Discussion Paper No. 1223

A REPUTATIONAL MODEL OF AUTHORITY

by

Nabil I. AL-NAJJAR*

May 1998
Revised: September 1998

Math Center website:
http://www.kellogg.nwu.edu/research/math

(*) Department of Managerial Economics and Decision Sciences, J.L. Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL, 60208. [al-najjar@nwu.edu]

I am indebted to Greg Greiff, Ramon Casdesus-Masanell, Jim Dana and Daniel Diermeier. I also thank Tim Feddersen, Marc Ventresca, and Brian Uzzi for their comments. This is a revised version of an earlier paper titled "The public-good nature of Reputation building," which benefited from the comments of seminar participants at Northwestern, Rochester, and Windsor. All remaining errors are my own.

Abstract:

The paper provides a model where authority relationships are founded on reputation. The viability of authority is the result of subordinates' free-riding on each other challenges, reducing the frequency of challenges, and making reputation worth defending. The party with authority secures subordinates' compliance through the payment of rents to influence the extent of their failure to act collectively and exacerbate the free-rider problem they face. The model provides a framework to explain how the magnitude and form of these rents depend on the primitives of the environment and on the authority's design of its reputation. Applications to efficiency wages, dictatorships, and the notion of legitimacy are considered.

1. INTRODUCTION

Many group interactions are characterized by one party's authority to direct or control the actions of others. Authority over subordinates and the discretionary powers it conveys are pervasive. It is, as noted by Simon (1951), a key characteristic of employment contracts.[1] In fact, a common view among economists is that authority over subordinates is what distinguishes interactions within firms from market exchange.[2]

Authority is also a central theme in other social sciences, as a major force in aligning divergent individual interests into coherent, organized structures like firms, unions, states and armies. The rise or decline of these organizations is often attributed to the success or failure of their internal authority structures.[3] Perhaps nowhere is authority more evident than in dictatorships, a political form prominent throughout history, where a small group of individuals mobilizes a much larger population into action (or inaction) to serve its interests. In all these examples, who possesses authority is a major determinant of how interactions are structured, resources are used and wealth is distributed.

Authority over subordinates, the power to direct them in ways often contrary to their immediate self-interest, would be of little value if its exercise is met with frequent and active challenges. As noted by Arrow (1974), the prohibitive cost of their frequent application means that sanctions "cannot be the sole or even the major basis for acceptance of authority":

> "Control mechanisms are, after all, costly. If the obedience to authority were solely due to potential control, the control apparatus would be so expensive in terms of resources used as to offset the advantages of authority."(p. 72)

Even when delegated from a more senior authority, or supported by legal rights, such as binding

---

[1] See also Arrow (1974). This view is also found in sociological studies of employment relationships. Halaby (1986), who elaborates on the central role of workplace authority, quotes Max Weber: "the hiring of any kind of service for wage or salary ... involves the subjection of the worker under a form of domination." (p. 636).

[2] This is a central theme of the transaction cost tradition, following Williamson (1985). Discretion and authority are central to the influence cost view of Milgrom (1988) and Milgrom and Roberts (1990), the property rights approach of Grossman and Hart (1986), and the firm as a carrier of reputation, as in Kreps (1990).

[3] See, for instance, Arrow (1974, p. 65), or Coleman (1990, ch. 6).

contracts or ownership rights, a viable authority must, at least partly, rest on subordinates' consent. An employer's authority, for instance, would not be viable if he had to spend most of his time and resources implementing disciplinary measures against employees. Indeed, the hallmark of successful leadership is structuring its authority so it is rarely, if ever, challenged. By contrast, the collapse of authority is often linked to a loss of will to face the mounting cost of dealing with an increased frequency of challenges.

That authority relationships ultimately rest on subordinates' consent, on their belief in its power rather than the overt and frequent use of such power, are important aspects often considered the defining features of such relationships. Subordinates' consent does not mean that they like compliance per se, but that they view it as optimal given their expectations about the authority's behavior. Noting the insufficiency of control over instruments of power as basis of authority, Arrow (1974, p. 72) concludes that "authority is viable to the extent that it is the focus of convergent expectations. An individual obeys authority because he expects that others will obey it."

But compliance based on "convergent expectations" poses a puzzle: individuals with conflicting interests would each like to establish authority over others, but only one can ultimately prevail. What makes one party more effective in building expectations about his resolve to defend his authority? What makes expectations converge in this party's favor?

This paper provides a model to address these issues. We consider a stylized environment where a single central player, called the *Authority*, interacts repeatedly with N identical long-lived agents. The Authority's power to direct agents is founded on its reputation, which we interpret as expectations about how challenges and compliance will be dealt with. Since subordinates have their own reputational concerns, they may be tempted to challenge—either to undermine the Authority's reputation, or to establish their own. The outcome of such contest is decided not just by the players' resources, but also by their expectations about their opponents' behavior. A reputational model of authority should therefore account for why one party has greater ability to carry reputation.

In this paper, the Authority's main advantage is that it interacts with every agent, while agents interact only with the Authority. This enables the Authority to control the frequency of challenges by making the collapse of its reputation a public good for the agents. Each agent then compares the private cost of a challenge with its incremental impact on making the Authority blink. Agents

will therefore free-ride on each others' challenges, reducing the frequency of challenges, and making reputation potentially worth defending.

Free-riding, however, is not sufficient to explain why the Authority's reputation dominates. The reason is the considerable arbitrariness that often characterizes expectation formation in a dynamic strategic setting. Maintaining authority seems to require just the opposite of this arbitrariness, namely subordinates' unquestioning submission and their inability to see viable alternatives to compliance. To achieve such degree of conviction, the Authority's reputation should be robust to the uncertain, often arbitrary, nature of the way expectations are formed. We model this by restricting the Authority to reputations that are worth defending against any agents' behavior that is rational relative to some conjecture they have about the future.

When a reputation is robust in this sense, a compelling logic determines players' expectations, leading to the game's unique equilibrium: the public-good nature of reputation implies that a large fraction of agents free-ride, so the Authority defends its reputation regardless of what the remaining agents do. But if all agents are convinced the Authority will defend its reputation, they all comply and the actual frequency of challenges drops to zero.[4] Subordinates conclude that compliance is 'the obvious thing to do', without the need for precise, prior knowledge of others players' actions and expectations.

Thus, the model uses the indeterminacy and arbitrariness of expectations as a source of restrictions on allowable reputations, hence on the authority structures they can support. To put this in perspective, it is useful to contrast the analysis with contracting in static agency settings. Such contracts require that agents be left just indifferent between taking the desired action and some other inferior action.[5] The implausibility of this prediction, especially in long-term relationships, is hard to overlook: agents left just indifferent between challenging and complying have nothing to lose in the short-run, and might have something to gain in the long-run if their challenges help overturn the Authority. But then the Authority is exposed to the risk of being challenged frequently, escalating

---

[4] This logic departs from the reputation literature pioneered by Milgrom and Roberts (1982) and Kreps and Wilson (1982) in several important ways, discussed in detail in Section 4.5.

[5] See, *e.g.*, Grossman and Hart (1982). Technically, the question is whether the incentive constraint binds. A more detailed discussion is included in Sections 3 and 4.

the cost of defending its reputation, and undermining its viability.

By contrast, an authority built on a robust reputation reflects the idea that subordinates view compliance as 'the obvious thing to do.' Such effective focusing of expectations comes at a price, however: to ensure it, the Authority must hedge against subordinates' incentives to challenge, even though these challenges never actually occur. In particular, since no free-riding has to occur if subordinates are indifferent between challenging and complying, robust reputations always require that agents receive rents for compliance. The Authority's problem is then to design its reputation to balance the need to exacerbate subordinates' collective action problem, with the desire to minimize the cost of implementing the necessary rewards and sanctions. The result is a simple framework that links the primitives of the environment to the Authority's reputation choice, and to the magnitude and form of the rents needed to secure compliance.

These rents provide a more realistic description of authority relationships, and an insight into the factors responsible for the immense variety of rewards and sanctions used to support them. For instance, extreme sanctions applied to "make an example of someone" are widely practiced in settings where agents have no viable outside options. These include some of the more (morally) perverse examples of authority relationships, such as dictatorial regimes, Mafia families and insurgency movements.[6] Sanctions involving seemingly gratuitous violence and cruelty are widely documented and, given their cost, would be difficult to explain in the absence of reputational concerns. It is unlikely that agents who live under such conditions are left just indifferent between compliance and challenge. Rather, extreme measures are effective means of exacerbating the free-rider problem among potential challengers. This makes it possible for a small group of individuals to control much larger populations using surprisingly small expenditures of resources, and despite the fact that the dominant group could not withstand a collective act of the subordinates.

When agents have viable outside options, as in employment contexts, the same logic implies that authority is maintained through rewards that make agents strictly prefer compliance to challenging. To an outside observer, this might appear to be a form of gift exchange or efficiency wage: employees

---

[6] See Gambetta (1993, *e.g.*, pp. 43-46, and p. 245) for the role of reputation in the operation of Mafia families, viewed as economic enterprises in the business of protection and intimidation. On insurgents' methods of extracting compliance and the role of reputational concerns, see Leites and Wolf (1970), pp. 99-112.

strictly prefer to work above the contractually specified minimum level, and the employer rewards employees with rents for complying. The efficiency effects of gift exchange and higher wages are well-known.[7] What the model offers is a new and distinct rationale for their use as means of raising subordinates' cost of challenging the employer's authority.

