Governance and Management of Autonomous Organizations[*]

Daniel Ferreira [†] Jin Li [‡]

February 2, 2025

Abstract

An organization is *autonomous* if it has the right or power of self-government. Self-government implies that autonomous organizations cannot rely on outside parties for contract enforcement; all contracts must be relational or self-executing. We present a model in which an autonomous organization commits to a governance structure that allocates managerial power to some members. We show that the organization faces a trilemma: the goals of autonomy, decentralization, and efficiency conflict with one another. Thus, the optimal governance structure of an autonomous organization is centralized. Centralization implies inequality in power and payoffs among members of autonomous organizations.

Keywords: organizations, power, governance, relational contracts, self-executing contracts
JEL classifications: C73, D23, D82, J33

[*] We thank Ulf Axelson, Dan Barron, Kim-Sau Chung, Alessandro De Chiara, Wouter Dessein, Matthias Fahn, Bob Gibbons, Bentley MacLeod, Niko Matouschek, Luis Garicano, Peter Kondor, Radoslawa Nikolowa, Jose Parra-Moyano, Andrea Prat, João Ramos, Scott Schaefer, Ilya Segal, and seminar participants at the BSE Workshop on Organizational Economics, CEPR Paris Symposium, Chinese University of Hong Kong, Hong Kong Baptist University, LME Workshop, London School of Economics, NBER Organizational Economics workshop, OEW 21, SITE (Dynamic Games, Contracts, and Markets), Tsinghua BEAT Workshop, University of Utah, 8th Workshop on Relational Contracts, and Zhejiang University Digital Conference for helpful comments and conversations. We thank Hanzhe Li, Hanmo Wang, and Yunchou Zhang for their excellent research assistance. All remaining errors are ours.

[†] London School of Economics, CEPR and ECGI, d.ferreira@lse.ac.uk.

[‡] HKU Business School, jli1@hku.hk.

1 Introduction

In his reappraisal of Coase's (1937) "The Nature of Firm," Steven N. S. Cheung illustrates the team production problem with the following example:

> *My own favorite example is riverboat pulling in China before the communist regime, when a large group of workers marched along the shore towing a good-sized wooden boat. The unique interest of this example is that the collaborators actually agreed to the hiring of a monitor to whip them.* (Cheung (1983), p. 8).

Riverboat pulling is a collective endeavor in which each member would rather free-ride on the efforts of others. This example illustrates that team members may benefit from the existence of an outside party whose only role is to monitor them.

Third-party monitoring is efficient only if the third party is honest, competent, and inexpensive. In the riverboat example, the whip holder may extract bribes from the workers in exchange for lighter whipping. The monitor could also under-punish shirkers (perhaps due to laziness) or punish the wrong workers. Finally, the monitor could demand too high a fee to perform her duties. In any of these cases, the workers may prefer not to employ a monitor.

This paper studies the problem of incentivizing team members when hiring a third-party monitor is infeasible or undesirable. Specifically, we consider the optimal allocation of power among the members of an *autonomous organization*. The Merriam-Webster Dictionary defines "autonomous" as "having the right or power of self-government." In the riverboat pulling example, self-government means that power (i.e., the whip) must be assigned to a team member or no one.

We present a model of an autonomous organization that produces a common good with individual inputs (i.e., effort) from its members. Members can write contracts specifying effort provisions. However, crucially, the enforcement of such contracts must be carried out by the organization members themselves. Some members have the power to punish members who do not fulfill their contractual obligations. This power is, however, limited by members' right to exit the organization.

The model is as follows. Two ex-ante identical agents contribute individual inputs to the production of a non-excludable good. Because inputs are costly, agents have incentives to underprovide such inputs. External enforcement is not feasible, implying that contracts in autonomous organizations must be relational. Because the game is infinitely repeated, the efficient outcome can be sustained by the threat of exit only if agents are sufficiently patient. If, instead, the discount factor is low, agents play the inefficient static Nash equilibrium.

We augment this canonical relational contract setup in two ways. First, we allow the organization to designate one of the members as a *manager*. In each period, the manager decides how much effort the other player (the *subordinate*) must provide. If the subordinate does not follow the manager's instructions, the manager can punish the subordinate. As in the riverboat pulling example, the manager holds a "whip" that she can use to discipline the subordinate. Because the subordinate always wants to avoid being whipped, when facing an instruction from the manager, she has two options: follow the instruction or leave the organization.

Second, we allow the organization members to agree upon and commit (through a self-executing contract) to a *governance structure*: a set of rules that allocate power (i.e., the whip) to different members contingent on the history of the game. Our problem is the optimal design of such a governance structure. The organization may choose a fully decentralized governance structure (power is spread evenly across members over time), a fully centralized structure (power is assigned to one member forever), or any combination of these polar cases.

We show that an optimal autonomous organization must be fully centralized: the same agent must hold the whip in all periods. This power asymmetry leads to asymmetric payoffs across members. In equilibrium, powerful members (i.e., managers) abuse their power and enjoy higher ex-post payoffs than those without power (i.e., subordinates). This abuse of power takes the form of asking subordinates to "overwork," under the threat of punishment. Because managers benefit from the extra effort that subordinates exert, managers must exert effort themselves to prevent the subordinates from leaving the organization. This is the *Paradox of Power*: by letting the strong party abuse his/her power, the weak party gains power over the strong party via exiting. We show that, in optimal organizations, the strong party does not benefit from becoming more powerful. In contrast, the Paradox of Power implies that the weak party benefits from becoming weaker.[1]

Our main result is what we call the *Organizational Trilemma*: the goals of autonomy, decentralization, and efficiency typically conflict with one another. If we insist on autonomy, there is a trade-off between efficiency and decentralization. If decentralization is a goal in itself, efficiency or autonomy must be sacrificed. If efficiency is the goal, we cannot

[1] This is related to the idea of Judo Strategy: using the strength of your opponent to your own advantage (see, e.g., Yoffie and Kwak (2001)). The Paradox of Power is an extreme example of this force.

have both autonomy and decentralization.

Our model applies to organizations that cannot – or do not want to – rely on external enforcement of contracts. Such organizations may be able to commit – at least partially – to an allocation of power among its members. One such example is the so-called Decentralized Autonomous Organizations (DAOs), which serve as our main motivation and application. The typical DAO is a blockchain-based entity that raises funds from its members and allocates such funds towards a common goal. Members decide on the allocation of funds collectively, typically through voting. DAOs resemble the canonical notion of an autonomous organization because they rely mostly on a combination of self-executing and relational contracts, with little use of externally enforced contracts. DAOs typically seek greater autonomy than traditional organizational forms for two reasons. First, blockchain technology makes designing and implementing self-executing contracts (also called "smart contracts") easier. Second, contract enforcement by outside authorities is often infeasible because they do not have the expertise, information, or recourse to a legal framework for adjudicating disputes.

Self-executing contracts can be very powerful, but their application is limited to "on-chain" actions, i.e., actions that occur on the blockchain where the DAO lives. Anything that requires off-chain actions cannot be fully automated and is thus subject to governance risk. Unlike the idealized vision of DAOs often found in Internet descriptions, real-world DAOs govern off-chain transactions not by code but by relational contracts (i.e., reputation and trust). For example, before an on-chain vote, DAO members often discuss proposals on internet forums (in platforms such as Discord) and conduct rounds of off-chain votes (using tools such as Snapshot).[2] Thus, in real-world DAOs, autonomy is achieved by designing self-executing contracts that complement relational contracts. Self-executing contracts are thus a *governance structure* that supports relational contracts.[3] In our model, the organization uses self-executing contracts to commit to a history-contingent power allocation.

Our results suggest that DAOs face the Organizational Trilemma: A truly decentralized autonomous organization will be inefficient. Our model thus helps us understand many of the practical difficulties encountered by DAOs. As we discuss in the Internet Appendix, real-world DAOs have been plagued by issues such as centralization (and often abuse) of power, lack of contractual enforcement, and poor performance. Our model also shows

[2] Snapshot is a voting platform that allows DAOs built on Ethereum to vote off-chain. Off-chain voting avoids Ethereum transaction fees (called *gas fees*). For more about off-chain discussion and voting, see https://t.ly/dGISZ.

[3] This notion of governance structure is similar to that of Williamson (2002).

that powerful actors have incentives to show restraint and behave benevolently in the early days of an organization, only to abuse their power once the organization is sufficiently mature. Thus, our model offers a cautionary note for participants of blockchain projects with powerful players, such as founders, foundations, core developers, and companies. Trust in blockchain "benevolent dictators" cannot be justified by observing their behavior in the early stages of a project.

In an extension, we show that efficiency can be restored if the organization members could commit to a different output division. Our main result is robust to this extension: Efficiency typically requires the concentration of power in the hands of one player. In addition, the powerful party must also be given most of the output. That is, the party holding the "stick" must also be offered a "carrot." In autonomous organizations, efficiency requires inequality in both power and payoffs.

This paper is related to several strands of literature. Our basic model setup is one of moral hazard in teams à la Alchian and Demsetz (1972), Holmström (1982), and the extensive literature that ensued (see, for example, Bolton and Dewatripont (2004) for a textbook treatment). Unlike Holmström (1982), we focus on solving the moral hazard problem within the team rather than relying on an external enforcer.

This paper is also related to the literature on relational contracting, especially papers that consider how to design relationships to foster cooperation; see, for example, Baker, Gibbons, and Murphy (1994, 2002, 2023), Che and Yoo (2001), Halonen (2002), Kvaloy and Olsen (2006, 2009), Rayo (2007), Mukherjee and Vasconcelos (2011), Deb et al. (2016), Barron and Guo (2021), Fahn and Zanarone (2022), and Troya-Martinez and Wren-Lewis (2023). Among these papers, the most related is Rayo (2007), which studies ownership concentration, examining how ownership structure depends on the observability of the effort. Rayo (2007) shows that when efforts are observable, dispersed ownership emerges to minimize total discretionary bonuses. In contrast, we abstract from monetary transfers and show that power and payoff should remain concentrated even with observable effort. Power and payoff concentration motivate the manager to exert effort.

While the theoretical literature on blockchain economics is large, it mainly focuses on the properties and limitations of specific blockchain protocols. In contrast, our paper focuses on a simple collective action problem, which may or may not live on a blockchain. That is, our autonomous organization is not necessarily a blockchain organization. Despite these fundamental differences, our paper shares similarities with blockchain economics papers that study the limits of decentralization. Biais et al. (2019) present an analysis of the proof-of-work protocol as a repeated game and show the existence of inefficient equilibria with persistent forks. Budish (2023) shows that the cost of sustaining trust in blockchain protocols is prohibitively high. His analysis casts doubt on the ability of autonomous blockchains to deter dishonest behavior without the help of governments or other third parties. Ferreira, Li, and Nikolowa (2023) show that the proof-of-work protocol creates incentives for ownership concentration in the industries that support the mining ecosystem. Han, Lee, and Li (2023) presents a theory of DAO governance based on conflicts between small and large token-holders.

Our paper is also related to the extensive literature on power and authority in organizations; see, for example, Simon (1951), Chwe (1990), Aghion and Tirole (1997), Rajan and Zingales (1998), Piccione and Rubinstein (2007), Van den Steen (2010), Acemoglu and Wolitzky (2011), and Rantakari (2023) (see also Bolton and Dewatripont (2013) for a survey). Most works in this literature study static power allocations. Two notable exceptions are Li, Matouschek, and Powell (2017) and Baker, Gibbons, and Murphy (1999), which, like our paper, examine power in dynamic settings. Unlike our paper, these papers define power as ownership of decision rights and show how informal power emerges persistently from given formal rights. We define power as the ability to enforce behavior through punishment and study the optimal allocation of formal power itself. Our analysis reveals that formal power – like informal power – should be persistent to effectively support relational contracts.

2 Model

We present a model of an autonomous organization. The organization produces a non-excludable good with inputs from its members. The model setup does not try to match the workings of any particular real-world organization. Instead, the model aims to illustrate the fundamental tension between decentralization and efficiency.

2.1 Setup

Consider an organization (to be formally defined later) with two members, called *players*, $i \in \{1,2\}$ (she and he), who interact repeatedly and share a common discount factor, δ. In each period $t \in \{1,2,...,\infty\}$, if both players participate in the organization, they jointly produce output y_t, which is equally shared between them (in Subsection 3.3, we consider an extension with endogenous output shares).[4] At each t, Player i chooses effort $e_{it} \in \{0,1,2\}$, where the cost of effort is $c(e_{it}) = ce_{it}$, $c > 0$. Player i's output is

$$y_{it} = \begin{cases} 0 & \text{if } e_{it} = 0 \\ B & \text{if } e_{it} = 1 \\ B+b & \text{if } e_{it} = 2. \end{cases} \tag{1}$$

Total output is $y_t = y_{1t} + y_{2t}$. All information is public. We assume the following:

Assumption 1. $2c > B > c > b > 0$.

This assumption implies that the first-best effort levels are $e_{it}^{FB} = 1$, for $i \in \{1,2\}$. However, B is not large enough, thus choosing $e_{it} = 1$ is not a dominant strategy in the single-stage game. Given this technology, players can *shirk* ($e_{it} = 0$), *work* ($e_{it} = 1$) or *overwork* ($e_{it} = 2$).[5] For expositional simplicity only, we also assume:

Assumption 2. $B + \frac{1}{2}b \geq 2c$.

This assumption implies that both players earn strictly positive payoffs when one player works and the other overworks, thus making participation constraints easier to meet. Assumption 2 is unnecessary for our analysis and is made only to reduce the number of cases to consider. We will keep it for most of the analysis but drop it for the extension in Subsection 3.3, where it is no longer needed.

We augment this standard relational contracts setup by introducing the concept of *power*. We focus primarily on *autonomous organizations*: power must be allocated to a member of the organization or no one. In Subsection 3.2.5, we also consider non-autonomous organizations in which external contract enforcement is feasible. It is immediate that if external enforcement is costless, efficiency can be achieved. Thus, the interesting case is when the organization must be autonomous, either because external enforcement is costly or because the organization's members derive direct utility from autonomy.

Our notion of power is similar to Simon's (1951) notion of authority.[6] At the beginning

[4] We believe that the case of non-excludable output (i.e., exogenous output shares) is the most relevant in practice. For example, if the organization is a "layer-1" blockchain on which applications are built, any improvements that lead to lower transaction costs will benefit all application developers and users on the upper layers. Due to the public-good nature of such blockchains, it may not be feasible to contractually exclude some parties from using them.

[5] A variation of this technology with continuous effort that delivers similar qualitative results is as follows: $y_{it} = [B + b(e_{it} - 1)]\mathbb{1}_{e_{it} \geq 1}$ and $c(e_{it})$ is a convex cost function where $2c(1) > B > c(1)$ and $c'(e) > b$ for $e > 1$.

[6] Simon (1951) defines authority in the context of an employment relation between a boss (B) and a worker (W): "*We will say that B exercises authority over W if B permits W to select x*" (p. 294).

of period t, one of the players is designated as the *manager*. If Player i is the manager, she recommends an action $\hat{e}_{-it} \in \{0,1,2\}$ for the other player (the *subordinate*) and also an action $\hat{e}_{it} \in \{0,1,2\}$ for herself. The subordinate and the manager then decide whether to exit or stay in the organization. We denote their participation decisions by $d_{it} \in \{0,1\}$, where 0 indicates exit and 1 indicates staying. If the subordinate stays but chooses an effort level different from $\hat{e}_{-it}$, the manager can reduce the subordinate's payoff by $D > B$. If the manager reduces the subordinate's payoff, she obtains a small incremental payoff $\varepsilon \in [0,D)$.[7] For simplicity, we set $\varepsilon = 0$; all results are similar for ε positive but small.

