MECS 570-1: Organizational Economics I

Michael Powell

Boundaries of the Firm, III: Property Rights; Foundations

1 Property Rights (Updated: Feb 2 2017)

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, the 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 passively. We will talk about why this is an important distinction to make when we talk about the influence-cost model in the next section.

Description 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 and e_D 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, $s \in S = S_C \cup S_{NC}$, with $S_C \cap S_{NC} = \emptyset$ and $\Pr[s \in S_{NC}] = \mu$. 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 $G = \{UI, DI, NI, RNI\}$. 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 $w \in W = \{w : E_U \times E_D \times S_C \to \mathbb{R}\}$ to D. The contract specifies a transfer $w(e_U, e_D, s)$ to be paid from D to U if they trade good $s \in S_C$. If the players want to trade a good $s \in S_{NC}$, 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 $g \in G$ and a contract $w \in W$. 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, $s \in S$ is realized.

4. If $s \in S_C$, D buys good s at price specified by w. If $s \in S_{NC}$, 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 $e_U^* : G \times W \to \mathbb{R}_+$ and $e_D^* : G \times W \to \mathbb{R}_+$, and a pair of offer rules $w_U^* : E_D \times E_U \times S_{NC} \to \mathbb{R}$ and $w_D^* : E_D \times E_U \times S_{NC} \to \mathbb{R}$ 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 $s \in S_{NC}$; 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 As always, 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. We will return to this issue later. 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

$$\begin{aligned} F_U &= f_{UU} + f_{DU} \\ F_D &= f_{UD} + f_{DD}. \end{aligned}$$

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 $s \in S_{NC}$, 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):

$$\begin{aligned}
\pi_U &= V_U(e_U, e_D) + w_U(e_U, e_D) = V_U(e_U, e_D) + V_D(e_U, e_D) - v_D(e_U, e_D; g) \\
\pi_D &= V_D(e_U, e_D) - w_U(e_U, e_D) = v_D(e_U, e_D; g).
\end{aligned}$$

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):

$$\begin{aligned}
\pi_U &= V_U(e_U, e_D) + w_D(e_U, e_D) = v_U(e_U, e_D; g) \\
\pi_D &= V_D(e_U, e_D) - w_D(e_U, e_D) = V_U(e_U, e_D) + V_D(e_U, e_D) - v_U(e_U, e_D; g).
\end{aligned}$$

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

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

Substituting in the functional forms we assumed above, these problems become:

$$\max_{\hat{e}_U} \frac{1}{2}\left(F_U \hat{e}_U + F_D e_D\right) + \frac{1}{2}\left(\left(h^g_{UU} - h^g_{DU}\right)\hat{e}_U + \left(h^g_{UD} - h^g_{DD}\right) e_D\right) - \frac{1}{2}\hat{e}_U^2$$

and

$$\max_{\hat{e}_D} \frac{1}{2}\left(F_U e_U + F_D \hat{e}_D\right) + \frac{1}{2}\left(\left(h^g_{DU} - h^g_{UU}\right) e_U + \left(h^g_{DD} - h^g_{UD}\right) \hat{e}_D\right) - \frac{1}{2}\hat{e}_D^2.$$

These are well-behaved objective functions, and in each one, there are no interactions between the managers' investments, so each manager has a dominant strategy, which we can solve for by taking first-order conditions:

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

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^g_{UU}, 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 $h^g_{UU} > 0$. 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 $h^g_{DU} > 0$.

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:

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

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

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

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.

PROPOSITION (Hart 1995). 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 Holmstrom points out in his 1999 JLEO article, "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." Holmstrom 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 Holmstrom 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 GHM setups is that the former holds that maladaptation in the contract execution

interval is the principal source of inefficiency, whereas GHM 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 Program 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) show 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. We will discuss it in more detail later.