This last point deserves some elaboration: agents' collective interest in undermining the Authority's reputation does not preclude welfare enhancing authority relationships. In the context of employment, compliance should be interpreted broadly to include such things as refraining from disruptive activities, bargaining, haggling, renegotiation or generally encroaching upon an employer's authority. These are examples of influence activities[8] that are socially wasteful and whose proliferation would be inconsistent with efficiently operating organizations. A credible authority can create efficiency gains that improve individuals' ex ante welfare (before joining the organization) by countering their ex post incentive, as subordinates, to challenge it by engaging in such wasteful activity. Stated differently, one solution to the problem of minimizing influence cost may be to create authority structures that succeed by exploiting subordinates' inability to act collectively. Individuals' ex ante welfare improves if some of the efficiency gains achieved in this manner are distributed through higher wages, better career opportunities and working conditions.

This argument may clarify our assumption of a centralized, star-shaped structure of interaction. This structure, taken here as exogenous, provides a tractable, stylized benchmark that captures key features of authority in important examples like employment relationships, franchising, and centralized political systems. It is natural to ask what forces give rise to such special structure. A commonly made argument[9] is that centralization economizes on information and decision making costs. The argument of the last paragraph provides a separate mechanism to account for the value of centralization through its role in strengthening authority structures.

The model also clarifies why size might matter in establishing and maintaining authority. Size matters not because maintaining authority against a larger number of agents is intrinsically more

---

[7] See the surveys by Akerlof (1984), Weiss (1990), and Bewley (1998) for the literature on gift exchange and efficiency wages. More references may be found in Section 5.1.

[8] See Milgrom (1988) and Milgrom and Roberts (1990).

[9] For instance, Arrow (1974, p. 68).

valuable, but because it tends to exacerbate the collective action problem they face. The model therefore predicts that discretionary powers to direct resources, resolve disputes and meet new contingencies, will tend to be vested in the larger party. This is broadly consistent with the stylized fact that such powers tend to be vested in employers rather than employees, franchisors rather than franchisees, and so on.[10] Size differences can also account for the asymmetry between the challengers and the Authority's incentives to invest in reputation: the Authority fully internalizes the benefits from challenging agents, so free-riding works in one direction only. The model therefore suggests an explanation for why, in an environment in which incompatible authorities might conceivably arise, the authority of the larger player will tend to prevail.

The free-riding argument implies that the relevant measure of size is the number of *independent* decision-making agents. Mechanisms facilitating collusion, such as organizations and other collective bodies, offset the free-riding problem, thus limiting the scope of authority and/or increasing the rents needed to secure compliance. The model therefore offers a simple rationale for the time-honored precept of "divide-and-conquer," or the widely documented practice of atomizing individuals and dissolving independent institutions in totalitarian regimes. Dividing opponents need not have any intrinsic value to the Authority. Rather, its value stems from increasing the effective number of independent agents, and the consequent worsening of their free-riding problem.

---

[10] Hadfield (1990) reports on the considerable authority of franchisors in franchising relationships. Franchisees may be interpreted as subordinates in the model of this paper. Of course, franchisees retain considerable discretionary power; see the discussion of delegation below.

Related Issues and Literature

Despite their pervasiveness in economic interactions, and despite Simon's (1951) early emphasis of their role in employment contracts, there has been relatively little formal analysis of authority relationships in competitive models of exchange, or in models of contracting under asymmetric information. One reason is that control over subordinates in these models is provided through exogenously given, formal mechanisms, like the rules governing exchange in a market, or complete contracts enforced through a third party. As noted by Arrow (1974), the need for authority arises precisely when such mechanisms fail to satisfactorily perform allocative tasks in organizations.

The major departure from this tradition is the transaction cost literature (*e.g.*, Williamson (1987)). This literature emphasizes interactions where well-defined property rights and complete contracts which anticipate all possible contingency are unavailable, thus forcing the issue of what enables one party to direct the actions of others. Prominent in this tradition is Grossman and Hart's (1986) model of ownership rights over the use of physical assets as a source of authority.[10] Asset ownership in their model does not convey direct control over subordinates, whose right to control their own actions is inalienable. Rather, ownership gives indirect authority through its effect on the set of options available once new contingencies unfold. Closer to the spirit of this paper is Kreps' (1990) model in which the rights of control and the discretionary powers they entail are founded on reputation. While closer in spirit, there are too many differences for a meaningful comparison. I further comment on Kreps' paper in Section 6.

Subordinates in many organizations, such as in franchising and employment settings, retain considerable discretionary power in directing day-to-day operation. Different organizational structures may be thought of as reflecting alternative patterns of internal delegation of authority. A recent contribution that examines these issues is Aghion and Tirole (1997). They use an incomplete contracting setting with information asymmetries to evaluate the costs and benefits of delegation. Our focus in this paper is on the source of authority. The issue of delegation is fundamental to understanding how authority is structured, and is obviously complementary to studies of its source.

---

[10] This idea is also part of the Marxist paradigm that capital ownership is the source of domination of the working class. See Selznick (1969, pp. 65-67) for the role of property rights in Marxist ideology.

The concept of authority is often associated with that of discretion, namely the right to direct agents' actions as new contingencies unfold. The employer's authority in Simon (1951), for instance, enables him to pick, ex post, an action at his own discretion. The present model is consistent with the exercise of discretion if we interpret compliance as "obeying the Authority's directives", where these directives may be state-contingent. While discretion is not inconsistent with the model, our focus will be on achieving compliance rather than investigating the nature of discretionary power conveyed by authority. It should also be noted that, while closely linked, discretion and authority are distinct concepts which can be usefully treated separately.[11]

By contrast with the economics tradition, questions about the source of authority, who acquires it and how it is maintained have been central topics of research in other social sciences. The political science and sociology literatures examine authority in relation to a host of other concepts, such as power, leadership, discretion, and legitimacy. Uniform, generally accepted definitions of these concepts and the links between them are, unfortunately, lacking. Dennis Wrong's (1979) classic essay, *Power*, provides a comprehensive account of the literature on authority and power. Authority relationships are also discussed extensively in Coleman (1990), as well as in the literature on organizations and employment relationships (*e.g.*, Selznick (1969), Halaby (1986) and Scott (1998)).

Sociologists draw a distinction between authority and power. While both involve the ability to direct subordinates to take actions which may be contrary to their immediate self-interest, one of the defining features of authority relationships is that subordinates' compliance is mostly voluntary or consensual. Wrong (1979, p. 37), who asserts this view, quotes Weber: "every genuine form of domination implies a minimum of voluntary compliance, that is, an interest ... in obedience." See also Coleman (1990, p. 71). One contribution of the present paper is to provide a mechanism explaining why subordinates form expectations under which voluntary compliance is optimal, and what price has to be paid to secure such compliance.

---

[11] Discretion may be achieved by means other than authority, *e.g.*, through an enforceable contract stating that the agent will do what his superior tells him to do. Here, the superior has discretion supported not by direct authority over the subordinate, but by whatever source of power used to enforce the contract. An authority relationship may also exist without significant scope for discretion: what a dictator wants from his subjects may be as simple as: do not revolt. In this case, establishing authority need not involve discretion.

Finally, sociology and political science studies often invoke reputation and free-riding in informal arguments about authority. For instance, in discussing authority within revolutionary movements, Coleman (1990, p. 482) criticizes the view of such organizations as monolithic entities because, from the perspective of the insurgents, "revolution is a public good, and, like any public good, it gives rise to the free-rider problem." See also Wintrobe's (1998, p. 106) discussion of dictatorships. This paper's contribution relative to these informal insights is the idea that subordinates' failure to act collectively is the result of an artificial public good problem, one created by the Authority for the purpose of manipulating their incentives. This makes explicit the Authority's incentive to design its reputation to exacerbate the subordinates' free-riding problem. The closer connection between reputation and free-riding makes it possible to address questions about how the primitives shape the scope of authority, and the form and magnitude of the rents needed to support it.

2. THE MODEL

2.1. Single-Period Interaction

We consider a setting of repeated interaction between an Authority and N identical agents. A single-period interaction is modeled as an agency problem (Grossman and Hart (1983)), with a few significant departures, explained below.

The agent chooses either to comply or challenge, denoted a^* and a^0 respectively. Here, we interpret a^0 as those aspects of performance that are either explicitly contracted, or can be guaranteed through some other direct means, such as binding contracts, use of force, coercion, or compliance with a more senior authority. The difference between a^* and a^0 thus reflects those aspects of performance that cannot be guaranteed through such direct means. For instance, compliance may take the form of following an employer's discretionary directives which are not part of the formal employment contract, while challenges correspond to behavior ranging from overt defiance down to more subtle forms of resistance.[12]

Each action $a \in \{a^*, a^0\}$ generates a signal $b \in \{b_1, \ldots, b_K\}$ with probability $\pi_a(b) > 0$. The authority's action is a contingent scheme, later interpreted as its 'reputation', of the form $\tilde{x} = (x_1, \ldots, x_K)$ under which an 'ex post' action $x_k \in I\!R$ is taken by the Authority, and represents a reward or a sanction in response to signal b_k.