Formally, the manager's identity is given by $g_t \in \{0,1,2\}$, with $g_t = i$ designating Player $i \in \{1,2\}$ as the manager and $g_t = 0$ denoting the case of no manager. We call g_t the *whip*.[8] The manager's identity must be determined at the beginning of each period before other actions are taken. This is because the subordinate must follow the manager's orders. The model's implications would differ if the manager's identity could be determined after the effort decisions. In that case, it could be optimal to randomize the ownership of the whip. However, in such a model, no one would give orders, and we could not interpret whip ownership as managerial power. Thus, our model is relevant to practical situations where the identity of a monitor must be known in advance.

Managerial power – here represented by the whip – is a scarce resource. Accordingly, we assume that only one whip is available. In Subsection 3.2.6, we briefly discuss the case of multiple whips. In reality, there are many practical reasons for managerial power to be concentrated in the hands of a few. In the case of DAOs, the ability to ban or exclude members requires special administrative rights for managing forums or editing the organization's protocol (e.g., signatures). Even in large blockchain projects such as Bitcoin and Ethereum, only a handful of core developers have the keys to modify the blockchain protocol. In addition, if punishment must be visible and sequential, the presence of a second whip would not alter the equilibrium payoffs unless exit is restricted for more than one period.

Whip assignment affects the set of feasible actions: The manager suggests actions for both players $(\hat{e}_{1t}, \hat{e}_{2t})$, while the subordinate does not suggest actions. If no player is the

[7] For example, in proof-of-stake blockchains, a validator who detects some irregularity may seize some of the tokens staked by a block producer.

[8] As in Van den Steen (2010), our notion of power is interpersonal. The manager can request the subordinate to deliver a minimum level of performance. If the subordinate underperforms, the manager can punish the subordinate ex-post. To avoid punishment, the subordinate must either perform according to expectations or exit the organization before the punishment stage.

manager (i.e., $g_t = 0$), no one suggests any action. To keep the space of actions constant, we define

$$e_{it}^a := \begin{cases} (\hat{e}_{1t}, \hat{e}_{2t}) & \text{if } g_t = i \\ \varnothing & \text{otherwise.} \end{cases} \tag{2}$$

We use e_t^a to denote the vector (e_{1t}^a, e_{2t}^a). Similarly, we define $d_t := (d_{1t}, d_{2t})$ and $e_t := (e_{1t}, e_{2t})$. Player i's end-of-the-period payoff is

$$u_{it}(e_t^a, d_t, e_t) = \begin{cases} d_{1t}d_{2t}\left(\frac{y_{1t}+y_{2t}}{2} - ce_{it} - D\mathbb{1}_{e_{it}\neq\hat{e}_{it}}\right) & \text{if } g_t = -i \\ d_{1t}d_{2t}\left(\frac{y_{1t}+y_{2t}}{2} - ce_{it}\right) & \text{if } g_t \neq -i, \end{cases} \tag{3}$$

where $\mathbb{1}_x$ is the indicator function. This payoff function assumes that the manager can commit to punishing the subordinate in case of a deviation from expected play. This commitment is trivially achieved because punishing the subordinate is payoff-neutral to the manager. Our analysis would be qualitatively unchanged if the manager extracted some small payoff $\varepsilon > 0$ from the punishment, perhaps because the manager could seize some bond posted by the subordinate, or if the manager enjoys a psychological benefit from punishing a worker who shirks. Of course, if punishing were costly to the manager, the model would work only if the manager had access to a commitment device.

Within each period t, there are five dates (players choose their actions simultaneously within each date):

Date 1. The outcome of a public randomization device x_t is realized.

Date 2. The whip g_t is assigned to one player ($g_t = 1$ or $g_t = 2$) or no player ($g_t = 0$). Then, players choose e_{it}^a.

Date 3. Players decide whether to exit ($d_{it} = 0$) or stay ($d_{it} = 1$).

Date 4. Players choose $e_{it} \in \{0, 1, 2\}$.

Date 5. Output $y_t \in \{0, .., 4\}$ and payoffs (u_{1t}, u_{2t}) are realized.

The role of x_t is to allow the whip allocation and actions to depend on some publicly observed external signal. The existence of a public randomization device is a common assumption in the repeated games literature and is made to convexify the set of equilibrium payoffs. Without loss of generality, we assume that x_t is uniformly distributed on the unit interval.[9]

We define the *history* at time t as $h^t = \{x_1, g_1, e_1^a, d_1, e_1, ..., x_{t-1}, g_{t-1}, e_{t-1}^a, d_{t-1}, e_{t-1}\}$.

[9] For clarity of exposition, we deviate from the literature and place the public signal at the beginning of the period. The analysis is identical if the public signal is at the end of the period.

We define a *governance structure* as $G = \{G_t\}_{t=1}^{\infty}$, where $G_t : (h^t, x_t) \to (g_t)$. A governance structure maps each history and signal realization to a whip allocation. That is, a governance structure fully determines the allocation of managerial power among players.[10]

We interpret G as a commitment technology. For example, if a manager can take actions that entrench herself in an organization, the governance structure allows this manager to retain power for a long time. Note that g_t can depend on past actions but not current ones. The idea is that (e_t^a, d_t, e_t) are observable but not "contractible." However, players can write public reports on past actions, which then become part of future histories (these reports are credible in equilibrium because players can leave the organization when they see a fake report). Because e_t is not (immediately) contractible, *punishment after a deviation cannot be automated*, thus creating a need for managerial power. Thus, the whip allocation at t is the only meaningful decision that can be automated.[11]

We can also interpret G as a contingent *self-executing contract*. In the DAO interpretation, G completely specifies the conditions under which some DAO members would gain special administrative rights. That is, G is implemented *on-chain*. Because past actions (such as participation in forums) are off-chain, they cannot trigger automated punishment. In a truly autonomous organization, the designated manager would observe off-chain behavior, record it, and then punish those who misbehave.[12]

Let Γ denote the set of all governance structures and $G_0 \in \Gamma$ denote the governance structure such that $g_t = 0$ for all $t \in \{1, ..., \infty\}$. We call G_0 the *default governance structure*. Under the default governance structure, no player has power over the other player; i.e., there is no whip. Let γ_0 denote the game associated with the default governance structure: a set of members, action spaces for each member, and their payoff functions, for the case where $g_t = 0$ always. We call γ_0 the *primitive game*. Because each $G \neq G_0$ is associated with different action and payoff spaces, we can think of G as a particular modification of the primitive game. We call the modified game, $\gamma(G)$, the *game induced by G*.

For given $\gamma(G)$, at each t, we denote player i's (pure) actions by $S_{it} = (e_{it}^a, d_{it}(e_t^a),$

[10] Our notion of governance structure relates to Williamson's (2002) view of governance structure as a set of mechanisms that support an ongoing contractual relationship. See also Baker, Gibbons, and Murphy (2023).

[11] In our setup, the manager must initiate the punishment. That is, punishment does not automatically occur given a state. This is in line with many blockchain projects. For example, in proof-of-stake protocols, block producers and validators must "stake" some of their tokens, which remain frozen for a given period. If a node discovers that some player broke the rules, it can punish that player by "slashing" their stake.

[12] In the DAO interpretation, managerial power is not voting, but instead the administrative rights that can be used to monitor off-chain decisions. In that context, centralization or decentralization must refer to such rights and not to the use of voting.

$e_{it}(e_t^a, d_t))$, where subscript t indicates that the actions are conditional on (h^t, x_t).[13] Player i's strategy is an infinite sequence of such actions, $S_i = \{S_{it}\}_{t=1}^{\infty}$. Let $S = \{S_t\}_{t=1}^{\infty}$ denote a *strategy profile*, where $S_t : (h^t, x_t) \to (e_t^a, d_t, e_t)$ is a mapping from the history at time t to a set of actions for both players. We define $\Psi(G)$ as the set of Subgame Perfect Equilibrium (SPE) strategy profiles for game $\gamma(G)$.

We can now define an organization:

Definition (**Organization**). *An organization is a triplet $\langle \gamma_0, G, S \rangle$ consisting of a primitive game γ_0, a governance structure $G \in \Gamma$, and an equilibrium profile $S \in \Psi(G)$.*

That is, an organization consists of a primitive game, a governance structure that modifies the rules of the primitive game, and a particular suggestion for how the members should play the modified game. We include equilibrium strategies in the definition of organization to allow for soft or intangible aspects, such as culture, to be part of the design of organizations. Because we will keep the primitive game fixed for most of the analysis (the only two exceptions are the non-autonomous organization described in Subsection 3.2.5 and the extension to endogenous output shares in Subsection 3.3), to economize notation, we will often denote an organization simply by $r = \langle G, S \rangle \in \Gamma \times \Psi(G)$.

2.2 Benchmark: The First Best

Let $S_0 \in \Psi(G_0)$ denote an equilibrium under the default governance structure (i.e., an equilibrium of the primitive game). We denote this organization by $r_0 = \langle G_0, S_0 \rangle$. Under G_0, cooperation can be sustained only by the threat of exit. Assumption 1 implies that the first-best effort levels are $e_{1t}^{FB} = e_{2t}^{FB} = 1$. Under the first-best, the normalized average payoff of each player is $B - c$. We restrict attention to equilibria with grim-trigger strategies, in which both players leave if any player deviates from the equilibrium play. That is, if at time t, Player i chooses an off-the-equilibrium-path action, for all $t' > t$ players choose $d_{1t'} = d_{2t'} = 0$. This restriction to grim-trigger strategies is without loss of generality: taking the outside option results in minimax payoffs – the harshest punishment possible – thereby providing the strongest incentive against deviation and facilitating the sustainability of any equilibrium. Under such strategies, the first-best payoffs can be sustained as an SPE if and only if

$$B - c \geq (1 - \delta)\frac{B}{2}. \tag{4}$$

[13] Given public correlation, the restriction to pure actions is without loss of generality.

The left-hand side of (4) is Player i's (normalized) average payoff from working ($e_{it} = 1$) forever, and the right-hand side is the (normalized) value of shirking ($e_{it} = 0$) today followed by the dissolution of the organization. Thus, the first-best can be sustained under the default governance structure if $\delta \geq \frac{2c-B}{B} =: \delta^{FB}$. As a result of the Folk Theorem, any cooperative outcome can be sustained if the discount factor is sufficiently high. The interesting case is $\delta < \delta^{FB}$, which we now assume.

Assumption 3. $\delta < \delta^{FB} := \frac{2c-B}{B}$.

3 Organization Design

In this section, we consider the problem of designing an optimal organization. We start from the primitive game γ_0, which we modify by choosing a governance structure. We then select an equilibrium strategy profile for the modified game. Formally, we consider the problem of a (hypothetical) planner who chooses an organization to maximize the normalized discounted sum of payoffs:

$$\max_{\langle G,S\rangle \in \Gamma \times \Psi(G)} (1-\delta)E\left[\sum_{t=1}^{\infty} \delta^{t-1}\left(u_{1t}+u_{2t}\right) \mid G,S\right]. \tag{5}$$

The economic interpretation is that the (benevolent) organization designer chooses a set of immutable rules, here summarized by G. These rules are enforced automatically, e.g., they are embedded in the organization's code. Our problem is to determine the set of rules that maximizes the organization's surplus, assuming that the designer also selects the best SPE associated with such rules. Alternatively, we could assume that the designer chooses only the governance structure while players coordinate on the surplus-maximizing equilibrium.[14]

Our focus on optimal organizations allows us to simplify the setup without any loss of generality. First, from now on, we restrict the space of feasible whip allocations to $\{1,2\}$, except for the case of the default governance structure, in which case we set $g_t = 0$ always. To see that this restriction is without loss, consider an organization such that, for some (h^t, x_t), we have $g(h^t, x_t) = 0$. Let $e(h^t, x_t)$ denote the associated equilibrium efforts.

[14] The existence of an "organization designer" should not be taken literally; it is just a metaphor for the forces that drive an organization towards efficiency. It could be conscious design, efficient bargaining, competition, or adaptation. Of course, if the organization's founders design it to maximize their own payoffs, in some cases, they might not solve (5). Thus, we interpret our results as an upper bound on the efficiency of an organization.

Suppose instead that we set $g(h^t,x_t) = 1$. It is immediate that by setting the manager's announcement to $e_1^a(h^t,x_t) = e(h^t,x_t)$, the effort vector $e(h^t,x_t)$ can be sustained as an equilibrium under this new governance structure. Because nothing changes in all other periods, this equilibrium is payoff-equivalent to the original one. Thus, from the organization designer's perspective, there is no reason to choose $g_t = 0$ following any realization of (h^t,x_t).

Second, we only consider organizations such that, in equilibrium, if i is the manager, $e_i^a(h^t,x_t) = e(h^t,x_t)$, for all (h^t,x_t). In other words, the manager always recommends the equilibrium effort levels. It is easy to see that any equilibrium in which $e_i^a(h^t,x_t) \neq e(h^t,x_t)$ is payoff-equivalent to an equilibrium that differs from the original one only by setting $e_i^a(h^t,x_t) = e(h^t,x_t)$. This simplification implies that we can ignore e_t^a when characterizing an equilibrium.

Third, because we will consider only grim trigger strategies, the optimal participation decision is $d_{it} = 1$ unless a player has deviated in the previous period, in which case the optimal decision is $d_{it'} = 0$ for all $t' \geq t$.

With these simplifications, we can think of the organization as an institution consisting of a (possibly rotating) manager and a subordinate. The manager makes decisions concerning productive efforts. The subordinate either carries out the manager's instructions or leaves the organization. The effort choices (or orders) depend only on (h^t,x_t), where the history is now more succinctly described as $h^t = (x_1,e_1,...,x_{t-1},e_{t-1})$. Any given strategy S_i for $i \in \{1,2\}$ can now be described by a sequence of effort functions $e_{it} = e_i(h^t,x_t)$ and participation decisions $d_{it} = d_i(h^t,x_t)$.

3.1 Stationary Organizations

In this subsection, we consider the case of stationary autonomous organizations. Stationarity helps with the intuition for the results. It also simplifies the analysis; the main results in this subsection are useful for solving the general case of optimal nonstationary autonomous organizations, which we present in the following subsection.

We first consider stationary organizations under the default governance structure. We say that an organization $\langle G_0,S_0\rangle$ is *stationary* if the equilibrium actions (on the equilibrium path) in period t are independent of the history, h^t. Under the default governance structure, we have the following result.

Proposition 1 (Equilibrium under Default Governance). *If $S_0 \in \Psi(G_0)$ is a stationary equilibrium, then $e_{1t} = e_{2t} = 0$ for all t.*

Proposition 1 shows that (under Assumption 3) the unique stationary equilibrium of the primitive game is such that both players shirk. That is, there is a unique *default organization* $r_0 = \langle G_0, S_0 \rangle$ where $e_{1t} = e_{2t} = 0$ always.[15]

We now consider an autonomous organization with $G \neq G_0$. For this organization to be stationary, we also require its governance structure to be i.i.d.:

$$g_t = g(x_t) := \begin{cases} 1 & \text{if } x_t \leq p \\ 2 & \text{if } x_t > p \end{cases}, \tag{6}$$

where $p \in [0,1]$ is the probability that Player 1 has the whip at any given t (recall that we have restricted the space of whip allocations for $G \neq G_0$ to $\{1,2\}$). Thus, under stationarity, we can fully describe a governance structure by p. In a stationary organization, we can write Player i's equilibrium effort decision as $e_i(h^t, x_t) = e_i(x_t)$ and define $e(x_t) := (e_1(x_t), e_2(x_t))$. The formal definition of stationarity is as follows.

Definition (Stationarity). *An organization is stationary if its governance structure is stationary and the equilibrium effort profile $e(x_t)$ is independent of h^t.*

Note that conditional on g_t, effort can be stochastic through its dependence on x_t. For future use, we define $\mu_i(g_t)$ as the probability distribution of e_{it} over $\{0,1,2\}$ conditional on g_t.

