Governance and Management of Autonomous Organizations*

Daniel Ferreira † Jin Li ‡

March 11, 2024

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 monitoring or contract enforcement. We present a model of the optimal power allocation in such an organization. The organization commits to a *governance structure* that allocates managerial power to agents. Members with power ("managers") can punish members without power ("subordinates"). This power is, however, limited by the subordinates' right to exit the organization. There are three main results. First, the goals of autonomy, decentralization, and efficiency conflict with one another. We call this result the *Organizational Trilemma*. Second, there is a *Paradox of Power*: an agent can be made worse off by their own power. Third, optimal governance structures in autonomous organizations are centralized and *populist*: the powerful party shows restraint in early periods, only to abuse their power in later periods.

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

*We thank Ulf Axelson, Dan Barron, Kim-Sau Chung, Matthias Fahn, Bob Gibbons, Peter Kondor, Radoslawa Nikolowa, Jose Parra-Moyano, Scott Schaefer, and seminar participants at the Chinese University of Hong Kong, Hong Kong Baptist University, LME Workshop, London School of Economics, OEW 21, Tsinghua BEAT Workshop, University of Utah, 8th Workshop on Relational Contracts, and Zhejiang University digital conference for helpful conversations. We thank Hanzhe Li 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 an efficient solution only if the third party is honest, competent, and inexpensive. In the riverboat example, the whip holder may choose to extract bribes from the workers in exchange for lighter whipping. The monitor could also underpunish 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 third-party 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 to 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. Specifically, 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 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 two polar cases.

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 compromised. If efficiency is the goal, we cannot have both autonomy and decentralization.

The Organizational Trilemma implies that, in autonomous organizations, the optimal (i.e., surplus-maximizing) governance structure is centralized. That is, in an optimal structure, the power to punish others should reside with a few selected members. 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 are incentivized to exert effort in order to keep the organization intact. 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. In some situations, if the strong party becomes more powerful, the strong party exerts more effort in equilibrium without changing the behavior of the weak party. Thus, the Paradox of Power implies that an agent can be made worse off by their own power.

To illustrate the Paradox of Power and the Organizational Trilemma, we first solve an organization design problem under the assumption of stationarity. We then consider the general organization design problem without imposing stationarity. As in the stationary case, we show that the optimal organization is fully centralized: the same agent holds the whip in all periods. However, the equilibrium actions are nonstationary. In the early periods of the game, both the manager and the subordinate exert the same amount of effort. In later periods, the manager abuses his/her power and forces the subordinate to overwork. The powerful agent behaves like a "populist dictator:" benevolent at first but abusive later.

Our model is useful for understanding the strengths and limitations of 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, in real-world DAOs, off-chain transactions are governed 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).[1] 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.[2] 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 illustrate in the next section, 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 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.

This paper is related to several strands of literature. Our basic model setup is one

[1] 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.

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

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). A unique feature of our paper is that we introduce interpersonal power to the relationship, and the central question of our analysis is how to allocate power. In addition, by interpreting our history-contingent power allocation as a self-executing contract, we show how "smart contracts" can support relational contracts in the absence of external enforcement.

While the theoretical literature on blockchain economics is large, it mostly 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 orga-

nizations; 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 of the works in this literature study static power allocations, even when the environment is dynamic. Our model, in contrast, studies dynamic power allocations, highlighting the importance of the persistence of power.

2 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.[3] "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.[4] Because most DAOs are not legal entities, DAO members usually cannot resort to the legal system to enforce contracts

[3] 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.

[4] 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).

among members.[5] 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.[6] 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."[7]

The existence of managers implies that real-world DAOs are not as decentralized as theoretical DAOs.[8] 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.[9] 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."[10] In November 2023, without holding a vote, the Aragon team decided to dissolve the DAO's governing body and return most of its assets

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

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

[7] 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."

[8] 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.

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

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

to tokenholders. DAO members voted to sue the Aragon Team, which shows that full autonomy is often a myth.[11]

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, 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.[12]

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. But 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.