To simplify the analysis, assume the N agents are identical and that their signals and actions are independent. Agent n's payoff depends on the action he takes, a_n, and on the contingent reward/sanction $x = x(b_n)$ in the form: $u(a_n) + g(x(b_n))$. The Authority's payoff with agent n depends on his signal b_n and the corresponding x in the form: $v(b_n) - c(x(b_n))$. The Authority maximizes the average expected payoff with the N agents: $\frac{1}{N}\sum_n (v - c)$.

There is a basic conflict of interest over the agent's choice of action: Each agent generates a surplus $S > 0$ which accrues to the Authority if he complies but that he retains if he challenges. Thus,

---

[12] For example, a^0 may represent showing up to work on time, yet being a trouble-maker, while a^* represents showing up on time and being a cooperative member of the group.

compliance generates expected benefits $Ev(b \mid a^*) = S$ and $u(a^*) = 0$, while challenging generates $Ev(b \mid a^0) = 0$ and $u(a^0) = S$.

One departure from the standard agency model is the interpretation of the ex post action x_k. Depending on the context, x_k may correspond to a direct monetary reward or penalty representing how output is shared, as in agency contracts. It may, however, equally represent social prestige or political favor to reward compliance, a sanction imposed for violating a law, a reprimand or dismissal of an employee, or fighting an attempt to enter the incumbent's market. Thus, indexing ex post actions on the real line should be viewed as a convenient representation, rather than necessarily portraying a one dimensional variable.

With this motivation, we distinguish between rewards and sanctions by assuming that g is strictly increasing, that c is convex with unique minimum at 0, and that $g(0) = c(0) = 0$. We interpret $x > 0$ as a reward and $x < 0$ as a sanction: agents prefer higher rewards and dislike harsher punishments while the Authority dislikes both rewarding and punishing the agent. Thus, rewards are transfers to the agent, while punishments are costly to both parties. This contrasts with the standard agency model, which usually does not include sanctions in the sense described here.

2.2. Static Benchmark

A useful benchmark is a static setting in which the Authority induces compliance by offering a binding contract.

To examine this possibility, note that the Authority always retains the option of offering the trivial scheme $\tilde{x}^0$ in which $x(b_k) = 0$ for all b_k. For any non-trivial scheme $\tilde{x}^* \neq \tilde{x}^0$, we impose the incentive and participation constraints:

$$Eg(\tilde{x}^* | a^*) \geq S + Eg(\tilde{x}^* | a^0) \qquad (IC)$$

$$Eg(\tilde{x}^* | a^*) \geq u_0, \qquad (PC)$$

where u_0 is the agent's reservation value.

Consider now the constrained maximization problem: select $\tilde{x}$ to solve

$$\max_{\substack{\tilde{x}\in\{\tilde{x}^0,\tilde{x}^*\}\\ a\in\{a^0,a^*\}}} \frac{1}{N}\sum_n E[v(b_n) - c(x(b_n)) \mid a], \quad \text{subject to } PC, IC.$$

I refer to the solution to this problem as the static agency solution. Since rewards and sanctions are costly, the Authority chooses $\tilde{x}^* \neq \tilde{x}^0$ only if compliance under $\tilde{x}^*$ generates a net benefit: $S - Ec(\tilde{x}^*|a^*) \geq 0$, an assumption which is maintained throughout.

The static agency setting, though useful as a benchmark, suffers from its inability to account for the source of the Authority's commitment power. In typical agency models, the ability to commit to $\tilde{x}^*$ is exogenous, yet in many settings it must be accounted from within the interaction itself. One problem may be that exogenously enforceable contracts are simply impossible (as in the case of a dictator interacting with his subjects). A more subtle issue arises if the Authority can make a short-term commitment to $\tilde{x}^*$, but cannot commit not to revert to $\tilde{x}^0$ in the future. Subordinates might challenge in the hope of weakening the Authority's resolve to uphold this commitment. Thus, even if short-term commitments are available, they will not necessarily be effective against subordinates with a long-term interest in the future of the relationship.

2.3. Repeated Interaction and Reputation

We now turn to the case in which compliance is based on the Authority's reputation. Assume that the Authority and the same N agents interact repeatedly. The beginning of a stage is identified by a history h containing information about all past decisions of the Authority and agents' signals from previous rounds.

Reputation: By a 'reputation' we will mean agents' expectations about how the Authority will respond to their behavior. This is modeled as a rule $\tilde{x}^*$ that maps agents signals into rewards and sanctions. We also identify the trivial rule $\tilde{x}^0$ with the collapse of reputation. Clearly, there are a priori many possible reputations, one corresponding to each possible rule. For the moment, we take $\tilde{x}^*$ as given, and focus on what makes $\tilde{x}^*$ robust against agents' incentives to challenge it. In Section 4 we address the issue of 'reputation design' in which the Authority optimally designs its reputation.

We impose two types of rigidities on the use of reputation by the Authority. First, the rule $\tilde{x}^*$ is the same across agents (although the ex post action taken $\tilde{x}^*(b_n)$ can, of course, vary). Second, if the

Authority ever blinks (plays $\tilde{x}^0$), then it receives a payoff of 0 in the current as well as in all future periods. That is, once it blinks, the Authority can never recover its reputation.[13]

Timing: In the body of the paper I assume that, at a history h, the Authority either plays $\tilde{x}^*$ (with probability denoted $\sigma(h)$), or blinks $\tilde{x}^0$ (with probability $1 - \sigma(h)$); agent n observes $\tilde{x}^0$ or $\tilde{x}^*$ and decides to either comply a^* (with probability $\alpha_n(h)$) or challenges. This timing gives the Authority a slight edge because it can make a short-term commitment to $\tilde{x}^*$. Agents are sure that $\tilde{x}^*$ will be implemented in the current period, but may still believe that his actions today could undermine the Authority's resolve to defend its reputation in the future. A separate, more elaborate, argument (available from the author on request) shows that qualitatively similar results obtain when the Authority and the agents choose their actions simultaneously and independently from each other.

Symmetrization: To make the analysis tractable, we impose the following symmetrization of the agents' strategy: at the beginning of stage h, a randomly drawn list of agent names $C(h) \subseteq N$ is announced publicly and designated as challengers for that period. The draw is symmetric in the sense that agents have equal probability of being selected. This allows us to treat agents asymmetrically within any given stage, yet ensuring that they all have the same ex ante value function. Note that equilibria and best responses are defined relative to all deviations, not just those satisfying this symmetrization.

2.4. Agents' Predictions and Optimization

At any history h at which agents face $\tilde{x}^*$, their action choice generates a (random) next period history h'. Agents' beliefs about the Authority's future behavior may be represented by state-contingent continuation strategies $\{\hat{\sigma}_{h'}\}$, one for each possible next period history h'.

Conceptually, agents' behavior is driven by two distinct factors: (1) *prediction* about the Authority's future behavior $\{\hat{\sigma}_{h'}\}$; and (2) *optimization*, given these predictions. In a standard equilibrium model of behavior, agents optimize against the actual strategy followed by the Authority in the future. That is, in equilibrium, agents optimize *and* their predictions are correct. An equilibrium model, however,

---

[13] Formally, if h is a history such that $\tilde{x}^0$ has been played, then the action set of the Authority is reduced to $\tilde{x}^0$ in all future periods.

does not explain where these predictions come from or what will ensure that they coincide with the actual strategy. In the analysis below, I use a more conservative description of agents' behavior, requiring only that they optimize relative to *some* prediction.

To formalize this, consider a history h at which agents face $\tilde{x}^*$ and hold a conjecture $\{\hat{\sigma}_{h'}\}$. Any strategy α generates values $U(h)$ and $U(h')$ satisfying:

$$U(h) = (1-\delta)E(u-g) + \delta EU(h'),$$

where δ is a discount factor common to all agents and the Authority. The symmetrization assumption ensures that agents have the same continuation values $U(h')$.

Call α a *best response* to $(\tilde{x}^*, \{\hat{\sigma}_{h'}\})$ at h if each agent maximizes his expected discounted payoff given $(\tilde{x}^*, \{\hat{\sigma}_{h'}\})$ and the strategies of other agents. A mixed action $\alpha(h)$ is *locally optimal at* h if there is $(\tilde{x}^*, \{\hat{\sigma}_{h'}\})$ and a best response $\hat{\alpha}$ such that $\alpha(h) = \hat{\alpha}(h)$; α is *locally optimal* if it is locally optimal at every h. Let $\mathcal{A}$ denote the set of locally optimal agent strategies.

In other words, α is locally optimal if actions can be rationalized as agents' best response to some theory about the Authority's behavior. We require agents' predictions at each stage to agree, as they always do in an equilibrium. Except for this requirement, local optimality is a minimal rationality restriction on agents' behavior. For instance, the predictions that rationalize their actions may differ from the actual strategy followed by the Authority, and may be inconsistent across periods. Although the formal definitions differ, local optimality is in the spirit of rationalizability (Bernheim (1984) and Pearce (1984)). The idea is to require behavior to be motivated by some hypothesis about how the future will unfold, but without requiring the hypothesis to be correct.