We define the *centralization index* of an organization with a stationary governance structure as $C(p) := |2p-1|$. Stationarity implies that the centralization index is constant across periods, histories, and strategy profiles. An organization with a stationary governance structure is fully decentralized if $p = 0.5$ and fully centralized if $p = 1$ or $p = 0$. For brevity, when considering fully centralized organizations, we focus only on the case in which Player 1 is the manager ($p = 1$); the case in which $p = 0$ is exactly symmetrical.

The following lemma shows a link between decentralization and shirking.

Lemma 1 (Shirking Lower Bound). *In a stationary organization, the player with the (weakly) lower payoff shirks when he is a manager.*

Lemma 1 implies that if $C(p) < 1$ (i.e., the organization is not fully centralized), some players will not work whenever they are managers. Specifically, the player with

[15] This result generalizes to nonstationary organizations as well, but the proof is rather tedious.

the (weakly) lower payoff must shirk when he is the manager. To see why, assume that Player 2 has a lower payoff than Player 1. Player 2's payoff must be strictly lower than the first-best payoff ($B-c$). Because $\delta < \delta^{FB}$, Player 2's loss in future payoff is smaller than the gain from saving the cost of effort. It is thus impossible to induce Player 2 to exert effort unless there is additional punishment for not doing so. But once Player 2 is the manager, no further punishment can be imposed on him. As a result, Player 2 must shirk whenever he is the manager.

For a stationary organization with governance p, Lemma 1 implies a *shirking lower bound*: in any equilibrium, the probability that at least one player shirks at any given t is no lower than $\min\{p, 1-p\} \equiv \frac{1-C(p)}{2}$. Note that the shirking lower bound is decreasing in the centralization index. This result illustrates the cost of decentralization: in more decentralized organizations, the shirking lower bound is tighter.

We now introduce two special organizational structures. First, we say that a stationary autonomous organization is *identity-blind* if effort levels do not depend on players' identities:

Definition (Identity-blindness). *A stationary organization is identity-blind if effort choices are such that $\mu_i(1) = \mu_{-i}(2)$ for $i \in \{1,2\}$.*

That is, the organization is identity-blind if the conditional effort distributions $\mu_1(g_t)$ and $\mu_2(g_t)$ are symmetric. Second, we say that a stationary organization is *power-blind* if a player's identity alone determines his/her effort choice.

Definition (Power-blindness). *A stationary organization is power-blind if effort choices are such that $\mu_i(1) = \mu_i(2)$ for $i \in \{1,2\}$.*

Identity-blindness and power-blindness are different types of symmetry with respect to the "flipping" of a power allocation. In an identity-blind organization, when the power allocation flips, the effort choices of the players also flip. In a power-blind organization, the effort choices remain unchanged when the power allocation flips. As the next result shows, if a stationary autonomous organization is either power-blind or identity-blind, then no player can be induced to work.

Proposition 2 (Symmetry leads to shirking). *If a stationary autonomous organization with $C(p) < 1$ is either power-blind or identity-blind, then $e_1(x_t) = e_2(x_t) = 0$ for all x_t.*

To see why this result holds, start with the power-blind case. Suppose Player 2 has a (weakly) lower payoff than Player 1. Proposition 1 implies that Player 2 shirks when he has the whip. Power-blindness then implies that Player 2 always shirks. Thus, Player 1 also shirks when she has the whip because her continuation payoff is insufficient to induce her to work. Again, power-blindness implies that Player 1 always shirks. Next, consider the identity-blind case. Proposition 1 implies that (say) Player 2 shirks when he is the manager. Identity-blindness then implies that both players shirk whenever they become managers. When managers don't work, the total payoff is lowered to such an extent that it is impossible to induce any worker to work.

Proposition 2 implies that no one works unless changes in whip assignments differentially affect players' behavior. That is, at least one player must change behavior when the whip changes hands, and if both players do so, such changes cannot be symmetric. In other words, one player must work harder than the other, either as a manager or as a subordinate.

The next result establishes a necessary and sufficient condition for managers not to shirk in equilibrium.

Lemma 2 (Managers' Incentive Constraint). *In a stationary organization in which $g(x_t) = i$ for some x_t, $e_i(x_t) > 0$ can be enforced if and only if Player i's equilibrium payoff is*

$$u_i \geq \frac{1-\delta}{\delta}\left(c - \frac{B}{2}\right) =: \underline{u}. \tag{7}$$

Lemma 2 implies that, in an equilibrium where some managers work, their payoff must be at least $\underline{u}$. From now on, we assume (without loss of generality) that Player 1 has the (weakly) higher payoff in equilibrium. If Player 1 works when she is the manager, she must obtain no less than $\underline{u}$ in equilibrium, which (from Assumption 3) implies that she should receive more than half of the joint payoff created. Thus, in any stationary organization where managers work, *the equilibrium must be asymmetric*: Player 1 must have a strictly higher payoff than Player 2.

The next result shows that abuse of power must occur in any stationary organization that improves upon the default organization.

Lemma 3 (Equilibrium Abuse of Power). *Consider a stationary organization that improves upon the default organization. Suppose Player* 1 *has the highest equilibrium payoff.*

1. *Whenever Player 1 shirks, Player 2 must also shirk.*
2. *Player 2 must overwork with strictly positive probability when he is a subordinate.*

This result implies that optimality requires Player 1 to abuse her power as manager by asking the subordinate to overwork with some probability. However, for that to happen, Player 1 must also exert some effort when she is the manager. That is, there is a nonzero-measure set of public signals for which $g(x_t) = 1$ and $e(x_t) = (e, 2)$ with $e > 0$.

Let $R(p)$ denote the set of all stationary organizations with governance p. An organization $r \in R(p)$ is *p-optimal* if

$$r \in \arg\max_{r \in R(p)} E\left[u_1 + u_2 \mid r\right]. \tag{8}$$

That is, a p-optimal organization maximizes the total payoff for a given governance structure p. Let $u_i(p)$ denote Player i's payoff and $u(p) = u_1(p) + u_2(p)$ the total payoff under a p-optimal organization. The next result constrains the set of enforceable effort profiles in p-optimal organizations.

Lemma 4 (Enforceable Effort Profiles). *In a p-optimal organization, the only effort profiles that can be played with positive probability are* $(0,0)$*,* $(1,0)$*,* $(1,1)$*, and* $(1,2)$.

The next result shows how payoffs change as centralization increases.

Proposition 3 (The Paradox of Power). *Suppose* $u(p) > 0$ *and let* $p' > p$*. Then,* $u_1(p') = u_1(p) = \underline{u}$ *and* $u_2(p') > u_2(p)$.

Proposition 3 shows that, in a p-optimal organization, Player 1 does not benefit from increasing her power; only Player 2 does. We call this phenomenon the *Paradox of Power*. Intuitively, if power is sufficiently centralized in Player 1's hands, her continuation payoff is high. Player 2 can thus use the threat of exiting to induce Player 1 to work. It is optimal for Player 1 to work whenever possible. Thus, any increase in Player 1's payoff above the minimum required in (7) is used to induce her to work more often. All benefits of increased centralization accrue to the weaker party.

Figure 1 illustrates the Paradox of Power. For expositional simplicity, we consider the case in which the equilibrium involves effort profiles $(1,2)$, $(1,0)$ and $(0,0)$ only. Assume also (for expositional simplicity) that given g_t, the players must play pure actions (i.e., mixing between actions is not allowed; the proof of Proposition 3 makes none of these simplifying assumptions). The figure shows the three relevant payoff profiles on the $u_1 \times u_2$ plane, where $u_i = (1-\delta)\sum_{t=1}^{\infty} E\delta^{t-1}u_{it}$. Suppose an equilibrium involves playing effort profiles $(1,2)$ and $(1,0)$ with probabilities p and $1-p$, where $p > 0.5$ is the probability that Player 1 is the manager. If the expected payoff profile is at point I, Player 1's expected payoff is $u_1 = p(B + \frac{b}{2} - c) + (1-p)(\frac{B}{2} - c) = \frac{1-\delta}{\delta}(c - \frac{B}{2}) = \underline{u}$. Thus, for this p, Player 1 can be induced to work when she is a manager (see Lemma 2), and this is an equilibrium.

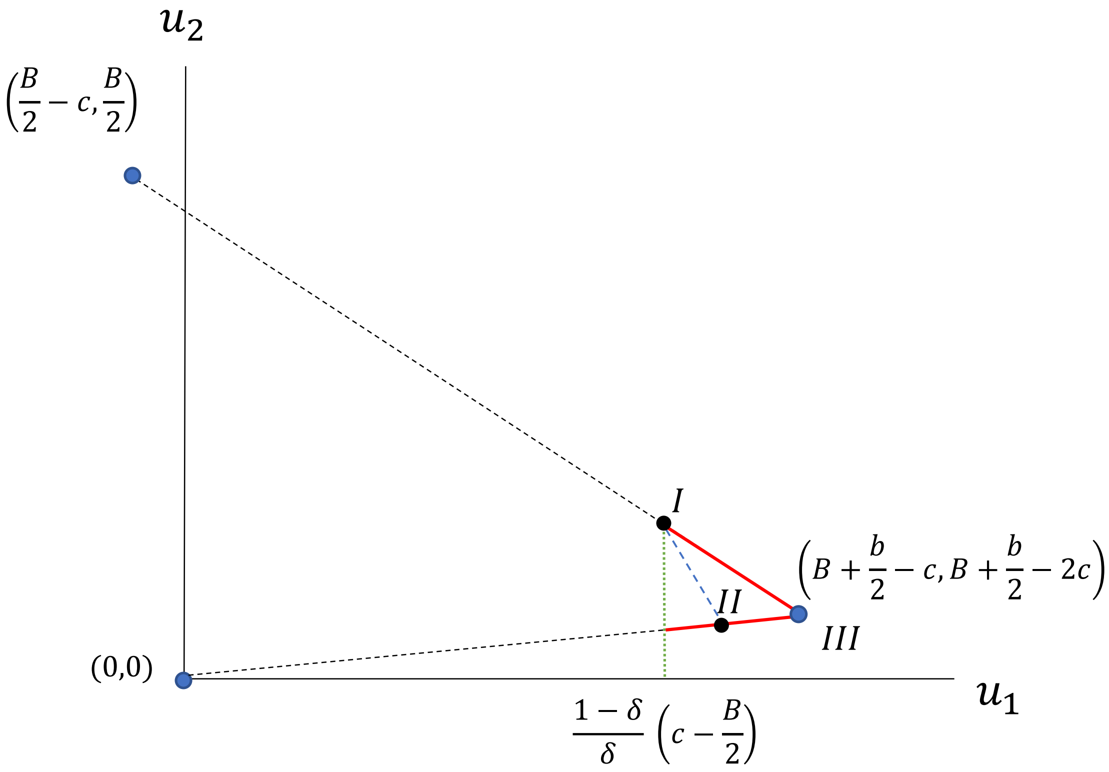

Figure 1: The Paradox of Power

For the same p, consider another equilibrium that involves $(1,2)$ and $(0,0)$ instead. For this equilibrium, the expected payoff profile is at point II. Player 1 is better off at II. However, the sum of payoffs (u_1+u_2) is higher at I. Note that Player 2 – the weaker party – is also better off at I. If Player 1 were less powerful (i.e., lower p), point I would not be enforceable, and the optimal organization (under the assumptions of this example) would be (nearly) at point II, which gives Player 1 a higher utility. Thus, Player 1 can be made worse off by having more power. Intuitively, because more power increases Player 1's continuation payoff, it eventually becomes possible to coordinate on an equilibrium where Player 1 always works. In that equilibrium, Player 2 gains power over Player 1 by threatening to exit in case Player 1 does not work.

This example illustrates why Player 1 may not benefit from having more power. Because of the restriction to pure actions, the equilibrium in this example is not necessarily p-optimal. Proposition 3 shows that, like the example, Player 1 does not benefit from having more power in a p-optimal organization. Unlike the example, in a p-optimal organization, u_1 is independent of p. In such an organization, it is optimal to randomize between $(1,2)$ and $(1,1)$ when $g(x_t) = 1$. In this case, an increase in p increases only u_2. Thus, the Paradox of Power is that making the strong party more powerful benefits only the weaker party.

The following proposition characterizes optimal stationary organizations.

Proposition 4 (Optimal Stationary Organizations). *A stationary organization that improves upon the default organization exists if and only if $\delta \geq \delta_1 := \frac{2c-B}{B+b}$ and is unique. The optimal stationary organization r^* must be fully centralized ($p = 1$) and randomize between effort profiles $(1,2)$ and $(1,1)$ with probabilities α^* and $1-\alpha^*$, where*

$$\alpha^* := \frac{2}{b}\left(\underline{u} - (B-c)\right).$$

Proposition 4 shows that, for sufficiently high δ (but still lower than δ^{FB}), one can design an organization that improves upon the default organization. Furthermore, in this case, the optimal organization is fully centralized. In equilibrium, Player 1's "abuses her power" with probability $\alpha^* > 0$. Thus, Player 1's payoff is higher than Player 2's payoff. Despite this asymmetry, both players would agree that this equilibrium is preferable to the default organization, which delivers zero payoff to both players. Player 2 is incentivized to overwork due to fear of being whipped. Player 1 works only because of her continuation value. Centralization is critical here because it efficiently delivers payoff asymmetry, which is necessary for providing sufficient continuation value to Player 1.

Figure 2 illustrates Proposition 4. Suppose that $\delta \in [\delta_1, \delta^{FB})$. Let $p = 1$ and suppose an equilibrium randomizes between action profiles $(1,2)$ and $(1,1)$ with probabilities α and $1-\alpha$. If the expected payoff profile is at point IV, Player 1's expected payoff is $u_1 = \alpha(B + \frac{b}{2} - c) + (1-\alpha)(B-c) = \frac{1-\delta}{\delta}(c - \frac{B}{2})$, which is the expression that defines α^*. Any payoff profile on the $III-IV$ line can be sustained for some $\alpha \geq \alpha^*$. Note that the total sum of the payoffs is maximized at point IV, implying that a fully centralized organization (i.e., $p = 1$) where the manager works in every period, and the subordinate randomizes between "work" and "overwork" with probabilities α^* and $1-\alpha^*$ is an optimal stationary organization.

Discussion. Proposition 4 shows that under full centralization, for sufficiently high $\delta < \delta^{FB}$, there exist equilibria in which neither player shirks. In such equilibria, the manager works in every period, while the subordinate alternates between working and overworking. By overworking, the subordinate increases the value of the relationship for the manager and induces her to work. One key feature of Proposition 4 is that, even if the players are ex-ante

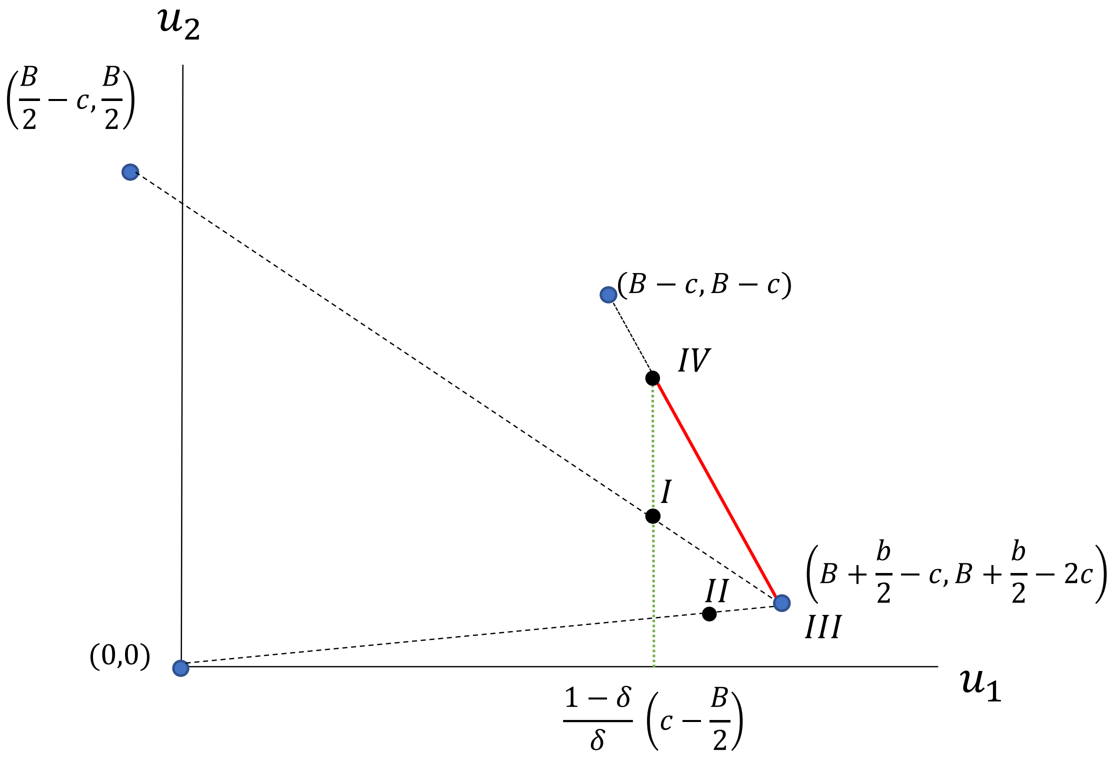

Figure 2: Implementable payoffs under full centralization

identical, the optimal stationary organization involves a hierarchical division of labor. At the top is the manager, who commands and controls the subordinate. Notably, there is no switching between roles: Once a player becomes manager, she holds the whip forever and becomes a designated enforcer.