2 Foundations of Incomplete Contracts (Updated: Feb 2, 2017)

The Property Rights Theory we discussed in the previous set of notes shows that 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. In this note, I will discuss some of the commonly given reasons for why contracts might be incomplete, and in particular, I will focus on whether it makes sense to apply these reasons as justification for incomplete contracts in the Property Rights Theory.

Contracts may be incomplete for one of three reasons. First, parties might have private information. This is the typical reason given for why, in our discussion of the risk–incentives trade-off in moral hazard models, contracts could only depend on output 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.

Another reason often given 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, IER 1985; Bajari and Tadelis, RAND 2001; and Battigalli and Maggi, AER 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 in this note 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. In this note, I will provide an example of the types of mechanisms that have proposed in the literature, and I will point out a couple of recent criticisms of these mechanisms.

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.

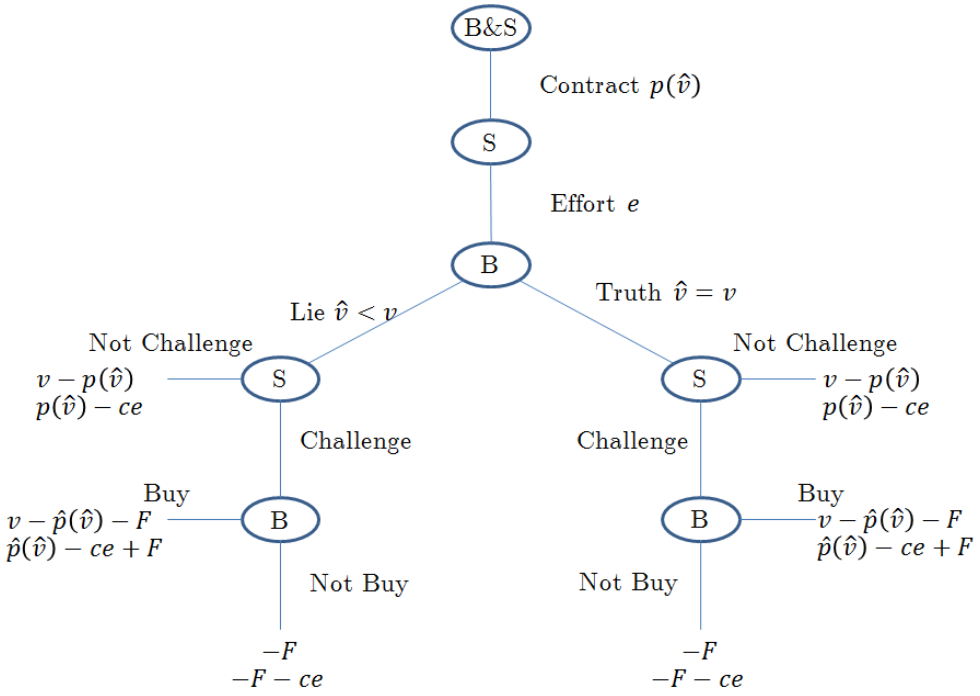
5. S can challenge B's announcement or not. The challenge decision is commonly observed and verifiable by a third party. If S does not challenge the announcement, trade occurs at price $p(\hat{v})$. Otherwise, play proceeds to the next stage.

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.



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 (Econometrica, 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 (RESTUD, 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, EER, 2002)

Recent theoretical work by Aghion, Fudenberg, Holden, Kunimoto, and Tercieux (QJE, 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, Fehr, Holden, and Wilkening (2016) take 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, Powell, and Wilkening (2015) take 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, Powell, and Wilkening are 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, Powell, and Wilkening provide suggestive evidence that preferences for reciprocity are responsible for their finding that B's often reject counter offers.

The findings of Aghion, Fehr, Holden and Wilkening and of Fehr, Powell, and Wilkening 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. Framing the implementation problem in the presence of these types of "behavioral" considerations and proving possibility or impossibility results could potentially be a fruitful direction for the implementation literature to proceed.