3. ANALYSIS

3.1. Agents' Incentives and Free-Riding

Consider a history h in which $\tilde{x}^*$ is chosen, so agents know that the Authority will defend its reputation for the current round.[14] Although a challenge imposes short-term costs, agents with long-term interest in the relationship may nevertheless challenge if they believe doing so sufficiently influences the Authority's future behavior to offset these short-term costs. If $\alpha(h)$ is a locally optimal strategy, with corresponding continuation values $U(h')$, then agent n complies only if:

$$Eg(\tilde{x}^*|a^*) \geq S + Eg(\tilde{x}^*|a^0) + \underbrace{\frac{\delta}{1-\delta}\left[E(U \mid a_n = a^0) - E(U \mid a_n = a^*)\right]}_{I_n}. \qquad (IC')$$

The term I_n reflects agents' perceived influence on the Authority's future behavior. Myopic agents necessarily have no influence (that is, $\delta = 0$ implies $I_n = 0$), in which case IC' reduces to the static constraint IC. On the other hand, agents who care about the future may be willing to incur the short-term cost of challenging if they believe their actions will be sufficiently pivotal to the Authority's future behavior. The remaining non-pivotal agents, on the other hand, free-ride on others' challenges. The next result shows that most agents free-ride:[15]

PROPOSITION 1: *For any $\tilde{x}^*$ at which (IC) does not bind, there a number K^* such that no more than K^* agents challenge, for any h, $\alpha \in \mathcal{A}$, and N.*

The number K^* is determined by the primitives (N, δ, c, g) and the Authority's reputation, $\tilde{x}^*$. This makes it possible to examine how changes in either the primitives or the reputation $\tilde{x}^*$ can impact the severity of the free-riding problem among agents. Section 4 provides a tighter characterization of K^* and uses it to examine the question of reputation design.

---

[14] Since the game is trivial once $\tilde{x}^0$ is chosen, to avoid redundancy I will assume that the Authority has selected $\tilde{x}^*$ when discussing the agents' decisions.

[15] The proofs of Propositions 1, 2 and 4 are in the Appendix.

3.2. Robust Reputations: Definition

If authority is to be founded on "convergent expectations", we must explain what mechanism generates these expectations. In an equilibrium model, formation of expectations is trivial: players know the strategies of their opponents, so their expectations coincide with the truth. An often noted weakness of equilibrium analysis is that it does not explain where these expectations come from, or what mechanisms ensure they indeed coincide with the strategies actually played.

A more robust mechanism derives players' expectations from the fundamentals of the game, rather than the strategies postulated by the modeler as a candidate equilibrium. I will therefore assume that agents' behavior is only locally optimal–*i.e.*, can be rationalized as a best response to *some* belief–without being specific about what this belief may be, or whether it coincides with the actual behavior of the Authority.

Formally, for a reputation $\tilde{x}^*$, let σ^* denote the strategy in which $\tilde{x}^*$ is played with probability 1 at every history. Then $\tilde{x}^*$ is *robust* if, starting from any history h, σ^* is a strict best response to every locally optimal strategy $\alpha \in \mathcal{A}$.

To put this definition in perspective, consider what happens at a reputation which is *not* robust: agents know the Authority might defend its reputation for some conjectures about their behavior, but not others. Their prediction about the Authority's behavior, and hence their own decision whether to comply or challenge, depends on their beliefs about what the Authority believes about them. Dependence on such a complex chain of reasoning means that predictions will not necessarily converge to anything robust enough to serve as foundation for authority.

On the other hand, a reputation is robust if this chain of reasoning always converges to the expectation that the Authority will defend its reputation. The mechanism generating agents' expectations here is simple: it is strictly optimal for the Authority to defend its reputation, provided only that it believes agents' behavior can be rationalized by *some* belief about its future actions.

This notion of robustness has an interesting interpretation in terms of focal points. Under a robust reputation, expectations are focal in the sense that, from knowledge of the reputation $\tilde{x}^*$ and the primitives N, δ, c, g, agents can determine that reputation will indeed be defended. While this reasoning ultimately pins down the equilibrium played, the important point to keep in mind is that

such reasoning can be carried out independently of any putative equilibrium suggested as a solution to the game.

3.3. Robust Reputations: Characterization

We now proceed to characterizing robust reputations in terms of the primitives of the environment. If a fraction of agents $\gamma \in [0,1]$ challenge at h, then the Authority's short-term average net gain from defending its reputation one more round is

$$(1-\gamma)[S - Ec(\tilde{x}^*|a^*)] - \gamma Ec(\tilde{x}^*|a^0).$$

Let $\gamma^* = \gamma^*(\tilde{x}^*)$ be the largest fraction for which this net gain is non-negative.[16] Then γ^* may be interpreted as the maximum fraction of challenges the Authority can tolerate indefinitely and still defend its reputation.

PROPOSITION 2: *A reputation $\tilde{x}^*$ is robust if*

$$\frac{K^*}{N} < \gamma^*. \qquad (R)$$

and only if

$$\frac{K^*}{N} \leq \gamma^*.$$

The intuition for the sufficiency part is that, by blinking, the Authority can always guarantee a payoff of 0 in the current as well as all future rounds. A reputation is then like an option, maintained only as long as its value is positive. But defending a reputation $\tilde{x}^*$ that satisfies (R) generates strictly positive expected payoff in the current round, while its option value for future rounds is non-negative. Defending $\tilde{x}^*$ is therefore optimal at all periods.

In analyzing the reputation design problem, the sufficient condition (R) creates inessential, technical complications because the set of reputations $\{\tilde{x}^* : \frac{K^*}{N} < \gamma^*\}$ is not necessarily closed. To ensure

---

[16] That is, γ^* is the unique number satisfying $(1-\gamma)[S - Ec(\tilde{x}^*|a^*)] = \gamma Ec(\tilde{x}^*|a^0)$.

existence of an optimal choice of reputation, it is more convenient to expand (R) by taking its closure $(\bar{R})$. Formally, we consider the set of reputations:

$$cl\left\{\tilde{x}^* : \frac{K^*}{N} < \gamma^*\right\}, \qquad (\bar{R})$$

where cl is the closure operation on subsets. By definition, any reputation satisfying $(\bar{R})$ can be arbitrarily closely approximated by one satisfying (R), hence necessarily robust reputation.

3.4. Equilibrium

Let σ^* and α^* denote the strategies in which $\tilde{x}^*$ and a^* are played in every stage h.

PROPOSITION 3: *If $\tilde{x}^*$ is a robust, then (σ^*, α^*) is the game's unique equilibrium.*

Proof: If (σ, α) is an equilibrium, then agents' strategy must be locally optimal, so $\alpha \in \mathcal{A}$. Robustness means that σ^* is strictly dominant relative to strategies in $\mathcal{A}$, so $\sigma = \sigma^*$. But then α^* is agents' unique, in fact strict, response to σ^*, so we must also have $\alpha = \alpha^*$.

Q.E.D.

The point to be emphasized here is that robustness implies not just the uniqueness of equilibrium, but also a mechanism explaining how expectations are formed, and why they converge in support of the Authority's reputation. The reasoning has an iterative dominance flavor; in particular, it does not require as precise a knowledge of opponents' strategies as would be required in equilibrium:

i) Agents choose a locally optimal strategy $\alpha \in \mathcal{A}$—*i.e.*, act in a way that can be rationalized relative to some (possibly false) conjecture about the Authority's behavior. By Proposition 1, all but possibly K^* agents free-ride;

ii) Robustness implies that σ^* is uniquely optimal against *any* $\alpha \in \mathcal{A}$;

iii) Knowing this, agents correctly conclude that challenging is pointless, so they all comply.

4. REPUTATION DESIGN

Observed authority relationships display an immense variety in terms of the intensity of the rewards and sanctions used, the circumstances under which they are applied, and the scope of activities over which authority is exercised. This variety presumably reflects differences in the players' objectives and the constraints imposed by the environment in which they interact. In this section, I provide a simple framework accounting for an authority's choice of reputation as a function of the primitives of its environment (*i.e.*, g, c, N, δ and so on).

4.1. The Reputation Choice Problem

Consider a game in which the Authority first selects one scheme $\tilde{x}^*$ to commit to, then the game proceeds as described in the last two sections (with the chosen $\tilde{x}^*$). The Authority either chooses to blink $\tilde{x}^0$ (an option which is always available), or some reputation $\tilde{x}^*$ that is robust. These assumptions reduce the problem of reputation choice to that of solving a constrained maximization problem:

$$\max_{\substack{\tilde{x}\in\{\tilde{x}^0,\tilde{x}^*\}\\ a\in\{a^0,a^*\}}} \frac{1}{N}\sum_n E[v(b_n) - c(x(b_n)) \mid a], \quad \text{subject to } PC, IC \text{ and } \bar{R}.$$

This differs from the static agency problem only by the addition of the robustness constraint ($\bar{R}$). We will say that the Authority chooses a non-trivial reputation if the outcome of this maximization problem is an $\tilde{x}^* \neq \tilde{x}^0$.