3 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.

[11] https://cointelegraph.com/news/aragon-dao-lawsuit-founders-patagon-management
[12] https://decrypt.co/197400/nouns-fork-disgruntled-nft-holders-exit-27-million-from-treasury

3.1 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. At each t, player i chooses effort $e_{it} \in \{0,1,2\}$, where the cost of effort is $c(e_{it}) = ce_{it}$, $c > 0$. The output of each player i 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* ($e_{it} = 2$). 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. This assumption is unnecessary for our analysis and is made only to reduce the number of cases to consider.

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 4.2, 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.[13] At the beginning of period t, one of the players may be 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, and $i \in \{1,2\}$. If the subordinate stays but chooses an effort level strictly lower than $\hat{e}_{-it}$, the manager can reduce the subordinate's payoff by $D > B$. Formally, the identity of the manager 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*.[14]

Managerial power—here represented by the whip—is a scarce resource. Accordingly, we assume that only one whip is available. 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 very small group of core developers have the keys to modify the blockchain protocol.[15]

Whip assignment affects the set of feasible actions for each player: The manager can suggest actions for both players $(\hat{e}_{1t}, \hat{e}_{2t})$, while the subordinate does not suggest actions. If no player is the manager (i.e., $g_t = 0$), no one suggests any action. To keep the space of

[13] 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).

[14] 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.

[15] 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.

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

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

where $\mathbb{1}_x$ is the indicator function. 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.[16]

We define the *history* at time t as $h^t = \{x_1, g_1, e_1^a, d_1, e_1, ..., x_{t-1}, g_{t-1}, e_{t-1}^a, d_{t-1}, e_{t-1}\}$. 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.[17]

[16]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.

[17]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

We interpret G as a contingent *self-executing contract*. In the DAO interpretation, G completely specifies under what conditions 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. Off-chain information must first be recorded on-chain, either by a member with special rights or by a third party (called an *oracle*). In a truly autonomous organization, the designated manager would observe off-chain behavior, record it in the underlying blockchain, and then punish those who misbehave.

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 a given game induced by G, at each t, we denote player i's (pure) actions by $S_{it} = (e^a_{it}, d_{it}(e^a_t), e_{it}(e^a_t, d_t))$, where subscript t indicates that the actions are conditional on (h^t, x_t).[18] Player i's strategy is thus 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 $S_t : (h^t, x_t) \to (e^a_t, d_t, e_t)$ 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

(2023).

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

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 exception is the non-autonomous organization described in Subsection 4.2), to economize notation, we will often denote an organization simply by $\langle G, S \rangle \in \Gamma \times \Psi(G)$.

3.2 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 $\langle G_0, S_0 \rangle$. Under G_0, cooperation can only be sustained 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 payoff of each player is $B - c$. We restrict attention to equilibria in 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 $d_{1t'} = d_{2t'} = 0$. Under such trigger strategies, the first-best payoffs can be sustained as an SPE if and only if

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

The left-hand side of (4) is player i's present value of working ($e_{it} = 1$) forever, and the right-hand side is the 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. $\delta < \delta^{FB} := \frac{2c-B}{B}$.

4 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 a 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 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.

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 of generality, consider an organization such that, for some (h^t, x_t), we have $g_t(h^t, x_t) = 0$. Let e_t^* denote the associated equilibrium efforts. Suppose instead that we set $g_t(h^t, x_t) = 1$. It is immediate that by setting the manager's announcement to $e_{1t}^a = e_t^*$ when (h^t, x_t) happens, the effort vector e_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, $e_{it}^a(h^t, x_t) = e_t(h^t, x_t)$, for all (h^t, x_t). In words, the manager always recommends the equi-

librium effort levels. It is easy to see that any equilibrium in which $e^a_{it}(h^t, x_t) \neq e_t(h^t, x_t)$ is payoff-equivalent to an equilibrium that differs from the original one only by setting $e^a_{it}(h^t, x_t) = e_t(h^t, x_t)$. This simplification implies that we can ignore e^a_t when characterizing an equilibrium.

