! Grades are important for various career paths that many students wish to take, be it grad school, parental respect, careers, whatever! We *hope* that you will come to class in order to further your understanding, *but you are not incentivised to do that.* So, trust me when I say, the frustrations go both ways!

4.3 The Corporate world

Two classic examples illustrate how misaligned incentives can have ruinous consequences. The first is Enron, and the second is Bernie Madoff.

4.3.1 Enron

How many of you have heard of Enron? In some ways it now occupies a similar place in the lexicon as 'Watergate' —an iconic scandal that has become a byword for corporate mismanagement. So, what happened?

In the 90s, Enron was *the* company. It was an energy manufacturer, founded in 1985, and claimed revenues of $101bn in the year 2000. Fortune named Enron 'America's Most Innovative Company' for 6 consecutive years.

What made Enron so astonishing was its ability to deliver energy at very low rates while still making steady profits. It turned out that the way they did this was by hiding their losses in so-called 'Special-Purpose-Vehicles', essentially just shell companies that didn't report into Enron's bottom line. This was obviously not sustainable, and in 2001, everything fell apart. The scandal was revealed, the company went bust, and the CEO, Jeff Skilling, was sentenced to 24 years in federal prison.

Where did everything go wrong? Jeff Skilling, who by all accounts is not a very nice man, had a responsibility to maximise the wealth of his shareholders. His incentive to do this was the stock price. The problem was that Jeff Skilling knew that if he blew up the stock price in the short-run, *he could make himself very rich.* Never mind that the company would blow in a few years, he'd be long gone! Of course, this was a risky strategy, and it didn't pan out how I imagine he'd hoped. But what this shows is that, providing *rough* incentive structures that *correlate* with the outcome you want does not preclude a guy like Jeff Skilling taking a hold of your company and rinsing it into the ground.

In the wake of the Enron scandal, the government passed the Sarbanes-Oxley act, a piece of legislation essentially designed to prevent the kind of accounting manipulations that allowed the Enron scandal to happen. In effect, this tightened the incentive scheme.

4.4 Bernie Madoff

Bernie Madoff ran the largest Ponzi scheme in history. Interestingly, he also invented 'payment for order flow', which perhaps explains why the government is so suspicious of its legitimacy! A Ponzi scheme is when I claim to have a great business idea that will deliver mega returns, but when you invest in me, I just use the money to pay the person I told yesterday that I would give them mega returns, while pocketing some for myself.

Ponzi schemes are by construction, doomed to fail. You cannot keep borrowing from people and moving money around and hope that it will last forever. I guess Bernie Madoff thought he would die before it ended, or that some other investment would pull off big time, but either way by the time everything came crashing down, there was a $65 **billion** hole in the accounts. Huge numbers of people lost massive amounts of money, and Bernie Madoff died in prison.

What happened? Bernie Madoff felt that the rewards from pumping his Ponzi scheme were worth the risk. But it wasn't even a case of risk! If something will definitely happen, *then it isn't a risk, it is an outcome.* Madoff acted in his *own* interests, *and was rewarded* (for a time) *for doing so!*

5 Summary

Getting people to act in your best interests is hard. The theory tells us that we should reward performance in accordance with its scale and alignment with the outcomes we want. Doing this badly can result in very bad outcomes. Indeed, we are still learning how to do this better.

The major innovation of paying managers in stock options was intended to overcome these agency problems. The thinking is that this brings the performance measure more in line with the outcome we want: more wealth for shareholders. But as we've seen with Enron and Bernie Madoff, if done poorly, this can still fail to prevent major agency failures.

One critical problem is that the *time horizons* of managers is typically different from that of shareholders. The average tenure of an American CEO *is just 7 years*. Many shareholders, like Buffett for example, hold stocks for decades, even passing them on generationally. Aligning the incentives of *short-termist* managers with the interests of shareholders is a very tricky problem! Indeed, it is an area that I do research in and find particularly interesting.