It is a foundational feature of organizations to have designated managers – someone specializing in observing and enforcing the behavior of others. In their seminal article, Alchian and Demsetz's (1972) explain this feature by highlighting the role of the owner-manager in measuring the workers' effort. Holmström (1982) formalizes this idea and shows how moral hazard in teams can be solved by employing an external enforcer. However, few models endogenize the choice of internal enforcers. In all prominent relational contracting models that consider internal enforcers, enforcement happens via exit: The manager can only enforce the behavior by threatening to leave the relationship (see, e.g., Rayo (2007)). Because all players can threaten to leave, in a sense, every player is an enforcer.

In our model, a manager enforces behavior by using her power, not the threat of exiting.

In his classic study of industrial bureaucracy, Gouldner (1954) presents an example of how a foreman controls the behavior of his subordinates: "*If I catch a man goofing off, I tell him in an a,b,c, way exactly what he has to do, and I watch him like a hawk 'til he's done it*" (Gouldner (1954), p. 159). While it is unclear why "watch him like a hawk" does the job, this managerial activity enforces behavior. Our model innovates by explicitly introducing this managerial activity in team-production problems, using the whip as a metaphor. Whoever carries out the managerial activity can enforce the other player's behavior.

Our model does not assume that the same person always acts as a manager. In principle, democratizing managerial activity could be a good idea. However, Lemma 1 shows that rotating managerial responsibilities is suboptimal because at least one of the managers will shirk. To maximize the organization's performance, the optimal governance structure requires the same player to carry out the managerial activity in each period. This creates a managerial *position*.

Another feature of the optimal organization is that the manager has more than half of the surplus. Therefore, the possession of power also gives the manager a higher payoff level. This payoff asymmetry arises because, without a high enough payoff to the manager, she will not put in the effort. In other words, because managers hold the "stick," "carrots" must also be offered to motivate them.

One may think that by rotating power, players could potentially discipline those who abuse their power when managers. Thus, why doesn't power rotation improve efficiency? The reason is that the only way of motivating a whip holder is to offer them sufficient continuation value. With equal output shares, abuse of power is necessary for providing sufficient continuation value to a manager. In other words, while power rotation can prevent abuse of power, *it is not efficient to do so*. Similarly, rotating the whip to punish those who shirked in previous periods does not work because players may leave between periods.

3.2 Optimal Organizations: The General Case

We now consider the general organization design problem without imposing stationarity. In a nonstationary organization, the equilibrium payoffs may change as play evolves, implying that following some history, the joint payoff may be lower than the ex-ante maximal joint payoff. In other words, the optimal equilibrium is not necessarily sequentially optimal. Consequently, knowledge about suboptimal equilibrium play can be useful in solving for the optimal organization.

We solve this problem using the recursive method developed by Abreu, Pearce, and Stacchetti (1990). This method focuses on characterizing the set of equilibrium payoffs rather than the equilibrium actions. Once the equilibrium payoff set is known, we can use it to derive the optimal equilibrium strategies and governance structures. This is done as a step-by-step process. For each equilibrium payoff, we find the equilibrium actions and governance structures associated with it. We then find the continuation payoffs associated with the equilibrium actions and, for the continuation payoffs, we find the associated governance structure and the equilibrium actions, and so on.

3.2.1 Preliminary Results

Here we generalize several of the results from the stationary case. Because the proofs in this subsection are straightforward variations of those for the parent results, for brevity we present them in the Internet Appendix. The next result generalizes Lemma 1.

Lemma 5 (Shirking Lower Bound: General). *At any (h^t, x_t), the player with the (weakly) lower continuation payoff shirks when he is a manager.*

Note that Lemma 5 implies Lemma 1: in stationary organizations, payoffs are independent of time, thus the continuation payoffs are the same as the equilibrium payoffs. The intuition is the same as before: a manager needs sufficient continuation payoff to work. This result implies that, unless an organization is fully centralized, shirking will occur with a strictly positive probability in equilibrium. The next result generalizes Lemma 2.

Lemma 6 (Managers' Incentive Constraint: General). *In an organization in which $g(h^t, x_t) = i$ for some (h^t, x_t), effort $e_i(h^t, x_t) > 0$ can be enforced if and only if Player i's continuation payoff is $u_i(h^t, x_t) \geq \underline{u}$.*

This result implies that, at any history and public signal realization, the designated manager must expect to receive more than half of the future total payoffs if she is to exert positive effort. Note that Lemmas 5 and 6 apply to any organization, not just to optimal ones. The next result shows the necessity of overworking.

Lemma 7 (Equilibrium Abuse of Power: General). *Consider an organization that improves upon the default organization.*

1. *If the player with the highest continuation payoff at (h^t, x_t) shirks, then the other player also shirks at (h^t, x_t).*

2. *At least one player must overwork with positive probability in equilibrium.*

3.2.2 Equilibrium Payoff Frontier

To solve for the optimal equilibrium, it suffices to characterize the equilibrium payoff frontier: Player 2's maximal equilibrium payoff for a given Player 1's payoff. Specifically, let u_1 denote Player 1's equilibrium (normalized) payoff. The equilibrium payoff frontier, which we denote as $f(u_1)$, is the value function of the following constrained maximization problem:

$$f(u_1) = \max_{\langle G,S\rangle \in \Gamma \times \Psi(G)} (1-\delta)E\left[\sum_{t=1}^{\infty} \delta^{t-1} u_{2t} \mid G,S\right] \tag{9}$$

$$\text{s.t. } (1-\delta)E\left[\sum_{t=1}^{\infty} \delta^{t-1} u_{1t} \mid G,S\right] = u_1. \tag{10}$$

We need to characterize only the equilibrium payoff frontier because it has a "self-generating" property: for any payoff pair on the frontier, its continuation payoff pair (the expected discounted payoffs of the players in the next period) will again stay on the frontier along the equilibrium play. Applying the self-generating property repeatedly shows that the continuation payoffs of the optimal equilibrium play remain on the frontier forever. Therefore, knowledge about the equilibrium payoff frontier is sufficient to describe the optimal governance structure and the equilibrium play.[16]

The standard method to characterize the equilibrium payoff frontier is to solve a functional equation (the Bellman equation). Doing so, however, is unwieldy in our setting because the equation would include many possible actions and governance structures. Instead, we solve for the equilibrium payoff frontier by deriving an upper bound and then showing that this upper bound can be supported as an equilibrium payoff. It thus follows that the upper bound is the equilibrium payoff frontier.

In this subsection, we restrict the analysis to discount factors in $[\delta_1, \delta^{FB})$. We show that this restriction is without loss of generality in Subsection 3.2.4. We have the following result:

Proposition 5 (Equilibrium Payoff Frontier). *For $\delta \in [\delta_1, \delta^{FB})$, the following holds.*

1. *$f(u_1)$ is symmetric along the 45-degree line.*

[16] The self-generating property arises because the players can publicly observe the subordinate's actions. When it is publicly known that the subordinate has carried out the equilibrium action, there's no need to punish him by reducing his payoff below the equilibrium payoff frontier. Therefore, the continuation payoffs associated with the optimal equilibrium play will stay on the frontier again.

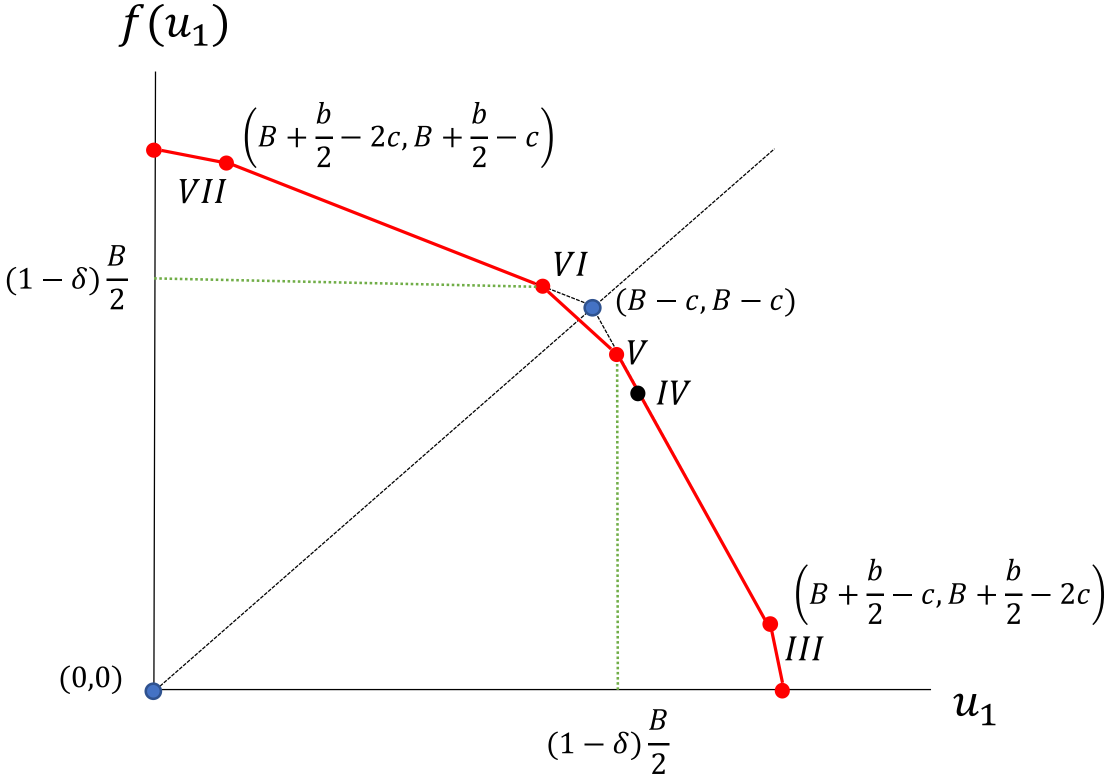

Figure 3: Equilibrium Payoff Frontier

2. *For $u_1 \geq \frac{1}{2}(1-\delta)B$, $(u_1, f(u_1))$ is on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$.*

3. *For $u_1 \in (f(\frac{1}{2}(1-\delta)B), \frac{1}{2}(1-\delta)B)$, $f(u_1)$ is a negative 45-degree line.*

Figure 3 illustrates the equilibrium payoff frontier. Part 1 of Proposition 5 shows that the payoff frontier is symmetric along the 45-degree line. This arises naturally because the roles of Player 1 and Player 2 are identical in our setup. Part 2 shows that for $u_1 \geq \frac{1}{2}(1-\delta)B$, the equilibrium payoff frontier is the line segment $III-V$. Point V is given by $\frac{1}{2}(1-\delta)B$ on the x-axis and $f(\frac{1}{2}(1-\delta)B) = B-c-\lambda(c-\frac{b}{2})$ on the y-axis, where $\lambda = \frac{2c-(1+\delta)B}{b}$ (point VI is the mirror image of V). That is, point V can be reached by requiring both players to work $(B-c, B-c)$ with probability $1-\lambda$ and, with probability λ, Player 1 to work and Player 2 to overwork $(B+\frac{b}{2}-c, B+\frac{b}{2}-2c)$. This line segment is also part of the feasible payoff frontier of the stage game. As a result, it is an upper bound for all equilibrium payoffs. Part 2 then shows that the equilibrium frontier reaches this upper bound.

Notice that part of the line segment $III-V$ can also be reached under the optimal stationary organization (Proposition 4). Figure 2 illustrates that the optimal stationary equilibrium (under full centralization) can reach the feasible payoff frontier for $u_1 \geq \frac{1-\delta}{\delta}(c-\frac{B}{2})$, which is the line segment $III-IV$. Part 2 of Proposition 5 shows that the feasible payoff frontier can be further extended to the left of IV to $u_1 = \frac{1}{2}(1-\delta)B$, which is the minimal payoff Player 1 must receive to sustain the first-best outcome under the default governance structure. This extension increases the total payoff. The further Player 1's payoff is to the left, the more often both players choose to work (rather than Player 1 working and Player 2 overworking), increasing the joint payoff. Part 2 then implies that, for $u_1 \in [\frac{1}{2}(1-\delta)B, \frac{1-\delta}{\delta}(c-\frac{B}{2})]$, the efficiency of the relationship can be improved by using a nonstationary equilibrium strategy, as we will show below.

Part 3 shows that when $u_1 \in (f(\frac{1}{2}(1-\delta)B), \frac{1}{2}(1-\delta)B)$, the equilibrium payoff frontier is a negative 45-degree line (the line segment $V-VI$). The joint payoff is the same anywhere on the frontier, and it is sustained by randomization between points V and VI. The necessity of randomization in this region has the same logic as that in stationary organization design. To induce both players to work, asymmetry in the payoffs is needed: one player must receive at least $\frac{1}{2}(1-\delta)B$. It is immediate from Figure 3 that the optimal organization must implement payoffs on the line segment $V-VI$. Such points can only be reached through randomization at $t=1$, when one of the players is chosen to receive payoff $\frac{1}{2}(1-\delta)B$. Without loss of generality, let this player be Player 1, so that $u_1^* = \frac{1}{2}(1-\delta)B$ and $u_2^* = f(\frac{1}{2}(1-\delta)B)$.

3.2.3 Centralization in Nonstationary Environments

So far, we have only defined centralization in stationary environments. In nonstationary environments, centralization may depend on patience. For example, an organization that alternates the whip between the two players every period is, intuitively, very decentralized. However, if the whip allocation rotates, say, every thousand periods, the extent to which this organization is decentralized is unclear. In such an organization, the player who has the whip earlier may use their power to extract more valuable benefits. Thus, we need some form of discounting power allocations over time. For organization r, we define the centralization index at (h^t, x_t) as

$$C(h^t, x_t, r) = (1-\eta)\left|\sum_{j=t}^{\infty} \eta^{j-t} E\left[\mathbb{1}_{g_{j+1}=1} - \mathbb{1}_{g_{j+1}=2} \mid h^t, x_t, r\right]\right|, \tag{11}$$

where $\eta \in (0,1)$ is an (arbitrary) "power discount factor." Function (11) is a generalization of our centralization index to nonstationary organizations. The centralization index ranges from zero (*full decentralization*) to one (*full centralization*), and it may vary across histories for the same organization. Note that, if $g_t = 1$ (or $g_t = 2$) always, the centralization index is 1. Conversely, if g_t is either 1 or 2 with equal probabilities for all t, the centralization index is zero. Thus, even without knowledge of η, we can use the centralization index to determine whether an organization is fully centralized or decentralized. For all other governance structures, the centralization index depends on the power discount factor. Lower η implies a lower tolerance to the "persistence of power:" governance structures with infrequent whip alternations are deemed to be more centralized. We treat η as an exogenously given parameter reflecting a commonly agreed definition of centralization. None of our results depends on η.[17]

3.2.4 Optimal Organizations

We now describe the optimal governance structures and equilibrium strategies. Our first result establishes the necessary conditions for an organization to improve upon the default organization.