We now turn to how the primitives of the problem (*i.e.*, N, δ, g, c and so on) influence optimal reputation choice. Since the role played by IC and PC is known from standard agency models, I focus on the new constraint ($\bar{R}$). For fixed δ and N, the effect of $\tilde{x}^*$ on the condition $\frac{K^*}{N} \leq \gamma^*$ can be broken down into its effect on the two sides of the inequality. Write γ^* as:

$$\gamma^* = \frac{S - Ec(\tilde{x}^*|a^*)}{S - Ec(\tilde{x}^*|a^*) + Ec(\tilde{x}^*|a^0)}$$

In terms of the effect on γ^*, rewards and sanctions enter symmetrically: increases in rewards or sanctions both decrease the maximum fraction of challenges the Authority is willing to tolerate. The effect of changes in reputation choice $\tilde{x}^*$ on the maximum fraction of challengers $\frac{K^*}{N}$ is, however, more subtle. We turn to this next.

4.2. Manipulating the Intensity of Free-Riding

Consider a locally optimal mixed action $\alpha(h)$ supported by continuation values $U(h')$. Condition IC' implies that, for an agent to challenge, his influence must exceed the (appropriately discounted) net cost of challenging:

$$[E(U \mid a_n = a^0) - E(U \mid a_n = a^*)] \geq \frac{1-\delta}{\delta} Eg(\tilde{x}^*|a^*) - S - Eg(\tilde{x}^*|a^0). \qquad (*)$$

We will show (Lemma A.1) that any $\alpha \in \mathcal{A}$ generates continuation values in the interval $[S, Eg(\tilde{x}^*|a^*)]$. It is therefore convenient to normalize (*) by dividing both sides by $S - Eg(\tilde{x}^*|a^*)$ to obtain:

$$d_n \geq r(\tilde{x}^*, \delta).$$

Thus, for agent n to challenge at α and U, his (normalized) influence, d_n, must exceed the critical pivotalness threshold, $r = r(\tilde{x}^*, \delta)$, necessary to offset the short-term cost of challenging.

Call agent n *r-pivotal* relative to (α, U) if $d_n \geq r$. Clearly, the more agents can be made pivotal, the greater the number of challenges the Authority might face. Our definition of robustness may be restated in terms of:

$$K^*(r) = \max_{\alpha \in \mathcal{A}} \#\{n : d_n \geq r\},$$

which is the maximum number of agents that can be convinced they will be r-pivotal.

Reputation choice influences agents' behavior through its effect on $r = r(\tilde{x}^*, \delta)$. Our problem, then, reduces to the questions: How many agents can be made r-pivotal? and, How does their number depend on r? Here I use general characterizations of influence and the maximal number of pivotal players found in Al-Najjar and Smorodinsky (1996) to identify the effect of reputation choice on K^*.

PROPOSITION 4: *$K^*(r)$ is decreasing in r, with $K^*(0) = N$ and $K^*(r > 1) = 0$. Furthermore, for large N and small $r > 0$, $K^*(r)$ decreases at the rate $\frac{1}{r^2}$.*

The intuition is that as r increases (to r', say), making an agent r'-pivotal is accomplished by shifting influence from other agents to agent n. This, however, reduces the number of r'-pivotal agents, causing $K^*(r')$ to decrease relative to $K^*(r)$.

4.3. Rents for Compliance

One consequence of the analysis is that the use of a robust reputation against agents with their own reputational concerns precludes the possibility that the incentive constraints bind:

PROPOSITION 5: *For $\delta > 0$, IC does not bind for any $\tilde{x}^*$ satisfying $(\bar{R})$. In particular, under the optimal reputation $\tilde{x}^*$, agents strictly prefer compliance to challenge, so the optimal static contract is inconsistent with a robust reputation.*

Proof: Condition (IC) means that $Eg(\tilde{x}^*|a^*) - S + Eg(\tilde{x}^*|a^0) = 0$. Thus, $r(\tilde{x}^*, \delta) = 0$ since $\delta > 0$. By Proposition 4, this implies $K^* = N$, so $\frac{K^*}{N} = 1 > \gamma^*$, violating $(\bar{R})$.

Q.E.D.

Define the *rents for compliance* as the difference between what an agent gets by complying rather than challenging, $G^* = Eg(\tilde{x}^*|a^*) - S - Eg(\tilde{x}^*|a^0)$ under the optimal reputation $\tilde{x}^*$. Proposition 5 says that the optimal reputation always involves positive rents.

The payment of these rents may appear puzzling: in the unique equilibrium of the game, none of the dynamic incentive constraints (IC') actually bind. Why would non-binding constraints be relevant in determining the Authority's choice of reputation? Why can't the Authority squeeze a bit more surplus by lowering the rents G^*—just like the principal does in a static agency relationship?

This puzzle is resolved by noting that robustness requires the Authority to hedge against any agents' behavior that can be rationalized with respect to *some* conjecture, not just those which happen to be part of the putative equilibrium postulated by the modeler. Robustness therefore reflects the Authority's uncertainty about the determinants of agents' expectations that leads it to take account of the dynamic constraints (IC') which differ from those in the unique equilibrium of the game.

As an example, consider a standard agency scheme $\tilde{x}$ which implements a^* (in the sense of satisfying PC and IC). Since agents are indifferent between complying and challenging, $\tilde{x}$ leaves them with no rents. Consider now the game in which the Authority chooses between $\tilde{x}$ and $\tilde{x}^0$.

One equilibrium of this game consists of repeated play of a^* and $\tilde{x}$. This, however, is supported by an expectational bubble: agents comply because they expect the Authority to defend its reputation,

which is optimal only because agents comply. These expectations are based on self-referential, self-confirming reasoning, rather than a mechanism which generates them from the fundamentals of the agents' environment. Thus, just as upholding authority is an equilibrium, so is its collapse: the repeated play of $(a^0, \tilde{x}^0)$ also is an equilibrium, supported by a different expectational bubble—this time agents challenge because they expect the Authority to blink, which is optimal because agents challenge.

Robustness rules out such circular chains of reasoning. This is not to say that the phenomena of sudden shifts between compliance and challenge is uninteresting. Rather, such shifts would be inconsistent with the notion of authority as a voluntarily accepted, stable arrangement which minimizes enforcement costs.

4.4. Effect of Size and Discount Factor

We now turn to evaluating the roles of N and δ as reflected in the necessary condition:

$$\frac{K^*(r(\tilde{x}^*, \delta))}{N} \leq \gamma^*(\tilde{x}^*). \qquad (\hat{R})$$

Unlike $\tilde{x}^*$ (which affects both sides of the inequality), N and δ enter much more neatly. Increasing N unambiguously reduces the fraction of challenges $\frac{K^*}{N}$ at the rapid rate of $\frac{1}{N}$. Thus, as N increases, the set of reputations $\tilde{x}^*$ satisfying $(\hat{R})$ approaches that of reputations satisfying IC. The robustness requirement becomes less and less binding, and the optimal reputation choice approaches the static agency contract.

The effect of the discount factor δ is similar. As agents become more myopic (*i.e.*, as $\delta \to 0$), r increases without bound: agents who care little about the future of the relationship challenge only if their impact on continuation values is substantial. But Proposition 5 tells us that there can be no more than $K^*(r)$ pivotal agents, and the higher is the pivotalness threshold r, the smaller is K^*, so more agents free-ride. In fact, as r exceeds 1, K^* drops to zero, so all agents free-ride.

In summary, we have a continuity result with respect to N and δ: increasing N raises the fraction of challenges the Authority may be willing to tolerate, while decreasing δ exacerbates the free-riding problem. Both lead to a reputation choice which approaches that generated by the static agency contract. Rents and excessive punishments consequently disappear.

4.5. Comparison with Models with Myopic Agents

It is useful to contrast our results with two benchmark models in the literature. Consider first the case in which the Authority (as an employer in an agency model, for instance) could commit to a contract in a static setting. A well-known result in agency theory is that the optimal contract should leave agents just indifferent between compliance and challenge, and between participating in the relationship or choosing their outside option (Grossman and Hart, (1982)). A contract for which either the incentive or participation constraint do not bind can be modified to increase the principal's expected payoff without violating these constraints.

This reasoning breaks down when applied to agents with their own reputational concerns: leaving agents just indifferent between complying and challenging means they suffer no cost from challenging, yet might gain if their challenges lead to a better continuation as a result of undermining the principal's reputation.

Second, consider the reputation models in the seminal works of Milgrom and Roberts (1982) and Kreps and Wilson (1982). There, an incumbent faces a sequence of myopic challengers with no long-term interest in the relationship, and hence no incentive either to test the Authority's resolve or to build their own reputation.

Since myopic agents completely discount the future impact of their behavior, their behavior is completely driven by short term incentives which are fully captured by the static incentive constraint IC. In particular, challengers always free-ride, and the public-good nature of reputation is trivial. The role of reputation is then simply as a device that replaces binding contracts (unavailable because, say, an incumbent cannot write a contract which commits him to fighting entry). The issue of reputation design reduces to following the optimal commitment strategy that replicates the static agency contract, generally leaving agents with no rents.