Third, because we will consider only trigger strategies, the optimal participation decision is $d_{it} = 1$ always 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 described by a sequence of effort functions $e_{it} = e_{it}(h^t, x_t)$ and participation decisions $d_{it} = d_{it}(h^t)$.

4.1 Stationary Organizations

In this subsection, we consider the case of stationary autonomous organizations. We impose stationarity only to facilitate the analysis and the exposition. As we will see in Subsection 4.3, the main messages from the results remain unchanged when we study the general case of optimal autonomous organizations.

We first consider stationary organizations under the default governance structure, G_0. We say that 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*

$\langle G_0, S_0 \rangle$ where $e_{1t} = e_{2t} = 0$ always.[19]

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.:

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

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 as 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 the

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

(weakly) lower (normalized discounted) payoff of the two must shirk when he/she is the manager. To see why this is the case, assume that Player 2 has a lower payoff than that of 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 μ_i and μ_j 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.

To streamline the exposition, we initially consider only the case where $\mu_i(g_t)$ is degenerate, i.e., Player i's effort choice is deterministic for a given g_t. In this case, we can write $e_i^{g_t} := e_i(g^{-1}(g_t; p))$ and $e^{g_t} := (e_1^{g_t}, e_2^{g_t})$. This restriction is without loss of generality for the next three results. We will remove this restriction later when it becomes binding.

From now on, we assume that Player 1 has the (weakly) highest payoff in equilibrium. The next result shows that abuse of power must occur in an optimal stationary organization.

Lemma 2 (Equilibrium Abuse of Power). *In an optimal stationary organization, the effort profile $e^1 = (1, 2)$ must be played when Player 1 is the manager.*

This result implies that maximization of the joint surplus requires Player 1 to abuse her power as manager by asking the subordinate to overwork. Notice that we cannot have both $e^1 = (1, 2)$ and $e^2 = (2, 1)$ because no identity-blind equilibrium with positive effort

exists. Thus, an optimal organization must be asymmetric. In addition, Lemma 1 implies that there is no equilibrium in which $e_2^2 > 0$. Thus, an optimal stationary equilibrium must have either $e^2 = (0,0)$ or $e^2 = (1,0)$.[20]

The next proposition shows the existence of asymmetric equilibria involving $e^1 = (1,2)$, provided the governance structure is not too decentralized.

Proposition 3 (The Paradox of Power). *Define* $\delta_1 := \frac{2c-B}{B+b}$. *There exist* $p_2(\delta) > p_1(\delta) > 0.5$ *such that, if* $\delta \in [\delta_1, \delta^{FB})$,

1. $e^1 = (1,2)$ *and* $e^2 = (0,0)$ *can be enforced by a stationary SPE if and only if* $p \geq p_1(\delta)$,
2. $e^1 = (1,2)$ *and* $e^2 = (1,0)$ *can be enforced by a stationary SPE if and only if* $p \geq p_2(\delta)$.

Both $p_1(\delta)$ *and* $p_2(\delta)$ *decrease in* δ.

Proposition 3 shows that one can design an organization that improves upon the default organization. For this to happen, the organization must be sufficiently centralized, i.e., p must be greater than some threshold $p_1(\delta) > 0.5$. In the equilibrium in Case 1, when $g_t = 1$, Player 1 works and Player 2 overworks; when $g_t = 2$, both players shirk. When Player 1 is the manager, she abuses her power and asks Player 2 to overwork. Thus, in equilibrium, 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 delivers payoff asymmetry, which is necessary for providing sufficient continuation value to Player 1.