Proposition 6 (Optimal Organization: Existence). *An organization that improves upon the default organization exists if only if $\delta \geq \delta_1$.*

This result shows that dropping the stationarity constraint does not increase the range of discount factors that sustain organizations with positive total payoff. The reason for this surprising result is that stationarity in subgames at $t > 1$ is not a binding constraint on the designer's maximization problem. Because the condition that determines δ_1 depends on Player 1's continuation payoff, imposing stationarity in the continuation game does not affect this condition.

Proposition 6 also implies that the restriction to $\delta \in [\delta_1, \delta^{FB})$ in Proposition 5 is without loss of generality; if $\delta < \delta_1$, it is not possible to improve upon the default organization and the payoff frontier collapses into a single point, $(0,0)$. The next result shows that optimality requires centralization.

[17] We condition the centralization index on (h^t, x_t) and measure centralization one period ahead. If we instead conditioned it on h^t only and measured centralization from time t, we could make any nonstationary organization fully decentralized at $t = 1$ by flipping a fair coin to decide who is the inaugural manager at $t = 1$. This is true even if this manager holds on to power forever.

Proposition 7 (Optimality Requires Centralization). *If $\delta \in [\delta_1, \delta^{FB})$, an optimal organization must be fully centralized: If $g(h^1, x_1) = 1$, then $g(h^t, x_t) = 1$ for all (h^t, x_t).*

This proposition shows that full centralization is necessary for optimality. The intuition is the same as in the stationary case: Because Player 1 has the largest equilibrium payoff, Player 2 would shirk as manager, which implies that Player 1 should always have the whip. Thus, optimal organizations – stationary or nonstationary – must be fully centralized *at any point* (h^t, x_t).

The optimality of centralization is more surprising for nonstationary organizations. In principle, there could be an additional force to discipline a manager – the threat of losing the whip in the future if she shirks today. However, this threat is moot because the threat of exit by the subordinate is immediate and, thus, more effective.

Another generalizable result is the Paradox of Power. Here we present a version of this paradox that is independent of η (the power discount factor). Similar to the stationary case, we say that an organization is *G-optimal* if it maximizes the total payoff for governance structure G. Let $u_i(G)$ denote Player i's payoff and $u(G) = u_1(G) + u_2(G)$ the total payoff under a G-optimal organization. Define $R(h^1, x_1)$ as the set of all organizations for which $g(h^1, x_1) = 1$ and $e(h^1, x_1) = (1, 1)$. The next proposition shows that moving from a not-fully-centralized organization to full centralization benefits only the weaker party.

Proposition 8 (The Paradox of Power, Revisited). *Consider two G-optimal organizations, r^c and r^d, with governance structures G^c and G^d, respectively. Let $r^c, r^d \in R(h^1, x_1)$, and suppose $C(h^1, x_1, r^c) = 1$ and $C(h^1, x_1, r^d) < 1$. Then, $u_1(G^c) \leq u_1(G^d)$, and $u_2(G^c) > u_2(G^d)$.*

As in the stationary case, the Paradox of Power follows from the payoff asymmetry implied by Lemma 6. Because condition (7) is a constraint on the maximization problem, it must be binding whenever possible. Because this constraint binds when the organization is fully centralized, only the weaker party (the one with the lowest equilibrium payoff) can benefit from a move to full centralization.

How do optimal nonstationary organizations look like? There are multiple equilibrium profiles that maximize the joint payoff. Thus, there are multiple optimal organizations. Despite this multiplicity, all optimal organizations share similar economic properties: they are fully centralized, deliver the same payoffs, begin with effort profile $(1, 1)$, and play profile $(1, 2)$ with positive probability for some $t > 1$. The next proposition characterizes one of these organizations.

Proposition 9 (**Populist Dictator**). *For $\delta \in [\delta_1, \delta^{FB})$, the following organization is optimal:*

1. *Player 1 is the manager in each period.*
2. *Both players choose to work in the first period until some (random) period T. From period T+1 on, Player 1 works, and Player 2 overworks.*

The structure of the optimal equilibrium play in Proposition 9 is akin to that of deferred compensation in optimal dynamic contract design. The payoff structure backloads the payoff of Player 1, and this avoids inefficient actions at the beginning of the relationship. Despite this similarity, the reason for backloading is somewhat different. The logic here is that of rent extraction because, for any stationary relational contract that requires overworking with positive probability, we can increase its efficiency by reducing the manager's payoff.

The optimal backloading of Player 1's payoffs is a property of all optimal organizations, and not just that described in Proposition 9. In particular, take any optimal stationary equilibrium. We can modify it by asking Player 2 to work with probability 1 in the first period and keep the rest of the equilibrium play unchanged. This modification remains an equilibrium. It increases the joint payoff of the players while reducing Player 1's payoff.

The rent-extraction logic implies that the relationship dynamics in this model differ from similar models of perfect information in the literature. When there is perfect information, the efficiency of the relationship improves over time (see Albuquerque and Hopenhayn (2004), Thomas and Worrall (2018), and Barron et al. (2022)). In contrast, the efficiency of the relationship in this model decreases over time.[18] To sustain efficiency in the earlier periods of the relationship, Player 2 overworks in the long run. As a result, while the manager's payoff increases over time, the worker's payoff decreases, and the efficiency of the relationship deteriorates.

The equilibrium described in Proposition 9 implies that the powerful party will refrain from abusing her power until time T. That is, the powerful party behaves as if she had no power in the early days of an organization. Such an apparent benevolence eventually disappears; after T, the powerful party begins asking the subordinate to overwork. The

[18] The performance of the relationship can decrease over time or cycle when there is private information; see, for example, Clementi and Hopenhayn (2006), Padro i Miquel and Yared (2012), Li and Matouschek (2013), Li et al. (2017), and Li et al. (2023). When the players can discover new production possibilities, however, it is possible that the performance of the relationship improves over time (Chassang (2010)).

equilibrium thus displays a form of "populism:" power is centralized in the hands of one agent, who initially behaves like a benevolent dictator, only to eventually show her true colors and abuse her power by forcing the subordinate to undertake inefficient actions. As the populist's mask comes off, the organization becomes less efficient.

The next corollary shows how $E[T]$ – the expected period of transitioning from managerial benevolence to abuse of power – depends on δ.

Corollary 1 (**Patience and Abuse of Power**). *In the equilibrium described in Proposition 9, $E[T]$ is increasing in δ.*

This result is very intuitive: If the manager is patient, she will work early in the game even if the opportunity to abuse her power comes only in the distant future.

3.2.5 The Organizational Trilemma

Autonomous organizations must enforce contracts or promises internally by allocating power to some members. By contrast, a non-autonomous organization may choose (or be forced) to allocate the whip to a third party, such as courts, regulators, or independent arbitrators. Suppose now that an unbiased third party exists; call it Player 3. Player 3 is not a member of the organization; thus, she cannot exert effort or enjoy a share of the output. If Player 3 observes the output, she can still play an essential role by promising to punish those players who shirk. We interpret Player 3 as an unbiased arbitrator (or court) that enforces the formal contracts written between Players 1 and 2.

Formally, consider an alternative primitive game, γ_0', that is identical to γ_0 except for a third player with no productive actions and constant zero payoff. Let the governance structure be $g(h^t, x_t) = 3$ for all (h^t, x_t). That is, the third party always has the whip. Consider an equilibrium where the third party expects both other players to exert the first-best effort level. The third party punishes any deviation by reducing the payoff of the deviating party by D (players can still avoid punishment by exiting at Date 3). It is easily seen that the first-best payoffs can be sustained as an SPE for any discount factor δ. Thus, a non-autonomous organization finds it easier to deliver efficient outcomes than an autonomous organization. Note that, in this example, the non-autonomous organization is also decentralized, in the sense that no organization member has power over one another. Thus, a non-autonomous organization can achieve efficiency under full decentralization.

One issue with this analysis is the assumption that the third party is honest, competent and inexpensive. A primary motivation for autonomy is a lack of trust in institutions. To capture this idea, suppose that, with probability ρ, the third party destroys the output y_t. This output destruction could be due to corruption (e.g., it is paid as a bribe), incompetence (e.g., excessive regulation), or the cost of the system (e.g., taxes to pay for the legal system). Now, a non-autonomous organization can sustain the first-best effort profile only if $(1-\rho)B \geq c$. Even in that case, the first-best payoffs can no longer be attained. We have the following result.

Proposition 10 (Non-autonomous Organization). *A non-autonomous organization where an external monitor holds the whip and destroys the output with probability ρ is optimal if and only if $2\rho B < \min\left\{2(B-c), \frac{B}{2}(1-\delta) - c - \lambda(c - \frac{b}{2})\right\}$, where $\lambda = \frac{2c-(1+\delta)B}{b}$.*

The following corollary – implied by Propositions 7 and 10 – summarizes the trilemma of decentralization, autonomy, and efficiency.

Corollary 2 (The Organizational Trilemma). *For $\delta \in \left[\delta_1, \delta^{FB}\right)$, we have the following tradeoffs.*

1. ***(Decentralization + Autonomy → Inefficiency).*** *A fully decentralized autonomous organization is suboptimal.*
2. ***(Autonomy + Efficiency → Centralization).*** *An optimal autonomous organization is fully centralized.*
3. ***(Decentralization + Efficiency → Non-autonomy).*** *An optimal organization is decentralized only if it is non-autonomous.*

The Organizational Trilemma implies that decentralized autonomous organizations must be inefficient unless players are sufficiently patient (i.e., Assumption 3 does not hold). To maximize the joint surplus, the organization must either become *fully* centralized or give up its autonomy. The latter option may also be inefficient because a third-party monitor may be expensive, dishonest or incompetent.

3.2.6 When Power is Not a Scarce Resource

In our analysis so far, we have assumed that power is a scarce resource, in the sense that only one player can have the whip at any point in time. This assumption is reasonable in many institutional settings. Here, we briefly consider the case in which the organization can assign whips to several players simultaneously. The case of multiple whips has a very simple solution if players can punish each other simultaneously. It is easy to see that, if both players have whips, the first-best effort profile can be implemented, even in a one-shot game. Having multiple whips is akin to decreasing the players' outside payoffs or their effort costs, so that cooperation is always achieved. In such a case, the optimal governance structure is trivially decentralized. Thus, our model is relevant in situations where power must be allocated only to a subset of the members of an organization. That is, power must be a scarce resource because of technological, institutional, or incentive constraints.

3.2.7 When There are $n > 2$ Players

While our analysis focuses on two players ($n = 2$), our findings readily generalize to settings with multiple players ($n > 2$). In the larger group, when the discount factor falls below the first-best threshold, a symmetric allocation of power and payoff results in the shirking of the manager (who has the whip). To induce the manager to work, she must receive a higher payoff and more power, reinforcing the idea that total payoff increases in power and payoff asymmetry. This higher payoff of the manager takes the form of abuse of power, yet with $n > 2$, there is greater flexibility in deciding which subordinates overwork. Applying the same reasoning, we can also establish the paradox of power and that centralization of power is optimal when $n > 2$.

3.3 Centralization and Pay Inequality: Endogenous Output Shares

In this section, we consider an extension of the model in which players can ex ante agree on a division of output such that Player i receives output shares ω_{it} at t, where $\omega_{it} \in [0,1]$ and $\omega_{1t} + \omega_{2t} = 1$. As with the whip allocation, players can commit to a contingent share rule $\Omega = \{\Omega_t\}_{t=1}^{\infty}$, where $\Omega_t : (h^t, x_t) \to (\omega_t)$. We keep the timing within each period as before: the output shares at t must be determined by (h^t, x_t); they cannot depend on e_t or y_t.[19] Because the organization is autonomous, Ω must be self-executing (i.e., hard-wired in the organization's technology). An organization is now defined as $\langle \gamma_0, G, \Omega, S \rangle$.

Under the default governance structure, it is easy to see that $\omega_{it} = 0.5$ for all t is weakly optimal. That is, if the whip is not available, the ability to choose an asymmetric share rule does not improve the organization's surplus. Things are different when the designer

[19] Allowing output shares to depend on y_t or e_t is identical to automating punishment, which amounts to solving the problem of autonomy by assumption, rendering the problem trivial and uninteresting. G and Ω may still depend on past effort levels indirectly, because players may "write reports" about the past actions they observed and feed these to the self-executing contracts that govern G and Ω.

can optimally choose a governance structure. The combination of an asymmetric output division and a whip allocation significantly expands the set of enforceable effort profiles. The next result describes the optimal organization (the proof is long and thus shown in the Internet Appendix).[20]

Proposition 11 (Optimal Organization with Endogenous Shares). *If $2B \geq 3c$, for any δ, an optimal organization sustains the first-best effort levels $(e_1, e_2) = (1,1)$. If $2B < 3c$, there exists $\delta^c \in (0, \delta^{FB})$ such that an optimal organization sustains the first-best effort levels $(e_1, e_2) = (1,1)$ if and only if $\delta \geq \delta^c$. In addition,*

1. *A fully centralized optimal organization exists for any $\delta \geq \delta^c$.*
2. *A fully decentralized optimal organization exists only if $\delta \geq \delta^d > \delta^c$.*

Proposition 11shows that, if $2B \geq 3c$, the first-best can be sustained regardless of the continuation payoffs. In that case, the organization can assign the whip and a higher output share to one player, so that this player works even in a one-shot game, while the subordinate works to avoid being whipped. When this is possible, we can implement the first-best with an organization that is fully decentralized, fully centralized, or anything in between.

Proposition 11 also shows that the optimality of centralized governance is robust to making output shares endogenous. It applies to all cases, stationary or not. It shows that whenever the first-best can be sustained, it can be sustained by a fully centralized organization. In contrast, the conditions for the optimality of fully decentralized organizations are more stringent than those for fully centralized ones, implying that centralization strictly dominates decentralization for some parameters, but not vice versa. The intuition here is the same as in the case of exogenous shares. An optimal organization must have asymmetric payoffs. The whip allocation helps with this goal. With endogenous shares, payoff inequality can also be created by an asymmetric share rule. Thus, for sufficiently high discount factors, it might be possible to implement the first-best with a fully decentralized organization. But governance decentralization is always an obstacle to payoff asymmetry.

We have the following result.

Corollary 3 (Asymmetric Share Rules). *Consider a fully centralized organization where Player i is the manager. If this organization implements the first-best effort levels $(1,1)$, it must have an asymmetric share rule: $\omega_{it} \geq \frac{c}{(1+\delta)B} > \frac{1}{2}$ for all t.*

[20] Here we no longer restrict the parameter set by Assumption 2. We made that assumption for simplicity of exposition only; it is unnecessarily too restrictive in the current application.

Corollary 3 shows that, in fully centralized organizations, managers should always be given a greater share of the profit. That is, to achieve efficiency, the organization needs to offer one of the players more power *and* a larger share of the output. This result echoes the famous monitor-as-owner result in Alchian and Demsetz (1972), formalized elegantly by Holmström (1982, 1999) in a moral hazard in teams model, and by Rayo (2007), who incorporates relational contracts in the model. In Holmström (1982, 1999), the owner is an external enforcer of contracts. In Rayo (2007), the owner also contributes to production and needs to be motivated.[21] In these models, the emergence of the owner requires the workers' efforts to be private information. When the efforts are publicly observed, Holmström (1999) shows that an external enforcer is unnecessary. And Rayo (2007) shows that dispersed ownership is optimal. The dispersed ownership result also arises in our model: equal output sharing is optimal under the default governance structure.