It is hard to interpret the short-run opponents model in many settings of interest, including contracting settings, such as employment or franchising, or indeed, even predation in environments where entrants have a long-term interest in the incumbent's market. The treatment of this paper reflects the problems arising when agents' incentives cannot be reduced to short-term considerations.

There are also important technical modeling differences in that I do not use the (by now standard) technique of perturbing the game by introducing types to obtain reputation results. Rather I rely on

a rationalizability-like restriction which requires the Authority to choose a reputation under which agents' beliefs become focal, in the sense robustness. This has several advantages, including providing an explicit mechanism for generating expectations as well as powerful, economically meaningful, restrictions on reputation choice in terms of the magnitude of rents which needed to secure compliance.

5. APPLICATIONS AND EXAMPLES

The predictive content of the model derives from the link it establishes between the exogenously given, observable features of the environment, and the Authority's optimal choice of reputation. As a test of the plausibility of the model's implications, I provide three examples of how alternative authority structures might arise as a result of optimal reputation choice.

5.1. Gift Exchange and Efficiency Wages

Neoclassical as well as agency theoretic predictions of employment compensation have long been criticized for their failure to reflect important aspects of real-world employment relationships. One set of stylized facts, reported by Bewley (1998), is that successful management treats employees better than the minimum, contractually agreed-upon level, and employees reciprocate by contributing effort, care and dedication in excess of their contractually specified levels. This practice, often reflected by concepts like morale, gift exchange, efficiency wages, and worker attachments[17] is particularly significant in view of the fact that most employment contracts are highly incomplete, so the explicitly specified dimensions of performance are, indeed, quite limited.

The model of this paper provides a new and distinct explanation for this practice. The employer offers salaries and work conditions above the minimum level (an efficiency wage) in exchange for employees' voluntary compliance with his authority. To see how this follows from the model, interpret a^0 and $\tilde{x}^0$ as the minimum, contractually specified levels of compliance and compensation, respectively. Proposition 5 implies that the use of a robust reputation to support this exchange requires agents to strictly prefer compliance to challenging by rents of magnitude G^*.

The predictive content of the model is in accounting for the magnitude and form of rents as a function of the primitives. Giving employees 'something to lose' in the form of an efficiency wage is, a priori, not the only way to provide incentives; coercion—giving agents 'something to fear'—could also make them strictly prefer compliance. How can the model account for the use of inducement rather

---

[17] See Akerlof (1982) for exposition of the main idea of morale and gift exchange, Weiss (1990) for a survey of the literature on efficiency wages, Halaby (1986) on worker attachment, and Bewley (1998) for a more detailed account of current theory.

than coercion? A special feature of employment relationships is the availability of a valuable outside option (workers can simply quit their jobs to seek employment elsewhere). Limited liability rules and other legal restrictions on the severity of sanctions represent additional exogenous restrictions on employers' ability to resort to sanctions to extract compliance. These special features of employment relationships, not shared by other authority structures such as dictatorships, imply that rents for compliance must take the form of additional rewards, resulting in the failure of the participation constraint to bind.[18]

Several suggestions appeared in the literature to explain why workers are treated better than the minimum, contractually-specified level. Some are based on the need to overcome monitoring and screening costs (Weiss (1990)), while others revolve around employees' sentiments towards the firm, and the positive psychological effects that a better treatment can have on workers' moral. The present model is not inconsistent with these explanation. Rather, it provides a complementary, and to my knowledge new, mechanism that leads to the payment of efficiency wages. This mechanism shows that qualitatively similar, if not observationally indistinguishable, predictions about efficiency wages and gift exchange may be a consequence of the use of reputation by welfare-maximizing individuals' to structure long-term interactions.

An interesting test of the model's prediction is provided by the traditional distinction between primary and secondary labor markets: "Primary sector jobs have stability, low quit rates, good working conditions, promotion according to a promotion ladder, acquisition of skills, and good pay. In contrast, secondary sector jobs have high quit rates, harsh discipline, little chance of promotion, low acquisition of skills, and poor pay." (Akerlof (1984), p. 79). Higher turnover in secondary sectors means lower expected duration of stay with any particular employer, hence a lower subordinates' discount factor δ. The difference between these two sectors provides a natural comparative static test of the model's predictions as subordinates' reputational concerns, measured by δ, vary. Lower δ weakens the robustness constraint ($\bar{R}$), so optimal reputation choice approaches that in a static agency setting, where agents have no reputational concerns (see Section 4.5). Rents for compliance, which may correspond to higher monetary pay, better working conditions and better treatment on

[18] In the notation of the model, attractive outside options correspond to higher values of u_0, while exogenous restrictions on the use of sanctions may be captured by a g function that rises rapidly with the severity of sanction (negative values of x).

the job, consequently disappear as δ approaches 0. The model's prediction of a negative correlation between the magnitude of rents and turnover rate appear broadly consistent the distinction between the primary and secondary sectors above.

Finally, the practice where parties voluntarily refrain from extracting the maximum private benefit from an exchange are pervasive not just in employment contracts and personnel management, but in nearly every other long-term social and political interaction where conflicting reputational concerns arise. It is folk wisdom that leadership and authority over subordinates with 'nothing to lose' is difficult, if not impossible, to maintain. Greed, in the form of milking subordinates for all they are worth, is not just morally questionable, but is often taken as sign of bad leadership. Cole, Mailath, and Postlewaite (1996), who examine the role of such rents in sustaining a class system through the threat of ostracism, report a telling quote from Alexander Solzhenitsyn:

> "You only have power over people so long as you don't take everything away from them. But when you've robbed a man of everything, he's no longer in your power–he's free again."

The model of this paper confirms that rents for compliance are a fairly general and robust consequence of using reputation as a foundation for authority relationships.

5.2. Dictatorships

Dictatorship is one of the most common forms of government throughout history.[19] Yet their existence and, often, remarkable stability raise several puzzles. The only resource under the direct physical control of the dictator is the ability to issue directives—cheap talk, with little or no direct consequence on subordinates' actions and welfare. What gives force to these directives is subordinates' willingness to carry them out, presumably due to a "convergence of expectations" in the dictator's favor.

Dictatorships are not monolithic organizations. It is physically impossible for a single individual to control a large population directly, so dictatorships inevitably rely on layers of subordinates charged

---

[19] See Wintrobe (1998), who provides an insightful analysis of the subject.

with controlling and monitoring lower layers. The stability of the dictator's rule depends on the distribution of rents for compliance throughout the hierarchy.

At the lowest level there are the masses of (typically) unorganized population. Although the resources available to a regime are usually trivial compared to the potential resources of the population, a large N implies that free-riding is rather severe. Individuals behave nearly myopically, allowing control of a large population by a much smaller group. The role of size in simplifying control of large populations has been noted by observers of dictatorial regimes. Commenting on Stalin's rule, Michael Polanyi wrote:

> "The stability of such naked power increases with the size of the group under its control, for a disaffected nucleus which might be formed locally by a lucky crystallization of mutual trust among a small number of personal associates would be overawed and paralyzed by the vast surrounding masses of people whom they would assume to be still loyal to the dictator. Hence it is easier to keep control of a vast country than of the crew of a single ship in mid-ocean." (quoted in Wrong (1979), p. 94)

The only countervailing force in a large population is the presence of mechanisms which organize and coordinate subordinates, such as independent institutions (opposition parties, religious institutions, and so on). Consequently, breaking or weakening organized resistance, "divide-and-conquer" strategies, and atomizing the population are, of course, the hallmark of many dictatorial regimes. Such practices exacerbate free-riding by increasing the number of independent decision-making agents.

A more subtle issue is the extensive use of sanctions rather than rewards. Using the model, it is easy to see that rewards are intrinsically more costly: conditional on compliance, a^*, sanctions are rarely applied, while rewards would have to be awarded frequently. Furthermore, unlike employment relationships, subjects of a dictatorship have little alternative but to live under the regime, so their outside options are limited. Clearly, limitations on the dictator's ability to impose sanctions always exist, varying widely from case to case and providing a useful comparative static test of the model's predictions. Taking all these factors into account, the model predicts that sanctions (to the extent feasible) dominate rewards, at least at the population level.

The model also suggests a possible reason for an often noted stylized fact that rewards relative to sanctions increase at higher levels of the hierarchy (*e.g.*, Wintrobe (1998)). The key factor here is that the number of subordinates shrinks at higher levels in the controlling hierarchy. While the number of (potentially) pivotal subordinates K^* does not necessarily change, their fraction relative

to members of that layer, $\frac{K^*}{N}$, increases rapidly as N shrinks. Each subordinate can potentially have a large impact, so rents must correspondingly increase. But why increase rewards rather than sanctions? A subtle issue which the model points to is the potentially destabilizing effect of sanctions: they increase the short-term cost of challenging, but also increase the potential gain from undermining the Authority's reputation. This should be contrasted with the unambiguous effect of rewards, which raise the short-term as well as long-term cost of challenging. This asymmetry between rewards and sanctions leads to an increase in positive inducements as one moves up the hierarchy.