Proposition 3 also shows that if the organization is sufficiently centralized ($p \geq p_2(\delta)$), Player 1 may work in every period. In such a case, Player 1 prefers the lower-surplus equilibrium in Case 1 ($e^1 = (1,2)$ and $e^2 = (0,0)$) to that in Case 2 ($e^1 = (1,2)$ and

[20] It is easy to see that $e^2 = (2,0)$ is dominated by $e^2 = (0,0)$ or $e^2 = (1,0)$: (i) it has a lower joint payoff, (ii) it makes Player 1's incentive constraint harder to meet, and (iii) Assumption 2 implies that Player 2's participation constraint is slack under $e^1 = (1,2)$ and either $e^2 = (0,0)$ or $e^2 = (1,0)$.

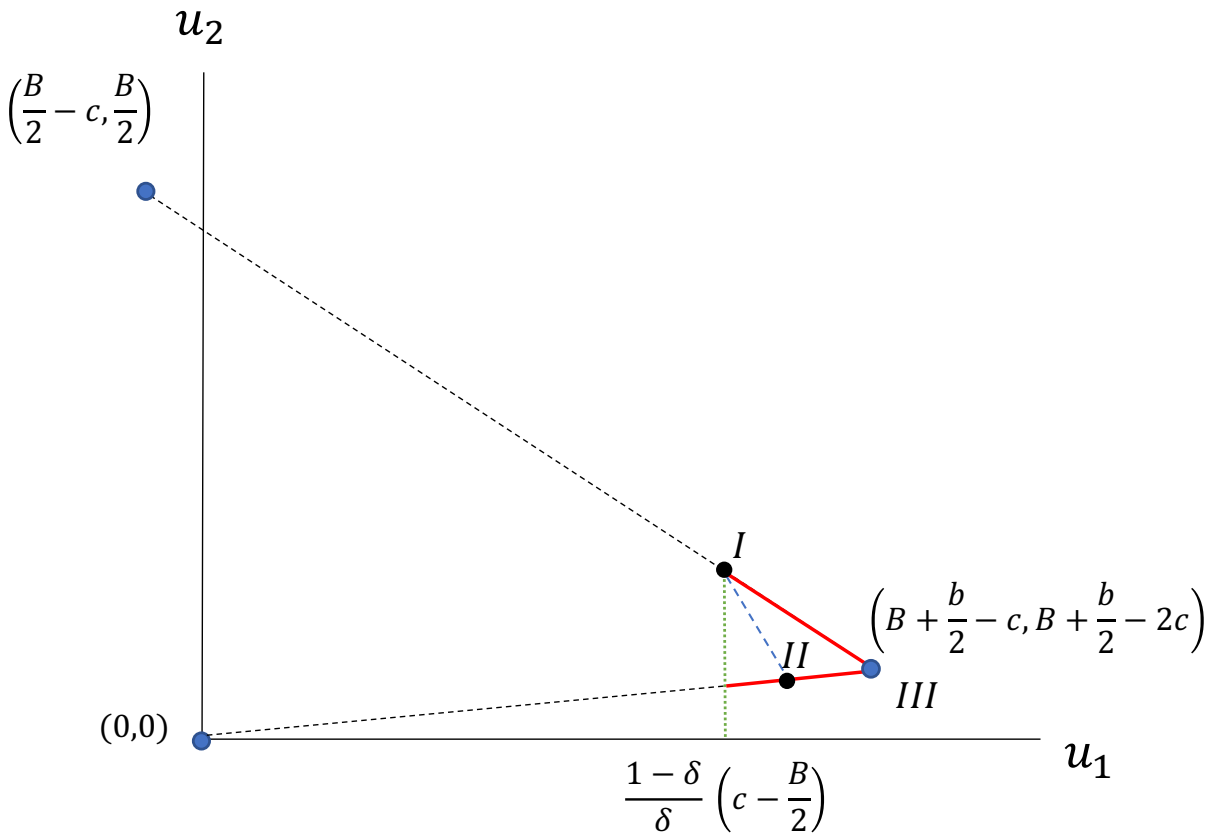

Figure 1: The Paradox of Power

$e^2 = (1,0)$). However, the organization members may coordinate instead on the higher-surplus equilibrium. Thus, Player 1 can be made worse off by her own power. We call this phenomenon *The Paradox of Power*. Intuitively, if power is sufficiently centralized in Player 1's hands, her continuation value is high. Player 2 can thus use the threat of exiting to induce Player 1 to work in all periods.