Proposition 11 shows that a single player typically arises as the owner-manager, even if efforts are publicly observable. We depart from existing relational contracting models by introducing a managerial activity: using the whip to enforce desired behaviors. Our model then shows that the optimal governance structure is to put this managerial activity in the hands of a single player. To motivate this designated manager, a larger share of the output should also go to her, making her the "owner." As mentioned earlier,"carrots" must be given to the player who holds the "sticks."

We end this section by connecting our model to the property rights theory of the firm à la Grossman and Hart (1986). This connection is inspired by Holmström (2016), which views Grossman and Hart (1986) as a theory of markets. Holmström writes

> "*...unlike Marx, who thought the primary purpose of the firm's concentrated power was to exploit the workers, in the logic of the (extended) property rights theory the purpose is to allow the firm to design more efficient organizational structures, in situations where markets function poorly, because of externalities (such as excessive hold-ups). Firms are far from democratic, but they are kept in check by the fact that workers, who own their human capital, can go elsewhere if they are unhappy—another illustration of the effectiveness of property rights.*"

[21] The emergence of the owner-manager in Rayo (2007) also depends on the heterogeneity in the observability of effort across the players. In his model, the ownership choice reduces the total (non-enforceable) bonus by taking advantage of the differences in the marginal rates of substitution between using the contractible output and the non-contractible signals of efforts to motivate the players.

Grossman and Hart (1986) shows how asset allocation keeps firms in check. However, firms also design organizational structures on top of the asset allocation. Our model shows that the way to keep the owner/manager in check is by giving her a higher share of the surplus. This, of course, enriches and creates inequality. However, without doing so, the threat of the worker's exit would not be effective.

4 Conclusion

In a relational contracts setup, we consider the optimal allocation of power among the members of an autonomous organization. We show that the goals of autonomy, decentralization, and efficiency conflict. This organizational trilemma results from the need for payoff asymmetry to incentivize the member who holds power.

Our model shows that, in the absence of external enforcement, self-executing contracts can be used to support relational incentives. Thus, the availability of self-executing contracts improves the performance of autonomous organizations. The flip side is that to realize such gains, an autonomous organization must use self-executing contracts as a tool for centralizing power. Centralization implies inequality in power and payoffs among members of autonomous organizations.

While centralization suggests that power rests solely with the manager, this isn't the case, as subordinates hold power through their right to exit. If the manager fails to perform, a subordinate may choose his outside option in the future. Giving more power to the manager enhances the credibility of the subordinate's exit threat and, paradoxically, strengthens the subordinate's influence over her behavior.

Our model connects naturally to Alchian and Demsetz's (1972) famous statement that power in firms and markets is fundamentally the same, as it stems from the ability to exit and withhold future business. Our analysis shows that the effectiveness of exit threats in providing incentives depends crucially on the proper allocation of power within firms. As Holmström (2016) points out, firms frequently adjust these power distributions through changes in job descriptions, reporting structures, and task assignments. These organizational adjustments not only serve strategic purposes but also influence broader incentive systems, shaping the behavior of both those who wield formal authority and those subject to it. Our paper provides a simple framework that serves as a foundation for understanding changes in organizational design.

A Proofs

Proof of Proposition 1. Denote the players' expected equilibrium payoffs as u_i for $i \in \{1,2\}$. Without loss of generality, assume that $u_1 \geq u_2$. Then, because $u_1 + u_2 \leq 2(B-c)$, we know that $u_2 \leq B-c$. Stationarity implies that $e_{2t} = e_2$, i.e., effort is independent of t. There are two cases to consider: either $e_2 = 1$ or $e_2 = 2$. When $e_2 = 1$, the incentive constraint is $(1-\delta)\left(c - \frac{B}{2}\right) \leq \delta u_2$. Because $u_2 \leq B-c$, the constraint implies that $\delta \geq (2c-B)/B = \delta^{FB}$, which contradicts Assumption 3, implying $e_2 \neq 1$. Assumption 1 implies that the payoff from a deviation is larger when $e_2 = 2$, thus Player 2 also deviates if he is required to choose $e_2 = 2$. Thus, $e_2 = 0$. Trivially, $e_1 = 0$ because $\frac{B}{2} - c < 0$ and $\frac{B+b}{2} - 2c < 0$. □

Proof of Lemma 1. Without loss of generality, assume that $u_1 \geq u_2$. Consider a period in which Player 2 has the whip. Using the same arguments as in the proof Proposition 1, we can show that Player 2 shirks. Thus, Player 2 always shirks when $g_t = 2$. □

Proof of Proposition 2. Without loss of generality, assume $u_2 \leq u_1$. Then, by Lemma 1, Player 2 shirks whenever he has the whip. Suppose a stationary organization is power-blind. Power-blindness implies that Player 2 must also shirk when he does not have the whip. Because Player 2 always chooses $e_2 = 0$, and $B/2 < c$, Player 1 will never choose $e_1 = 1$. Similarly, because $(B+b)/2 < 2c$, she will never choose $e_1 = 2$. It follows that $e_1 = 0$.

Next, suppose a stationary autonomous organization is identity-blind. Recall that Player 2 shirks when he has the whip (because $u_2 \leq u_1$). Now, suppose that he does not have the whip. If he is forced to put in $e_2 = 1$, his participation constraint is given by

$$(1-\delta)\left(\frac{B}{2} - c\right) + \delta u_c \geq 0,$$

where u_c is Player 2's continuation payoff. Because $u_2 \leq u_1$ implies $u_c \leq B-c$, this constraint requires $\delta \geq (2c-B)/B = \delta^{FB}$, which is a contradiction. If she is forced to put in $e_2 = 2$, his participation constraint is given by

$$(1-\delta)\left(\frac{B+b}{2} - 2c\right) + \delta u_c \geq 0.$$

Because we must have $u_c \leq B+b-2c$, this constraint requires $\delta \geq (4c-B-b)/(B+b) > \delta^{FB}$. This, again, is a contradiction. Because Player 2 shirks, identity-blindness implies that Player 1 also shirks. □

Proof of Lemma 2. Player i's IC constraint for working when x_t is

$$(1-\delta)\left(\frac{B}{2}+\frac{y_{-it}}{2}-c\right)+\delta u_i \geq (1-\delta)\frac{y_{-it}}{2}. \tag{A.1}$$

where y_{-it} is the output of Player $-i$ (i.e., not i). Rearranging yields (7). Thus, if (7) holds, $e_i(x_t)=1$ can be enforced when $g(x_t)=i$. Because the IC constraint for $e_i(x_t)=2$ implies (A.1), if either $e_i(x_t)=1$ or $e_i(x_t)=2$ is enforceable, (7) must hold. □

Proof of Lemma 3. Consider a stationary equilibrium and let (u_1,u_2) be the expected payoffs. Let $u_1 \geq u_2$. Then, by Lemma 1, Player 2 does not exert effort when he has the whip, i.e., if $g(x_t)=2$, then $e_2(x_t)=0$. Suppose Player 2 exerts positive effort at some x_t, where $g(x_t)=1$, but Player 1 shirks. Player 2's participation constraint requires $(1-\delta)(\frac{y_{2t}}{2}-ce_2)+\delta u_2 \geq 0$, which doesn't hold because $u_2 < B-c$ (Player 2 has the lowest payoff, and the maximum joint payoff is less than $2(B-c)$). Thus, we conclude that when Player 2 exerts nonzero effort, Player 1 must also exert nonzero effort ⇒ if Player 1 shirks, Player 2 must also shirk (Part 1).

If Player 2 never exerts effort, Player 1 can obtain a maximum surplus of $\frac{B}{2}-c<0$. Thus, Player 1 does not exert effort if Player 2 always shirks. For the organization to improve upon r_0, Player 1 must then exert nonzero effort for some (nonzero measure) set of public signals $H \in [0,1]$, and Player 2 must exert nonzero effort for some (non-zero measure) subset of H. Because Player 2 shirks as manager, Player 2 must exert nonzero effort for some $x_t \in H$ where $g(x_t)=1$.

Suppose $g(x_t)=1$ and $e_1(x_t)>0$ for some $x_t \in H$. Then, (7) implies that Player 1 must have $u_1 > B-c$ (using Assumption 3). Given that Player 2 shirks when manager, if Player 2 chooses $e_2(x_t) \leq 1$ for all x_t, the maximum payoff Player 1 can get is $p(B-c)$, where p is the probability that Player 1 is the manager in a stationary organization. Thus, (7) is violated, implying that Player 1 shirks. We conclude that Player 2 must overwork with strictly positive probability when subordinate (Part 2). □

Proof of Lemma 4. If Player 1 works when $g(x_t)=1$, Player 2 participation constraint for working is $(1-\delta)(B-c)+\delta u_2 \geq 0$, which holds because $B>c$. If Player 2 overworks instead of working, his participation constraint is $(1-\delta)(B+b/2-2c)+\delta u_2 \geq 0$, which again holds because of Assumption 2. Thus, there is no need for Player 1 to choose the inefficient action (overwork) when $g(x_t)=1$, because Player 2's behavior can be enforced if Player 1 works instead. There is also no need for Player 1 to overwork when $g(x_t)=2$ because Player 2 cannot be induced to work when manager and, when subordinate, his continuation payoff is irrelevant for his participation constraints. Thus, we conclude that effort profiles of the type $(2,e)$ are not part of a stationary equilibrium that maximizes the total payoff for a given p. Lemma 3 rules out profiles $(0,1)$ and $(0,2)$. Thus, the only remaining profiles are $(0,0)$, $(1,0)$, $(1,1)$, and $(1,2)$. □

Proof of Proposition 3. Lemma 4 implies that there are only four effort profiles that can be enforced by a p-optimal organization that improves upon r_0. Of these, Lemma 1 implies that profiles $(1,2)$ and $(1,1)$ can only be played if $g(x_t) = 1$, and $(1,2)$ must be played with some positive probability.

Clearly, it is not optimal to play profiles $(0,0)$ or $(1,0)$ with positive probability when $g(x_t) = 1$ because they can be replaced with $(1,1)$, which increases total payoffs and slacks Player 1's incentive constraints without affecting Player 2's participation constraints, as these are not binding when $(1,1)$ is played. Thus, let α denote the probability that profile $(1,2)$ is played when $g(x_t) = 1$, and $1-\alpha$ the probability of $(1,1)$. When $g(x_t) = 2$, let $(1,0)$ be played with probability β and $(0,0)$ with probability $1-\beta$. We can then write the payoffs as

$$u_1 = p\left(B - c + \alpha\frac{b}{2}\right) + (1-p)\beta\left(\frac{B}{2} - c\right)$$

$$u_2 = p\left(B - c + \alpha(\frac{b}{2} - c)\right) + (1-p)\beta\frac{B}{2}$$

$$u = u_1 + u_2 = p\left(2(B-c) + \alpha(b-c)\right) + (1-p)\beta\left(B-c\right).$$

Note that u decreases with α and increases with β, but u_1 increases with α and decreases with β. Player 2's participation constraints hold for any α and β (because $B + \frac{b}{2} - 2c > 0$ by Assumption 2). Thus, in a p-optimal organization, α and β are chosen to maximize u subject to Player 1's incentive constraint in (7). A solution must exist because the objective function is continuous in α and β and the set of (α,β) defined by (7) is compact. Because u_1 is linear in both α and β, (7) must bind. This implies that, in a p-optimal organization, $u_1(p) = \underline{u}$ and, thus, independent of p. Because u is strictly increasing in $p \in (0.5, 1]$ for any $\alpha,\beta \in [0,1]$, (by the Envelope Theorem) $u(p)$ strictly increases with p. Thus, since $u_1(p) = \underline{u}$, $u_2(p)$ must strictly increase with p. □

Proof of Proposition 4. Continuing from the proof of Proposition 3, note that u_1 is maximized at $p = \alpha = 1$. Thus, from Lemma 2, a necessary condition for Player 1 to work as manager is

$$B-c+\frac{b}{2}\geq\frac{1-\delta}{\delta}\left(c-\frac{B}{2}\right)=\underline{u},$$

which requires $\delta\geq\delta_1:=\frac{2c-B}{B+b}$, proving the necessity part.

To prove sufficiency, we first characterize the optimal stationary organization when $\delta\geq\delta_1$. The total payoff is

$$u_1+u_2=p\left(2(B-c)+\alpha(b-c)\right)+(1-p)\beta\left(B-c\right),$$

which is strictly increasing in p for any $\alpha,\beta\in[0,1]$. Player 1's payoff u_1 is also increasing in p for any $\alpha,\beta\in[0,1]$, which makes the incentive condition (7) easier to meet. Thus, increasing p increases total payoffs while slacking Player 1's incentive constraint (Player 2's participation constraints are unaffected), which implies that the optimal organization must be fully centralized, i.e., $p^*=1$.

Setting $p^*=1$, the total payoff decreases with α (β is irrelevant in this case). Thus, the optimal α is the lowest level that meets Player 1's incentive constraint:

$$B-c+\alpha\frac{b}{2}\geq\frac{1-\delta}{\delta}\left(c-\frac{B}{2}\right)=\underline{u}, \tag{A.2}$$

which then leads to

$$\alpha^*=\frac{2}{b}\left[\frac{1-\delta}{\delta}\left(c-\frac{B}{2}\right)-(B-c)\right]=\frac{2}{b}\left(\underline{u}-(B-c)\right). \tag{A.3}$$

Because $(1-\delta)/\delta$ decreases in δ and $2c>B$, α^* decreases in δ. To see that $\alpha^*\leq1$, evaluate (A.3) at δ_1 to find that it is 1. To see that $\alpha^*\geq0$, evaluate (A.3) at δ^{FB} to find that it is 0. Thus, if $\delta\geq\delta_1$, we can always choose the (unique) organization r^* where $p=1$ and $\alpha=\alpha^*$, so that the organization improves upon r_0. By construction, r^* is the stationary organization with the highest total payoff. □

Proof of Proposition 5. Part 1 is straightforward due to the symmetric structure of the game.

For Part 2, notice that the line segment between $(B-c,B-c)$ and $(B+\frac{1}{2}b-c,B+\frac{1}{2}b-2c)$ is a subset of the *feasible* payoff frontier of the stage game. Therefore, if a payoff pair on this line segment is attainable, it must coincide with $(u_1,f(u_1))$.

The proof of Proposition 4 shows that any payoff pair $(u_1,f(u_1))$ such that $u_1\geq\underline{u}$ can be enforced by a stationary SPE. Consider the point on the line segment where $u_1=\underline{u}$ (this is point *IV* in Figures 2 and 3). Proposition 4 implies that *IV* is sustained by the following equilibrium: Player 1 always holds the whip, and the effort profile randomizes between

$(1,2)$ and $(1,1)$ with probabilities α^* and $1-\alpha^*$. Thus, all payoff profiles on the line between points *III* and *IV* are part of the equilibrium payoff frontier.

The equilibrium payoff frontier can be extended to the left of point *IV* on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$. Without loss, suppose Player 1 is the manager at (h^t, x_t) and the equilibrium payoffs at (h^t, x_t) are on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$. To be on that line, no Player can shirk at (h^t, x_t).[22] Thus, Lemma 6 implies that Player 1's continuation payoff $u_1(h^t, x_t)$ must be weakly greater than $\underline{u}$ (otherwise she would shirk).