Finally, the analysis illustrates the model's ability to generate a wide range of authority structures. In Wrong's (1979) classification of authority relationships, gift exchange and efficiency wages are examples of *authority by inducement*, where a promise of reward is offered in exchange for compliance. Dictatorships are an example of *coercive authority*, using force "to establish credibility and thus to create a future power relation based on the threat of force that precludes the necessity of overt resort to it" (p. 41). Reputation plays critical role: force is "used less for the immediate effects on its victims than to establish or maintain a relation of coercive authority in the future." (p. 42). The model explains that the use of reputation as foundation for authority is capable of generating a variety of authority structures.

5.3. Legitimacy and the Boundaries of Authority

A central theme of the political science and sociology literature on authority is that submission to authority is never absolute, but limited by rules delineating the scope of activities over which authority may *legitimately* be exercised.[20] Simon's (1951) notion of acceptance set of activities over which an employer can exercise discretion, and Arrow's (1974) emphasis on the need for restrictions to ensure responsible exercise of authority, are very much in this spirit.

Although the boundaries delineating the scope of authority are rarely sharply drawn, they nevertheless represent real constraints on what an Authority can and cannot do. For instance, it may be legitimate for an employer to exercise discretion in assigning production tasks to employees, but

---

[20] In his study of authority in organizations, Scott (1998) defines: "authority is legitimate power." Similar, if less stark, statements can be found throughout the sociology literature.

not in asking them to wash his car or provide sexual favors. By shaping subordinates expectations, legitimacy consequently has considerable practical importance: collapse of authorities is often linked to subordinates' perception to their loss of legitimacy, while their stability is attributed to acquiring or retaining such legitimacy.

Legitimacy raises several complex and multi-faceted issues whose treatment is beyond the scope of this paper. However, a feature common to all notions of legitimacy is the presence of restrictions on the scope of activities over which authority may be exercised. My goal here is limited to show how the model explains the role of reputation in accounting for this key feature of legitimacy.

Imagine that subordinates' behavior consists of choosing levels of L separate activities that may be relevant to the Authority. For instance, one such activity may be a worker's performance on a production task while another represents idiosyncratic, personal aspects of his conduct. As in Section 2.2, we may also interpret the activities as compliance with state-contingent, discretionary directives. Identify an action a_k with a bundle of (levels) of these activities. For simplicity, assume that there is a finite set of such actions $A = \{a_0, a_1, \ldots, a_K\}$, with a^0 continuing to represent challenge, while $a_1, \ldots, a_K$ represent various types of compliance. For instance, different a_k's might represent different subsets of the L basic activities over which agents comply.

The reputation design problem now has two components. First, for any given possible reputation a_k, we can find the optimal scheme $\tilde{x}_k$ and corresponding rents G_k needed to support it. This is just the reputation design problem of Section 4, with a^* replaced by a_k. A second, new, dimension is the Authority's ability to choose among the various levels of compliance.

As a benchmark, assume that a_K is the action that would have been implemented in the optimal static agency contract,[21] under which the agent is left with no rents. If a_K is the only level of compliance open to the Authority, then the analysis of Section 4 implies that positive rents $G_K > 0$ must be paid to implement it. These rents could be substantial; G_K might be so high that no robust reputation can implement it.

A richer set of actions $\{a_1, \ldots, a_K\}$ ameliorates the Authority's predicament. By allowing for the option of a lower level of compliance, $a^* = a_k$, in exchange for a lower rent, G_k, the Authority can

[21] Or equivalently, using a reputation against myopic agents.

establish a reputation in situations which it might not otherwise have been able. Note that lower levels of compliance may reflect restrictions on the number of dimensions over which authority may be exercised.

The model therefore generates restrictions on the scope of authority based on this trade-off between the desire to achieve a high level of compliance and minimize the rents necessary to implement it. Since this trade-off depends on observables, such as the number of subordinates and the strength of their reputational concerns, it can potentially generate useful, non-vacuous predictions about observed authority structures. For instance, relative to the setting with myopic agents, a more limited scope of authority arises in long-term relationships where agents have substantial reputational concerns.

6. CONCLUDING REMARKS

When reviewing the argument that authority stems from control over the instruments of power, Arrow noted that this is "not a sufficient explanation of obedience to authority even at the immediate level ... [nor is it] ... a sufficiently deep one."[22] This paper provides some insight into the role of reputation as a basis for authority relationships and as an explanation of how these relationships are structured. Free-riding is suggested as a unifying principle that can account for parties' different abilities to carry reputations and to sort out the maze of resources and constraints underlying authority relationships.

Analytically, the paper sought to provide a model which is as simple as possible (hopefully, not simpler). Behavioral properties and comparative statics of inherently dynamic phenomena, like reputation and authority, are captured by a robustness constraint added to the standard static agency model. This makes it possible to disentangle those effects caused by subordinates' reputational concerns from the remaining ingredients of the model. Using the standard agency framework as a benchmark also makes it relatively simple to assess the impact of introducing dynamic considerations on existing models in the literature.

Several important questions about authority relationships, of course, remain. One important issue is raised by Kreps' (1990) view of corporate culture, embodied in a firm's reputation, as the 'principle' followed by the firm in such things as exercising discretion over subordinates and adjudicating disputes. The principle is not an explicit, state-contingent rule, but a general guide to be interpreted and re-evaluated as new, previously unforeseen, contingencies unfold. Applied to the present model, the legitimacy of an Authority over its subordinates consists of establishing, and subsequently adhering to, a set of principles that delineates the range of activities and subordinates' actions over which its power legitimately extends. The present model captures restrictions on the scope of authority reflected in explicitly stated rules (Section 5.3), but not restrictions embodied by general principles of the sort Kreps has in mind. Modeling such principles is considerably more difficult, apparently for the same sort of reasons that unforeseen contingencies and incomplete contracting are difficult to model. I hope to pursue this in future work.

---

[22] Arrow (1974, p. 71).

APPENDIX

PROOFS

I first prove Proposition 4, from which Proposition 1 easily follows. Unless indicated otherwise, h' will denote the immediate successor history to h. I begin with a simple lemma.

Lemma A.1: *$[S, Eg(\tilde{x}^*|a^*)]$ is the smallest interval containing the range of continuation values $U(h')$, uniformly over h, N, and $\alpha \in \mathcal{A}$.*

Proof: Fix a history h and a locally optimal strategy α. We show that S and $Eg(\tilde{x}^*|a^*)$ represent, respectively, the least upper bound and greatest lower bound on $U(h')$, uniformly over h, N, and $\alpha \in \mathcal{A}$. For a history h, let α_h denote an agents' strategy for the game starting at h, and $\{\hat{\sigma}_{h'}\}$ the conjectures at h to which $\alpha = (\alpha(h), \{\alpha_{h'}\})$ is a best response (that is, $\alpha(h)$ represents the actions at stage h, while $\alpha_{h'}$ represents the continuation of α_h on the history h').

Clearly, by complying in each period, any agent can guarantee himself an average continuation value of at least $Eg(\tilde{x}^*|a^*)$, so this is a lower bound on the continuation values. This bound is achieved by taking $\hat{\sigma}_{h'} = \sigma^*$ for all h'. In the other direction, any conjecture $\hat{\sigma}_{h'}$ consists of a (possibly state-contingent) sequence of a^0 and a^* choices, against which an agent's stage payoff is an average of S, $Eg(\tilde{x}^*|a^*)$, and $S - Eg(\tilde{x}^*|a^0)$. By our assumptions, such convex combinations are bounded above by S. This bound is achieved by taking the conjecture that the Authority chooses $\tilde{x}^0$, to which agents' best response guarantees them S each period.

Q.E.D.

Before proving Proposition 4, we need the following notation. If h' is the immediate successor to h, then it includes, in addition to h, information about the action of the Authority, $\tilde{x} \in \{\tilde{x}^*, \tilde{x}^0\}$, and the vector of agents' signals, $(b_1, \ldots, b_N) \in B^N$. We may therefore identify $h' = (b_1, \ldots, b_N; \tilde{x}^*, h)$.

Proof of Proposition 4: Fix a history h, a strategy $\alpha \in \mathcal{A}$, and assume that $\tilde{x}^*$ is played at h. Let H' denote the set of next period histories, and define the function $d : H' \to [0, 1]$ by $d(h') = \frac{U(h')}{S - Eg(\tilde{x}^*|a^*)}$. Playing $\alpha(h)$ implies a distribution over the agents' pure actions at that stage. Let $\{a_1, \ldots, a_N\}$ denote a vector of such actions (not necessarily symmetric). Then each signal

b_n has positive probability $\pi_{a_n}(b_n) > 0$ by assumption. Let $E(d(b_{-n}, b_n) \mid b_n = b, a_{-n})$ denote the expectation of d conditional on agent n's signal. Define $Z_n(a_1, \ldots, a_N) = \max_{b,b'} E(d(b_{-n}, b_n) \mid b_n = b, a_{-n}) - E(d(b_{-n}, b_n) \mid b_n = b', a_{-n})$. That is, Z_n is the maximum difference in the expected value of d corresponiding to any pair of agent n's signals, given this agent's uncertainty about other agents' signals. By Theorem 2 in Al-Najjar and Smorodinsky (1996), for any constant $r > 0$, there is an integer, $K(r)$ such that there can be at most $K(r)$ agents for whom $Z_n > r$, uniformly over all functions d (hence all continuation values U), profiles $(a_1, \ldots, a_N)$, and N.