Figure 1 illustrates this argument. The figure shows the three relevant payoff profiles from Proposition 3 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 $e^1 = (1,2)$ and $e^2 = (1,0)$ with probabilities p and $1-p$. If the expected payoff profile is at point I, Player 1's expected payoff is $u_1 = p(B+\frac{b}{2}-c) + (1-p)(\frac{B}{2}-c) = \frac{1-\delta}{\delta}(c-\frac{B}{2})$, which is the expression that defines $p_2(\delta) := \frac{2C-B}{\delta(B+b)}$. For $p =$

$p_2(\delta)$, if the equilibrium involves $e^1 = (1,2)$ and $e^2 = (0,0)$, the expected payoff profile is at point II. Player 1 is better off at II. However, the sum of payoffs ($E(u_1 + u_2)$) is higher at I. If Player 1 were less powerful (i.e., if $p < p_2(\delta)$), point I would not be sustainable. Thus, Player 1 can be made worse off by having more power. Intuitively, because more power increases Player 1's continuation value, it eventually becomes possible to coordinate on an equilibrium where Player 1 always works ($e_{1t} = 1$). In that equilibrium, Player 2 gains power over Player 1 by threatening to exit in case Player 1 does not work.

In Figure 1, the sum of the payoffs increases as we move along the $I - III$ line. Higher p allows for equilibrium payoffs closer to III, and $p = 1$ can sustain the action profile $(1,2)$ with probability 1 (point III), which dominates (i.e., has a larger sum of payoffs than) all points on the $I - III$ line. Thus, Proposition 3 implies the following result.

Corollary 1 (Centralization is Optimal). *If $\delta \in [\delta_1, \delta^{FB})$, an optimal stationary organization must be fully centralized ($p = 1$).*

This corollary shows that if $\delta \in [\delta_1, \delta^{FB})$, an organization can implement $(1,2)$ every period, but only under full centralization ($p = 1$). In fact, a fully-centralized stationary organization can do even better. To improve upon profile $(1,2)$, Player 2 must choose an effort level lower than 2 with some probability. Thus, we now remove the restriction that $\mu_i(g_t)$ is degenerate. The following proposition characterizes the optimal stationary organization.

Proposition 4 (Optimal Stationary Organizations). *Consider a stationary autonomous organization with $p = 1$. There exists $\alpha_1(\delta)$ such that for all $\delta \in [\delta_1, \delta^{FB})$, if $\alpha \in [\alpha_1(\delta), 1]$, the action profile that randomizes between $(1,2)$ and $(1,1)$ with probabilities α and $1 - \alpha$ can be enforced by some SPE. Furthermore, if $\delta \in [\delta_1, \delta^{FB})$, an organization with $p = 1$ and $\alpha = \alpha_1(\delta)$ is an optimal stationary organization.*