Let u_{min} denote the minimum equilibrium payoff to Player 1 at (h^t, x_t), conditional on: (i) her being the manager, (ii) nobody shirking, and (iii) her continuation payoff being on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$. Because nobody shirks, the minimum payoff to Player 1 happens when $e(h^t, x_t) = (1,1)$ and her continuation payoff is $u_1(h^t, x_t) = \underline{u}$. Thus,

$$u_{min} = (1-\delta)(B-c) + \delta \cdot \frac{1-\delta}{\delta}(c - \frac{B}{2}) = \frac{1}{2}(1-\delta)B.$$

Note that $(u_{min}, f(u_{min}))$ is on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$, because $(u_{min}, f(u_{min}))$ is a convex combination of $(B-c, B-c)$ and $(\underline{u}, f(\underline{u}))$, with weights $1-\delta$ and δ, respectively. Using Proposition 4, we can see that $(u_{min}, f(u_{min}))$ can be sustained by the following nonstationary SPE: Player 1 always holds the whip, and the action profile is $(1,1)$ in period 1 and then it randomizes between $(1,2)$ and $(1,1)$ with probabilities α^* and $1-\alpha^*$ from period 2 on. Player 1's IC constraint at $t=1$ is satisfied with equality and Player 2's participation constraint at $t=1$ also trivially holds. Thus, $(u_{min}, f(u_{min}))$ is on the frontier (this is point *V* in Figure 3, which is to the left of *IV*).

We now show that the frontier cannot be extended to the left of point *V* on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$. By construction, u_{min} is the minimum equilibrium payoff to a working manager if no one shirks. Thus, to the left of point *V* on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$, we either need (i) at least one shirking player or (ii) Player 2 to be the manager. If (i), the equilibrium payoff must place some weight on payoffs where at least one of u_1 or u_2 is zero or negative. But then the equilibrium payoff cannot be on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$, which is a contradiction. If (ii), Player 2's continuation payoff

[22] More rigorously, players cannot shirk for a set of (h^t, x_t) that occurs with strictly positive probability in equilibrium.

must be at least u_{min}, implying that $u_2 \geq u_{min} > B-c$. But the maximum u_2 on the line segment between $(B-c, B-c)$ and $(B+\frac{1}{2}b-c, B+\frac{1}{2}b-2c)$ is $B-c$, implying $u_2 \leq B-c$, which is a contradiction. We conclude that the frontier cannot be extended to the left of point V. This completes the proof of Part 2.

Note: The equilibrium payoff frontier may extend to the right of point *III* (as shown in Figure 3). We note that this region is irrelevant for finding the optimal organization because it is dominated by point *III*.

To prove Part 3, it is sufficient to show that no payoff pair can achieve a joint surplus greater than $\frac{1}{2}(1-\delta)B + f(\frac{1}{2}(1-\delta)B)$. Suppose to the contrary that this is not the case. Let (u_1', u_2') denote a payoff pair that maximizes the joint surplus. Without loss of generality, we assume that (u_1', u_2') is sustained by a pure action profile. We first show that (u_1', u_2') must be sustained by action profile $(1,1)$. To see this, decomposing u_1' and u_2' into current payoffs and continuation payoffs leads to

$$u_1' = (1-\delta)u_1(e_1', e_2') + \delta u_{1,c}', \text{ and } u_2' = (1-\delta)u_2(e_1', e_2') + \delta u_{2,c}',$$

where (e_1', e_2') is the action profile in period 1, and $u_{1,c}'$ and $u_{2,c}'$ denote the continuation payoffs. Notice that, because (u_1', u_2') maximizes the joint surplus, we have $u_{1,c}' + u_{2,c}' \leq u_1' + u_2'$. Therefore,

$$u_1(e_1', e_2') + u_2(e_1', e_2') \geq u_1' + u_2'.$$

Also note that, because playing the effort profile $(1,2)$ forever is an equilibrium, we have $u_1' + u_2' > 2B + b - 3c$, where the latter is the payoff sustained by the effort profile $(1,2)$. Since only effort profile $(1,1)$ gives a higher payoff than $(1,2)$, we then must have $(e_1', e_2') = (1,1)$.

However, because (u_1', u_2') is between points VI and V in Figure 3, we know that both u_1' and u_2' must be smaller than $\frac{1}{2}(1-\delta)B$. In this case, regardless of how the whip is allocated, one of the two players would prefer to shirk. This is a contradiction. □

Proof of Proposition 6. Suppose first that all managers shirk. The maximum total payoff is then $B-c$ (which happens only when all subordinates work). A subordinate's participation constraint when he works implies $(1-\delta)\left(c-\frac{B}{2}\right) \leq \delta u_i(h^t, x_t)$. Because $u_i(h^t, x_t) \leq B-c$, the constraint implies that $\delta \geq (2c-B)/B = \delta^{FB}$, which contradicts Assumption 3. Thus, in equilibrium, subordinates would also shirk if managers always shirk. It follows that to improve upon r_0, there must exist (h^t, x_t) where the manager does not shirk.

At any given (h^t, x_t), to obtain the maximum continuation payoff for the manager (say, Player *i*), the other player must overwork with probability 1 for all $t' > t$. But then Player *i* will always have the highest continuation payoff for all $t' > t$, and from Lemma 7, Player *i* must at least work in $t' > t$ for the other player to overwork in $t' > t$. Thus, the highest possible continuation payoff for a manager is $B + \frac{b}{2} - c$ (that is, Player *i* works and Player $-i$ overworks). From Lemma 6, a necessary condition for a manager not to shirk is

$$B - c + \frac{b}{2} \geq \frac{1-\delta}{\delta}\left(c - \frac{B}{2}\right) = \underline{u},$$

which requires $\delta \geq \delta_1 := \frac{2c-B}{B+b}$, proving the necessity part. Sufficiency is implied by Proposition 4. □

Proof of Proposition 7. An optimal organization must deliver total payoff $u_1 + u_2 = (1-\delta)\frac{B}{2} + f((1-\delta)\frac{B}{2})$, which is the total payoff at frontier points V or VI (or a convex combination of both). Without loss of generality, suppose that Player 1 is the manager at (h^1, x_1). Because all players must not shirk at any time to sustain equilibrium payoffs at points V or VI, Player 1 must not shirk at (h^1, x_1). For Player 1 not to shirk, her continuation payoff must be at least $\underline{u}$ (see Lemma 6). That is, the continuation payoff profiles at (h^1, x_1) must lie on the frontier points between III and IV. Such continuation payoff profiles require that only effort profiles $(1,1)$ and $(1,2)$ are played with positive probability at $t > 1$. In such a case, we must have $u_1(h^t, x_t) > u_2(h^t, x_t)$ for all $t > 1$. Lemma 5 then implies that Player 2 would shirk if he is a manager at any (h^t, x_t) for all $t > 1$. To avoid shirking, the whip should then be given to Player 1 for all $t > 1$. That is, optimality requires full centralization. □

Proof of Proposition 8. Write the total payoff of a G-optimal organization $r \in R(h^1, x_1)$ as

$$u(G) = (1-\delta)2(B-c) + \delta(u_1(h^1, x_1, G) + u_2(h^1, x_1, G)), \tag{A.4}$$

where $u_i(h^1, x_1, G)$ is Player *i*'s continuation payoff. Because r^c is fully centralized and G^c-optimal, Proposition 7 implies that r^c must also be unconstrained optimal, implying it must sustain the payoffs at point V in Figure 3. Thus, we must have $u_1(h^1, x_1, G^c) = (1-\delta)\frac{B}{2}$. Lemma 6 implies that $u_1(h^1, x_1, G^d) \geq (1-\delta)\frac{B}{2}$, thus we must have $u_1(h^1, x_1, G^d) \geq u_1(h^1, x_1, G^c)$. Proposition 7 implies $u(G^c) > u(G^d)$, so we must have $u_2(h^1, x_1, G^d) < u_2(h^1, x_1, G^c)$, implying $u_2(G^c) > u_2(G^d)$. □

Proof of Proposition 9. Consider the following strategy profile in which Player 1 always holds the whip. In period 1, the state of the dynamics is given by $(u_1, u_2) = (\frac{1}{2}(1-\delta)B, f(\frac{1}{2}(1-\delta)B))$, the action is fixed as $(1,1)$, and the continuation payoff is given by

$(u_1,u_2)=(\frac{1-\delta}{\delta}(c-\frac{B}{2}),f(\frac{1-\delta}{\delta}(c-\frac{B}{2}))$. In period 2, the state of the dynamics randomizes between $(u_1,u_2)=(\frac{1}{2}(1-\delta)B,f(\frac{1}{2}(1-\delta)B))$ and $(u_1,u_2)=(B+\frac{b}{2}-c,B+\frac{b}{2}-2c)$. If the former realizes, the state of the dynamics gets back to what happens in period 1. If the latter realizes, the state of the dynamics is absorbed by $(u_1,u_2)=(B+\frac{b}{2}-c,B+\frac{b}{2}-2c)$. This strategy profile constitutes an SPE because the proof of Proposition 5 has shown that in period 1, action $(1,1)$ can be enforced with continuation payoffs $(u_1,u_2)=(\frac{1-\delta}{\delta}(c-\frac{B}{2}),f(\frac{1-\delta}{\delta}(c-\frac{B}{2})))$, and action $(1,2)$ with $(u_1,u_2)=(B+\frac{b}{2}-c,B+\frac{b}{2}-2c)$ is an SPE due to Assumption 2. Following the SPE we construct, the dynamics in the long run fall into $(u_1,u_2)=(B+\frac{b}{2}-c,B+\frac{b}{2}-2c)$ with probability one, where the action is $(1,2)$ forever. □

Proof of Corollary 1. For any $t\in(1,T)$, let θ denote the probability that the effort profile $(1,1)$ is played. We must have

$$\frac{1-\delta}{\delta}\left(c-\frac{B}{2}\right)=\theta\left(\frac{1-\delta}{2}B\right)+(1-\theta)(B+\frac{b}{2}-c).$$

Implicit differentiation gives

$$\frac{\partial\theta}{\partial\delta}=-\frac{-\frac{1}{\delta^2}\frac{2c-B}{2}+\frac{\theta B}{2}}{B+\frac{b}{2}-c-\frac{1-\delta}{2}B}=-\frac{-\frac{1}{\delta^2}\frac{\delta^{FB}B}{2}+\frac{\theta B}{2}}{B+\frac{b}{2}-c-\frac{1-\delta}{2}B}>0$$

because $\delta\le\delta^{FB}$ and $\theta<1$. Thus, higher δ increases the probability of $(1,1)$, delaying the expected realization of T. □

Proof of Proposition 10. If the organization is non-autonomous, it can implement the first-best effort profile if $(1-\rho)B\ge c$. In this case, the joint surplus is $2(1-\rho)B-2c$. To be an optimal organization, this surplus must be greater than that in point V in Figure 3, which is $\frac{1}{2}(1-\delta)B+B-c-\lambda(c-\frac{b}{2})$, where $\lambda=\frac{2c-(1+\delta)B}{b}$. These two inequalities jointly imply $2\rho B<\min\left\{2(B-c),\frac{B}{2}(1-\delta)-c-\lambda(c-\frac{b}{2})\right\}$. □

Proof of Corollary 3. In a stationary organization with $p=1$, let ω denote Player 1's output share. To sustain $(e_1,e_2)=(1,1)$, we need $\omega\ge\frac{c}{(1+\delta)B}$ to satisfy Player 1's IC constraint. Note that, for $\delta<\delta^{FB}$, $\frac{c}{(1+\delta)B}>\frac{1}{2}$. □

References

ABREU, D., D. PEARCE, AND E. STACCHETTI (1990): "Toward a theory of discounted repeated games with imperfect monitoring," *Econometrica*, 58, 1041–1063.

ACEMOGLU, D. AND A. WOLITZKY (2011): “The economics of labor coercion,” *Econometrica*, 79, 555–600.

AGHION, P. AND J. TIROLE (1997): “Formal and real authority in organizations,” *Journal of Political Economy*, 105, 1–29.

ALBUQUERQUE, R. AND H. A. HOPENHAYN (2004): “Optimal lending contracts and firm dynamics,” *The Review of Economic Studies*, 71, 285–315.

ALCHIAN, A. A. AND H. DEMSETZ (1972): “Production, information costs, and economic organization,” *American Economic Review*, 62, 777–795.

BAKER, G., R. GIBBONS, AND K. J. MURPHY (1994): “Subjective performance measures in optimal incentive contracts,” *Quarterly Journal of Economics*, 109, 1125–1156.

——— (1999): “Informal Authority in Organizations,” *The Journal of Law, Economics, and Organization*, 15, 56–73.

——— (2002): “Relational contracts and the theory of the firm,” *Quarterly Journal of Economics*, 117, 39–84.

——— (2023): “From incentives to control to adaptation: Exploring interactions between formal and relational governance,” *Journal of Institutional and Theoretical Economics*, 179, 500–529.

BARRON, D. AND Y. GUO (2021): “The use and misuse of coordinated punishments,” *Quarterly Journal of Economics*, 136, 471–504.

BARRON, D., J. LI, AND M. ZATOR (2022): “Morale and debt dynamics,” *Management Science*, 68, 4496–4516.

BIAIS, B., C. BISIÈRE, M. BOUVARD, AND C. CASAMATTA (2019): “The blockchain folk theorem,” *Review of Financial Studies*, 32, 1662–1715.

BOLTON, P. AND M. DEWATRIPONT (2004): *Contract theory*, MIT Press.

——— (2013): “Authority in organizations: A survey,” in *The Handbook of Organizational Economics*, ed. by R. Gibbons and J. Roberts, Princeton: Princenton University Press, chap. 9, 342–372.

BUDISH, E. (2023): “Trust at scale: The economic limits of cryptocurrencies and blockchains,” Working Paper, U. of Chicago.

CHASSANG, S. (2010): “Building routines: learning, cooperation and the dynamics of incomplete relational contracts,” *American Economic Review*, 100, 448–65.

CHE, Y.-K. AND S.-W. YOO (2001): “Optimal incentives for teams,” *American Economic Review*, 91, 525–541.

CHEUNG, S. N. S. (1983): "The contractual nature of the firm," *Journal of Law and Economics*, 26, 1–21.

CHWE, M. (1990): "Why were workers whipped? Pain in a principal-agent model," *Economic Journal*, 100, 1109–1121.

CLEMENTI, G. L. AND H. HOPENHAYN (2006): "A theory of financing constraints and firm dynamics," *Quarterly Journal of Economics*, 121, 229–265.

COASE, R. H. (1937): "The nature of the firm," *Economica*, 4, 386–405.

DEB, J., J. LI, AND A. MUKHERJEE (2016): "Relational contracts with private subjective evaluations," *RAND Journal of Economics*, 47, 3–28.

FAHN, M. AND G. ZANARONE (2022): "Transparency in relational contracts," *Strategic Management Journal*, 43, 1046–1071.

FERREIRA, D., J. LI, AND R. NIKOLOWA (2023): "Corporate capture of blockchain governance," *Review of Financial Studies*, 36, 1364–1407.

GOULDNER, A. W. (1954): *Patterns of industrial bureaucracy*, Free Press.

GROSSMAN, S. J. AND O. HART (1986): "The costs and benefits of ownership: A theory of vertical and lateral integration," *Journal of Political Economy*, 94, 691–719.

HALONEN, M. (2002): "Reputation and the allocation of ownership," *Economic Journal*, 112, 539–558.

HAN, J., J. LEE, AND T. LI (2023): "DAO Governance," Working Paper, Seoul National U. and U. of Florida.

HOLMSTRÖM, B. (1982): "Moral hazard in teams," *Bell Journal of Economics*, 13, 324–340.

——— (1999): "The firm as a subeconomy," *Journal of Law, Economics, Organization*, 15, 74–102.

——— (2016): "Grossman-Hart (1986) as a Theory of Markets," in *The Impact of Incomplete Contracts on Economics*, ed. by P. Aghion et al., Oxford University Press.

KVALOY, O. AND T. OLSEN (2006): "Team incentives and relational employment contracts," *Journal of Labor Economics*, 24, 139–169.

——— (2009): "Endogenous verifiability and endogenous contracts," *American Economic Review*, 99, 2193–2208.

LI, J. AND N. MATOUSCHEK (2013): "Managing conflicts in relational contracts," *American Economic Review*, 103, 2328–51.

LI, J., N. MATOUSCHEK, AND M. POWELL (2017): "Power dynamics in organizations," *American Economic Journal: Microeconomics*, 9, 217–241.