Since agents in our model control only the distribution of their signals, not the signals themselves, Z_n is a also an upper bound on how large changes in the distribution on signals can have on the conditional expectation of d. That is, define $\hat{Z}_n(a_1, \ldots, a_N) = \max_{a^0, a^*} \mid E(d(b_{-n}, b_n) \mid a_n = a^0, a_{-n}) - E(d(b_{-n}, b_n) \mid a_n = a^*, a_{-n})$, then $\hat{Z}_n \leq Z_n$, so $K(r)$ is also an upper bound on the number of agents for whom $\hat{Z}_n > r$. We now let $K^*(r)$ denote the least upper bound on the number of agents for whom $\hat{Z}_n > r$.

That $K^*(0) = N$ and $K^*(r > 1) = 0$ is obvious. The fact that $K^*(r)$ decreases at the rate $\frac{1}{r^2}$ for large N and small $r > 0$ follows from the asymptotic formula, proven in Al-Najjar and Smorodinsky (1986) stating that $K^*(r) \simeq \frac{1}{\epsilon \pi r^2}$, where $\epsilon = \min_{a,b} \pi_a(b)$ is strictly positive by assumption.

Q.E.D.

Proof of Proposition 1: Follows immediately from Proposition 4.

Q.E.D.

In the next proof, we use $V(h; \sigma, \alpha)$ to denote the Authority's average discounted payoff evaluated at history h when the profile (σ, α) is played. If this profile is clear from the context, then we simply write $V(h)$.

Proof of Proposition 2: Suppose that $\tilde{x}^*$ satisfies (R). As a notational convention, we write $\sigma' \succ_\alpha \sigma$ if the Authority strictly prefers (in terms of present discounted payoffs) strategy σ' to σ when playing against α. Note that $\succ_\alpha$ is continuous relative to weak convergence of strategies.[23] We want

[23] A sequence of pure strategies $\{\sigma_m\}$ weakly converges to a strategy σ if for every integer K, there is M large enough such that, $m \geq M$ implies that σ_m and σ coincide on all histories of length at most K. This guarantees that discounted expected payoffs converge for any discount factor $0 \leq \delta < 1$.

to show that for any σ, such that $\sigma \neq \sigma^*$, and any $\alpha \in \mathcal{A}$, we have $\sigma^* \succ_\alpha \sigma$.

Fix $\sigma \neq \sigma^*$, and define the new strategy σ_1 as follows. Take a shortest-length history h with the property "$\sigma(h) < 1$, but $\sigma(j) = 1$ for all histories j preceding h" (that is, h is a 'first' history at which $\tilde{x}^*$ is not played with probability 1 under σ). Define $\sigma_1(h) = 1$; for any immediate successor h' to h such that $V(h') < 0$, $\sigma_1(h') = 0$; and for every other history, σ_1 coincides with σ.

Then $\sigma_1 \succ_\alpha \sigma$. To see this, note that the only change in values occur at histories starting with h, since σ_1 and σ coincide otherwise. At h, the changes in payoff can be decomposed into the current period change, and the change in the expected continuation values. Proposition 1 implies that for any $\alpha \in \mathcal{A}$ there will be at most K^* challenges, while condition (R) implies that $\frac{K^*}{N}$ is less than the ratio of challenges the Authority is willing to tolerate, γ^*. Thus, playing $\tilde{x}^*$ instead of $\tilde{x}^0$ at h generates strictly positive short-term gain. On the other hand, the continuation values $V(h')$ at successor histories h' either stayed the same, or strictly increased (this occurs for h' with $V(h') < 0$). Thus, switching to σ_1 strictly increases payoffs at h, weakly increases them at all successor histories h', and leaves them unaffected otherwise. Consequently, $\sigma_1 \succ_\alpha \sigma$.

If $\sigma_1 = \sigma^*$, then we are done. Otherwise, repeat the process above with σ replaced by σ_1 to obtain a new strategy σ_2 such that $\sigma_2 \succ_\alpha \sigma_1$. Continuing in this manner, if we ever obtain a strategy $\sigma_k = \sigma^*$, then we are done, otherwise we have an infinite sequence $\{\sigma_k\}$ that converges to σ^* weakly as $k \to \infty$ and has the property $\sigma_{k+1} \succ_\alpha \sigma_k$, for all k. By continuity of $\succ_\alpha$, $\sigma^* \succ_\alpha \sigma_k$ for all k, hence $\sigma^* \succ_\alpha \sigma$.

In the other direction, suppose that the reputation $\tilde{x}^*$ is one for which $\frac{K^*}{N} > \gamma^*$. From the definition of K^*, there must be a locally optimal strategy $\alpha \in \mathcal{A}$ for which this inequality holds at every history. Let σ^0 denote the strategy in which $\tilde{x}^0$ is played at each h. Then σ^0 strictly dominates σ^* when played against α, so $\tilde{x}^*$ is not robust.

Q.E.D.

REFERENCES

Aghion, P. and J. Tirole (1997): "Formal and Real Authority in Organizations," *Journal of Political Economy*, **105**, 1-29.

Akerlof, G. A. (1982): "Labor Contracts as a Partial Gift Exchange," *Quarterly Journal of Economics Quarterly Journal of Economics*, **XLVII**, 543-69.

Akerlof, G. A. (1984): "Gift Exchange and Efficiency-Wage Theory: Four Views," *American Economic Review, Papers and Proceedings*, **74**, 79-83.

Al-Najjar, N.I. and R. Smorodinsky (1996): "Pivotal Players and the Characterization of Influence," MEDS Department, Kellogg GSM, CMSEMS working paper no. 1174R, Northwestern University.

Arrow, K. J. (1974): *The Limits of Organization.* New York: Norton.

Bernheim, D. (1984): "Rationalizable Strategic Behavior," *Econometrica*, **52**, 1007-28.

Bewley, T. F. (1998): "Listening to Business: A Study of Wage Rigidity," Yale University.

Cole, H., G. Mailath, and A. Postlewaite (1996): "Class Systems and the Enforcement of Social Norms". Forthcoming in: *Journal of Public Economics.*

Coleman, J. S. (1990): *Foundations of Social Theory.* Cambridge: Belknap Press.

Gambetta, Diego (1993): *The Sicilian Mafia: The business of Private Protection.* Cambridge: Harvard University Press.

Grossman, S. and O. Hart (1983): "An Analysis of the Principal-Agent Problem," *Econometrica*, **51**, 7-45.

Grossman, S. and O. Hart (1986): "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, **94**, 691-719.

Hadfield, G. (1990): "Problematic Relations: Franchising and the Law of Incomplete Contracts," *Stanford Law Review*, **42**, 927-93.

Halaby, C. (1986): "Worker Attachment and Workplace Authority," *Americal Sociological Review*, 51, 634-49.

Kreps, D. (1990): "Corporate Culture and Economic Theory," in *Positive Perspectives on Political Economy*, ed. by Alt, J. and K. Shepsle. Cambridge: Cambridge University Press, 90-143.

Kreps, D. and R. Wilson (1982): "Reputation and Imperfect Information," *Journal of Economic Theory*, **27**, 253-279.

Leites, N. and C. Wolf Jr. (1970): *Rebellion and Authority: An Analytic Essay on Insurgent Conflict.* Santa Monica, California: The RAND Corporation.

Milgrom, P. (1988): "Employment Contracts, Influence Activities and Efficient Organization Design," *Journal of Political Economy*, **96**, 42-60.

Milgrom, P. and J. Roberts (1982): "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, **27**, 280-312.

Milgrom, P. and J. Roberts (1990): "Bargaining Costs, Influence Costs, and the Organization of Economic Activity," in *Positive Perspectives on Political Economy*, ed. by Alt, J. and K. Shepsle. Cambridge: Cambridge University Press, 57-89.

Pearce, D. (1984): "Rationalizable Strategic Behavior an the Problem of Perfection," *Econometrica*, **52**, 1029-50.

Scott, W. R. (1998): *Organizations (4th edition).* New Jersey: Prentice Hall.

Selznick, P. (1969): *Law, Society, and Industrial Justice.* New York: Russell Sage Foundation.

Simon, H. (1951): "A Formal Theory of Employment Relationship," *Econometrica*, **19**, 293-305.

Weiss, A. (1990): *Efficiency Wages: Models of Unemployment, Layoffs, and Wage Dispersion.* Princeton, New Jersey: Princeton University Press.

Williamson, O. (1985): *The Economic Institutions of Capitalism.* New York: Free Press.

Wintrobe, R. (1998): *Political Economy of Dictatorship.* Cambridge: Cambridge University Press.

Wrong, D. H. (1979): *Power: Its Forms, Bases, and Uses.* Chicago: The University of Chicago Press.