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

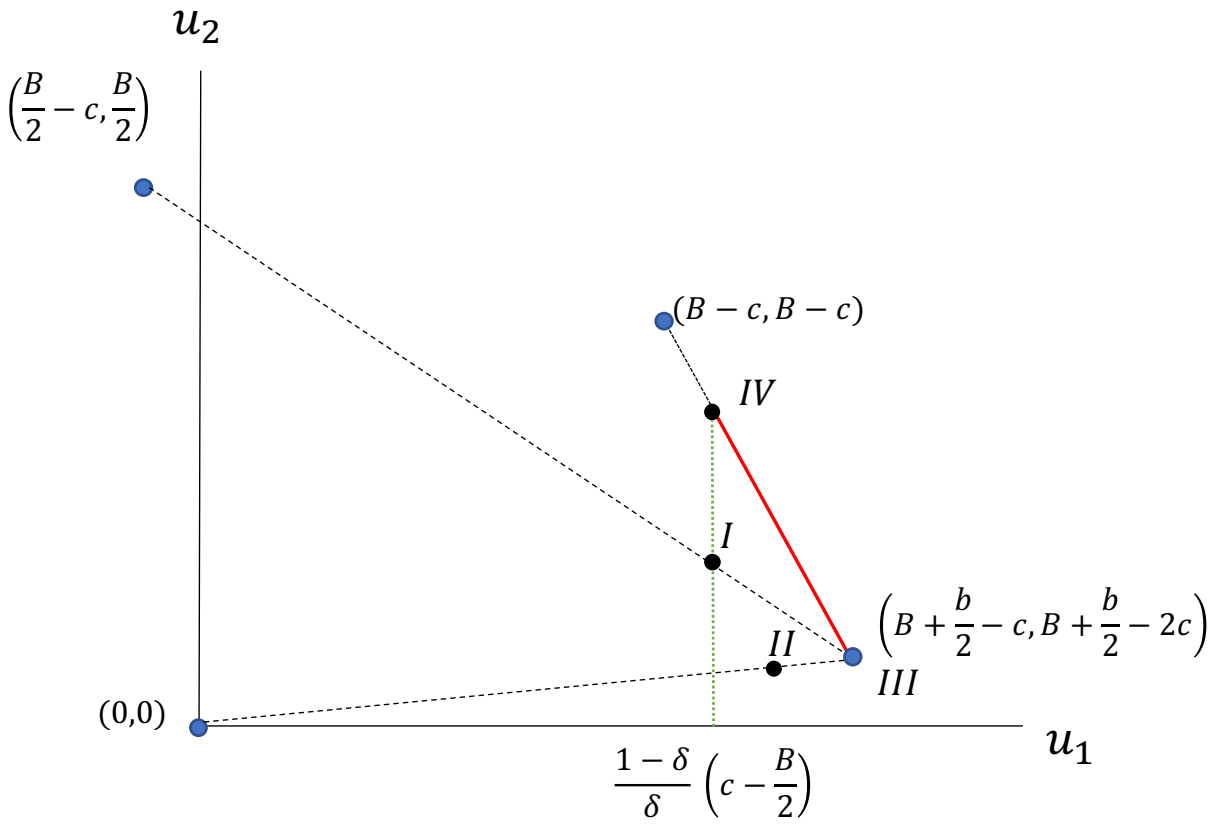

Figure 2: Implementable payoffs under full centralization

$u_1 = \alpha(B+\frac{b}{2}-c)+(1-\alpha)(B-c) = \frac{1-\delta}{\delta}(c-\frac{B}{2})$, which is the expression that defines $\alpha_1(\delta) := \frac{2C-(1+\delta)B}{\delta b}$. Any payoff profile on the $III-IV$ line can be sustained for some $\alpha \geq \alpha_1(\delta)$. 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 $\alpha_1(\delta)$ and $1-\alpha_1(\delta)$ is an optimal stationary organization.

Proposition 4 shows that under full centralization, for sufficiently high $\delta < \delta^{FB}$, there exists a type of equilibrium in which neither player shirks. In this equilibrium, the manager works in every period. The subordinate alternates between working and overworking. By overworking, Player 2 increases the value of the relationship for Player 1 and

induces Player 1 to work. Proposition 4 shows that overworking must occur. The probability of overworking depends on the discount factor. It can be shown that as the discount factor drops, the minimum probability of overworking increases.

The type of equilibrium in Proposition 4 has several features and lessons. First, the equilibrium shows that a designated enforcer is helpful to induce effort. Different from existing models of relational contracting, a key feature of our model is the introduction of a whip. This whip enables better enforcement of desired behaviors because the deviator can now also be punished by the whip. As discussed above, there can be many different ways to allocate the whip. The results in this section indicate that a designated enforcer is essential in stationary autonomous organizations. Without a designated enforcer, players must have a sufficiently high payoff to be induced to work when they have the whip. But since the sum of the players' payoffs is bounded by the first-best payoff level, no division of the payoffs is feasible to induce both players to work. By having one player always as the designated enforcer (manager), the governance structure eliminates the need to consider the incentive constraint of the other player as manager.