LI, J., A. MUKHERJEE, AND L. VASCONCELOS (2023): "What makes agility fragile? A dynamic theory of organizational rigidity," *Management Science*, 69, 3578–3601.

MUKHERJEE, A. AND L. VASCONCELOS (2011): "Optimal job design in the presence of implicit contracts," *RAND Journal of Economics*, 42, 44–69.

PADRO I MIQUEL, G. AND P. YARED (2012): "The political economy of indirect control," *Quarterly Journal of Economics*, 127, 947–1015.

PICCIONE, M. AND A. RUBINSTEIN (2007): "Equilibrium in the jungle," *Economic Journal*, 117, 883–896.

RAJAN, R. G. AND L. ZINGALES (1998): "Power in a theory of the firm," *Quarterly Journal of Economics*, 113, 387–432.

RANTAKARI, H. (2023): "Simon says? Equilibrium obedience and the limits of authority," *Journal of Law, Economics, and Organization*, forthcoming.

RAYO, L. (2007): "Relational incentives and moral hazard in teams," *Review of Economic Studies*, 74, 937–963.

SIMON, H. A. (1951): "A formal theory of the employment relationship," *Econometrica*, 19, 293–305.

THOMAS, J. AND T. WORRALL (2018): "Dynamic relational contracts under complete information," *Journal of Economic Theory*, 175, 624–651.

TROYA-MARTINEZ, M. AND L. WREN-LEWIS (2023): "Managing relational contracts," *Journal of the European Economic Association*, 21, 941–986.

VAN DEN STEEN, E. J. (2010): "Interpersonal authority in a theory of the firm," *American Economic Review*, 100, 466–490.

WILLIAMSON, O. E. (2002): "The theory of the firm as governance structure: From choice to contract," *Journal of Economic Perspectives*, 16, 171–195.

YOFFIE, D. B. AND M. KWAK (2001): *Judo Strategy: Turning your competitors' strength to your advantage*, Harvard Business School Press.

Governance and Management of Autonomous Organizations: Internet Appendix

Daniel Ferreira [*] Jin Li [†]

February 2, 2025

1 Management and Governance Issues in DAOs

While Decentralized Autonomous Organizations can take many different forms, the typical example is an organization that raises funds from its members to pursue some collective goals. A famous example is ConstitutionDAO, which raised over $40 million in an (ultimately unsuccessful) attempt to buy a copy of the U.S. Constitution in an auction. A DAO usually raises funds by selling tokens created on a "smart contract" platform such as Ethereum.[1] "Decentralization" means that all members have the right to participate directly in decision-making, such as how to spend treasury funds and how to govern the organization. Typically, decision-making rights are distributed as governance tokens. Most decisions are voted on by members who own the governance tokens.

DAOs face a traditional collective action problem: To achieve a common goal, the individual members must exert costly effort. For example, a DAO must often decide how to allocate its funds across multiple projects. Individual members must gather information to decide which projects to support. Because information acquisition is costly, members have an incentive to free-ride on the effort of others.[2] Because most DAOs are not legal entities, DAO members usually cannot resort to the legal system to enforce contracts among mem-

[*] London School of Economics, CEPR and ECGI, d.ferreira@lse.ac.uk.

[†] HKU Business School, jli1@hku.hk.

[1] A smart contract is a piece of code that automatically executes a transaction once prompted by a message. A famous analogy is that of a vending machine, in which a product is dispensed automatically once coins are inserted. Smart contracts can be "state-dependent," in the sense that a transaction is executed automatically if a particular state occurs.

[2] See, e.g., Hall and Oak (2023): "*Users of online systems expect convenience and are generally uninterested in participating in governing the platforms that they use. Rates of voting in online communities in the web3 space are generally quite low*" (p.1).

bers.[3] Thus, contract enforcement is mainly based on code (i.e., self-executing contracts) and relational incentives (i.e., trust and reputation).

While the ability to write code that automates contract execution is touted as the greatest strength of DAOs, in reality, only some transactions can be automated. Most DAOs depend on "off-chain" actions, which require human execution. In the example of ConstitutionDAO, someone must convert digital coins into fiat money, save them in a bank account, and physically bid in the auction. After a failed bid, there is also the non-trivial issue of returning (some of) the money to members and paying for operation costs. As a matter of fact, ConstitutionDAO never held a single vote using its token.[4] Because of these off-chain actions, most DAOs have a core team (or a foundation), who often have discretion over many decisions. These are essentially (in all but name) "managers."[5]

The existence of managers implies that real-world DAOs are not as decentralized as theoretical DAOs.[6] Examples of abuse of power by DAO managers abound. The foundation that manages the blockchain Arbitrum allegedly started to spend its funds even though its nearly $ 1 billion budget had not yet been approved by governance token holders.[7] The core team that runs Aragon—a DAO that builds tools for managing DAOs—banned some DAO members from its governance discussion forums. Commenting on the ban, CoinDesk contributor Danny Nelson concludes that "*their banishment from Aragon's Discord for asking 'probing questions' and using 'inappropriate language' highlights the disconnect between the censorship-resistant ideals of crypto governance and the reality that insiders hold considerable sway*."[8] In November 2023, without holding a vote, the Aragon team decided to dissolve the DAO's governing body and return most of its assets to tokenholders. DAO members voted to sue the Aragon Team, which shows that full autonomy is often a myth.[9]

DAO managers' power is not absolute. DAO members who are unhappy with management may leave the organization. A prominent example is Nouns, a DAO that invests in several projects that promote their branded NFTs. Unhappy with management decisions,

[3] See https://t.ly/Mf806

[4] https://www.vice.com/en/article/bvnze5/constitutiondao-is-shutting-down-after-unrelenting-chaos

[5] DAO founders and key players understandably avoid using titles such as CEOs, executives, and managers. Instead, they often refer to themselves as "core developers," "heads" and "leads."

[6] Ethereum—the "Layer-1" blockchain on which most DAOs are built—is also fairly centralized. For example, Fracassi, Khoja, and Schär (2024) show that ten individual developers contributed 68% of all implemented core Ethereum Improvement Proposals.

[7] https://t.ly/y2T6n.

[8] https://t.ly/A5Ru9. Note that, although notionally decentralized, Aragon has a "Head of Communications."

[9] https://cointelegraph.com/news/aragon-dao-lawsuit-founders-patagon-management

56% of Nouns NFT holders voted to leave the organization, taking about $27 million worth of treasury funds along with them. The defectors created a new DAO, with the same NFT artwork as the original, where each holder is allowed to "ragequit" and take some of the funds with them.[10]

In contrast to these examples, an example of a high-performing autonomous organization is the Ethereum network. Unlike most of the DAOs built on its blockchain, the governance of Ethereum is entirely off-chain. Ethereum has no governance tokens and holds no "on-chain" votes to decide on changes and upgrades. Instead, changes are adopted when some key players in the network agree, in a process often dubbed "rough consensus." In practice, this means that the Ethereum Foundation, particularly its effective controller, Vitalik Buterin, has most of the real decision-making authority. Thus, Ethereum's governance is highly centralized.

As these examples illustrate, real-world DAOs (as opposed to idealized DAOs) are rife with governance, management, and performance problems. Because of their alleged autonomy, external enforcement of contracts is limited. DAOs are typically centralized due to the power of core teams and foundations. These managers may be able to punish bad behavior, for example, by banning some members or canceling their tokens. However, they can also abuse their power and have discretion over the use of funds. Non-managing members have the option to quit, thus imposing costs on those who stay.

2 Omitted Proofs

Proof of Lemma 5. Let $u_i(h^t, x_t)$ denote Player i's normalized continuation payoff at (h^t, x_t) under equilibrium play. Suppose $u_1(h^t, x_t) \geq u_2(h^t, x_t)$ for some (h^t, x_t). If $g(h^t, x_t) = 2$ and $e_2(h^t, x_t) = 1$, Player 2's incentive constraint implies $(1-\delta)\left(c - \frac{B}{2}\right) \leq \delta u_2(h^t, x_t)$. Because $u_2(h^t, x_t) \leq B - c$, the constraint implies that $\delta \geq (2c - B)/B = \delta^{FB}$, which contradicts Assumption 3. Thus, if $g(h^t, x_t) = 2$, we must have $e_2(h^t, x_t) = 0$ (notice that it if $e_2(h^t, x_t) = 1$ is not incentive compatible, $e_2(h^t, x_t) = 2$ also is not) . We conclude that if Player i is the manager and is expected to work at (h^t, x_t), this player must have the higher continuation payoff at (h^t, x_t). □

[10] https://decrypt.co/197400/nouns-fork-disgruntled-nft-holders-exit-27-million-from-treasury

Proof of Lemma 6. Player i's IC constraint for $e_i(h^t,x_t)=1$ is

$$(1-\delta)\left(\frac{B}{2}+\frac{y_{-i}(h^t,x_t)}{2}-c\right)+\delta u_i(h^t,x_t)\geq(1-\delta)\frac{y_{-i}(h^t,x_t)}{2}, \tag{IA.1}$$

where $y_{-i}(h^t,x_t)$ is the output of Player $-i$ (i.e., not i). Rearranging implies $u_i(h^t,x_t)\geq\underline{u}$. Thus, if $u_i(h^t,x_t)\geq\underline{u}$ holds, $e_i(h^t,x_t)=1$ can be enforced when $g(h^t,x_t)=i$. Because the IC constraint for $e_i(h^t,x_t)=2$ implies (IA.1), if either $e_i(h^t,x_t)=1$ or $e_i(h^t,x_t)=2$ is enforceable, (IA.1) must hold. □

Proof of Lemma 7. Suppose $u_1(h^t,x_t)\geq u_2(h^t,x_t)$ for some (h^t,x_t). Then, by Lemma 5, Player 2 does not exert effort if $g(h^t,x_t)=2$. Suppose Player 2 exerts positive effort at some (h^t,x_t), where $g(h^t,x_t)=1$, but Player 1 shirks. Player 2's participation constraint requires $(1-\delta)(\frac{y_2(h^t,x_t)}{2}-ce_2(h^t,x_t))+\delta u_2(h^t,x_t)\geq 0$, which doesn't hold because $u_2(h^t,x_t)<B-c$ (Player 2 has the lowest payoff, and the maximum joint payoff is less than $2(B-c)$). Thus, we conclude that when Player 2 exerts nonzero effort, Player 1 must also exert nonzero effort ⇒ if Player 1 shirks, Player 2 must also shirk (Part 1).

For the organization to improve upon r_0, at least one player must exert nonzero effort for some (nonzero measure) set (h^t,x_t). Suppose no one ever overworks. Then, Lemma 6 implies that managers would always choose zero effort. Part 1 then implies that subordinates always choose zero effort. Thus, to improve upon r_0, some players must overwork with positive probability in equilibrium (Part 2). □

Proof of Proposition 11. Consider an organization where at (h^t,x_t) Player i has whip. Let ω_i denote Player i's output share and $u_i(h^t,x_t)$ denote Player i's normalized continuation payoff. To sustain $(e_1,e_2)=(1,1)$ at this period, we need to find ω_i that satisfies both Player i's incentive constraint and Player $-i$'s participation constraint. Player i's incentive constraint implies

$$\omega_i\geq\frac{c}{B}-\frac{1}{B}\frac{\delta}{1-\delta}u_i(h^t,x_t), \tag{IA.2}$$

and Player $-i$'s participation constraint implies

$$\omega_i\leq 1-\frac{c}{2B}+\frac{1}{2B}\frac{\delta}{1-\delta}u_{-i}(h^t,x_t). \tag{IA.3}$$

These constraints imply

$$2B-3c+\frac{\delta}{1-\delta}(2u_i(h^t,x_t)+u_{-i}(h^t,x_t))\geq 0. \tag{IA.4}$$

Note that, if $2B-3c\geq 0$, condition (IA.4) holds independently of $u_i(h^t,x_t)$ and $u_{-i}(h^t,x_t)$.

For example, if we set $\omega = \frac{c}{B}$, both (IA.2) and (IA.3) are satisfied for any δ, $u_i(h^t, x_t)$ and $u_{-i}(h^t, x_t)$. Thus, in such a case, $(1,1)$ can be implemented for any δ. Suppose $2B - 3c < 0$. For a given δ, Condition (IA.4) is maximally slack by setting $u_i(h^t, x_t) = 2(B-c)$, that is, by implementing effort profile $(1,1)$ in all future dates and giving all future surplus to Player i. This requires setting $\omega_i = 1 - \frac{c}{2B}$, which is greater than 0.5. The minimum discount factor for which the first best can be implemented in this case is $\delta^c := \frac{3c-2B}{2B-c}$. Thus, a necessary condition for an optimal organization to implement the first-best is $\delta \geq \delta^c$.

If $\delta \geq \delta^c$, the first-best can always be implemented by a fully centralized organization. To show this, note that for an organization that implements the first-best, we can rewrite (IA.4) as $2B - 3c + \frac{\delta}{1-\delta}(2(B-c) + u_i(h^t, x_t)) \geq 0$. Increasing $u_i(h^t, x_t)$ always slacks this condition. Let $u(\delta)$ denote the minimum $u_i(h^t, x_t)$ such that (IA.4) holds with equality. We then have

$$u(\delta) = \frac{3c - 2B}{\delta} - c.$$

Because $u(\delta) \leq 2(B-c)$ (the maximum feasible payoff), we obtain $\delta \geq \delta^c$. Thus, the first-best can be implemented by an organization where Player i holds the whip at every period and receives normalized payoff $u(\delta)$ if and only if $\delta \geq \delta^c$.

We now find a necessary condition for a fully decentralized organization to be optimal. Suppose that, at (h^t, x_t), Player 1 has the whip. Thus, the minimum continuation value for Player 1 so that both players work at t is $u(\delta)$. Suppose the equilibrium is fully decentralized. Thus, Player 1 will retain the whip in period $t+1$ with probability 0.5. If $g_{t+1} = 1$, Player 1's maximal payoff is $2(B-c)$ (i.e., the full joint payoff). If $g_{t+1} = 2$, Player 1's normalized payoff at that time is $(1-\delta)(1-\omega_2)2B - (1-\delta)c + \delta u_1$, where u_1 is her continuation value. For given u_1, the maximum payoff Player 1 can expect is when (IA.2) binds, which implies her payoff will be

$$(1-\delta)(1 - \frac{c}{B} + \frac{1}{B}\frac{\delta}{1-\delta}(2(B-c) - u_1))2B - (1-\delta)c + \delta u_1.$$

Player 1's payoff when subordinate at $t+1$ is maximal when $u_1 = 0$, which then gives

$$(1-\delta)(2B - 2c + \frac{\delta}{1-\delta}4(B-c) - (1-\delta)c = 2B - 3c + \delta(2B - c),$$

which is positive if and only if $\delta \geq \delta^c$. Thus, at best, Player 1's continuation value at time t is

$$\frac{1}{2}2(B-c) + \frac{1}{2}(2B - 3c + \delta(2B-c)),$$

Thus, for providing $u(\delta)$ to Player 1 at t we need

$$\frac{1}{2}2(B-c)+\frac{1}{2}(2B-3c+\delta(2B-c))\geq u(\delta)$$

which requires

$$4B-3c\geq\frac{2(3c-2B)}{\delta}-\delta(2B-c).$$

Solving it for the root, we find that

$$\tilde{\delta}=\frac{3c-4B+\sqrt{40Bc-16B^2-15c^2}}{2(2B-c)}.$$

It is easily verified that $\tilde{\delta}\in(\delta^c,\frac{3c-2B}{B})$. Thus, a fully decentralized organization cannot implement the first-best if $\delta<\tilde{\delta}$. □

References

FRACASSI, C., M. KHOJA, AND F. SCHÄR (2024): "Decentralized crypto governance? Transparency and concentration in Ethereum decision-making," Working Paper, U. of Texas at Austin and U. of Basel.

HALL, A. B. AND E. R. OAK (2023): "What kinds of incentives encourage participation in democracy? Evidence from a massive online governance experiment," Working Paper, Stanford U. and Yale U.