Another feature of the model is that, under full centralization, the manager has more than half of the surplus. This payoff asymmetry arises because without giving sufficient payoff to the manager, she will not put in effort. The possession of power, therefore, necessitates a high level of payoff. The positive association between "power" and "payoff" is reminiscent of Alchian and Demsetz's (1972) solution of "who monitors the monitor" in the sense that, there, the residual claimant (owner) carries out the role of the manager by measuring the effort of the subordinates. But the difference is that, here, the manager's role is not to measure output but instead to exert productive effort and enforce the contract. In addition, the manager is not the residual claimant: she gets half of the surplus.

Finally, note that even if the players are ex-ante identical, our model leads to a hierarchical division of labor. At the top, there is the manager. She is motivated by a "carrot:" the prospects of long-term rewards from a well-functioning organization. At the bottom, there is the worker. He is driven by a "stick:" the immediate penalties for failing to carry out the order given by the manager. The hierarchy places the worker under the manager's

control, effectively limiting his autonomy within the relationship. Through this perspective, our model suggests that one (personal) benefit of power is the freedom it grants.

4.2 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 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, we 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. Consider the governance structure $g(h^t, x_t) = 3$ 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 that it assumes 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), incompe-

tence (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 only sustain the first-best effort profile if $(1-\varrho)B \geq c$. Even in that case, the first-best payoffs can no longer be attained.

The following proposition summarizes the trilemma of decentralization, autonomy, and efficiency.

Proposition 5 (**The Organizational Trilemma**). *For stationary organizations in which* $\delta \in \left[\delta_1, \delta^{FB}\right)$*, we have the following tradeoffs.*

1. ***(Decentralization + Autonomy*** $\rightarrow$ ***Inefficiency).*** *A fully decentralized autonomous organization is inefficient and implements action profile* $(0,0)$ *in all periods.*

2. ***(Autonomy + Efficiency*** $\rightarrow$ ***Centralization).*** *An optimal autonomous organization is fully centralized.*

3. ***(Decentralization + Efficiency*** $\rightarrow$ ***Non-autonomy).*** *A decentralized organization is optimal if and only if it is non-autonomous and* $2\varrho B < \min\{2(B-c), \alpha_1(\delta)(c-b)\}$.

The Organizational Trilemma implies that decentralized autonomous organizations must be inefficient unless players are sufficiently patient (i.e., Assumption 3 does not hold). To restore efficiency, 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.

4.3 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.

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

$$\max_{\langle G,S\rangle \in \Gamma \times \Psi(G)} (1-\delta) E\left[\sum_{t=1}^{\infty} \delta^{t-1} u_{2t} \mid G, S\right] \tag{7}$$

$$\text{s.t. } (1-\delta) E\left[\sum_{t=1}^{\infty} \delta^{t-1} u_{1t} \mid G, S\right] = u_1. \tag{8}$$

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.[21]

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

[21]The self-generating property arises because the players can publicly observe the actions of Player 2. When it is publicly known that Player 2 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.

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.

We restrict the analysis to discount factors in $[\delta_1, \delta^{FB})$ to facilitate the comparison with the stationary case. We have the following result:

Proposition 6 (Equilibrium Payoff Frontier). *For $\delta \in [\delta_1, \delta^{FB})$, the following holds.*

1. *$f(u_1)$ is symmetric along the 45-degree line.*
2. *For $u_1 \geq \frac{1}{2}(1-\delta)B$, $(u_1, f(u_1))$ is on the line segments 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 6 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 $III - V$, which lies on the line segment whose payoff pair on the one end requires both players to work $(B-c, B-c)$ and on the other end requires 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 organizational design (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 6 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 joint payoff of the players.

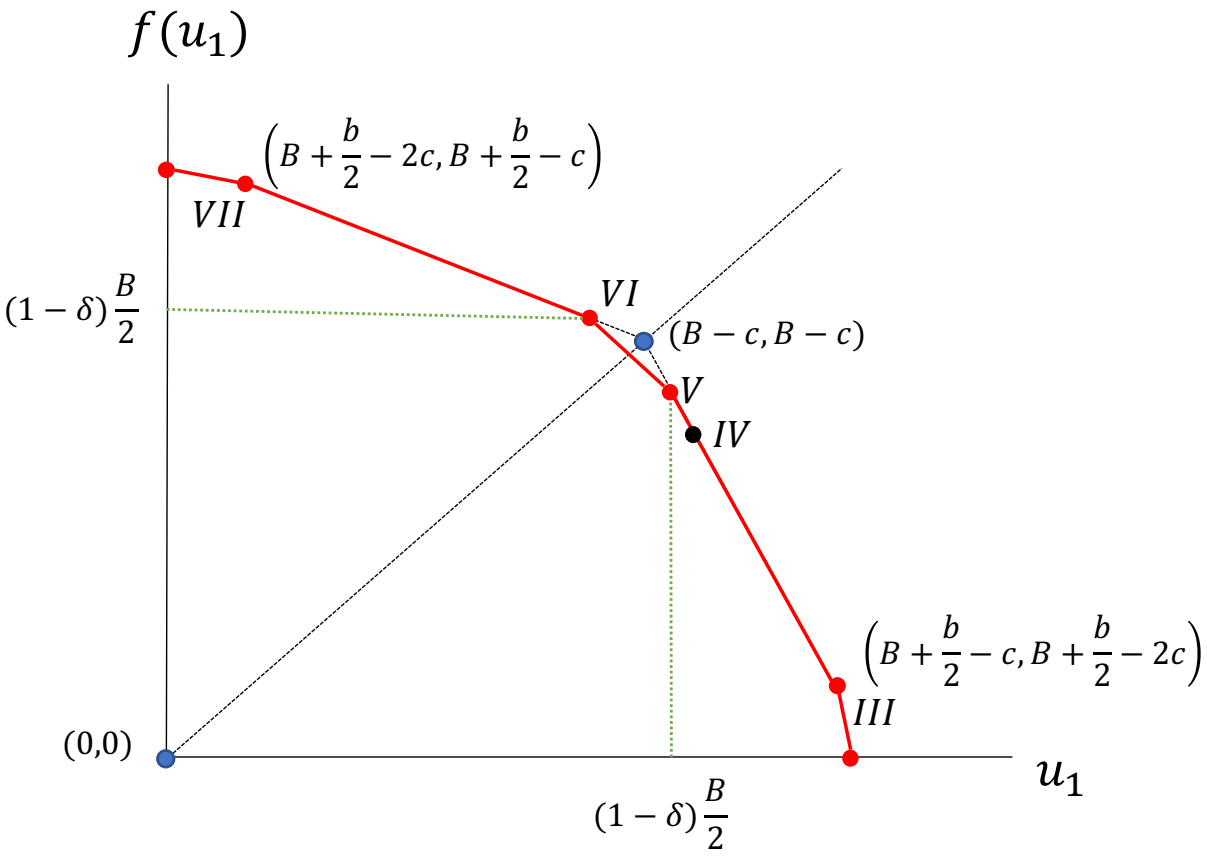

Figure 3: Equilibrium Payoff Frontier

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 surplus. 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, which we will discuss 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 payoffs on the frontier are all the same, and they are sustained by randomization between points V and VI. In this region, randomization is needed because the players cannot choose pure actions to reach the equilibrium payoff pair. The payoffs of both players in this region are less than $\frac{1}{2}(1-\delta)B$,

which is the minimal payoff to induce working under the default governance structure. As a result, regardless of who becomes the manager in this period, the manager cannot be motivated to work. 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. Part 3, therefore, implies that to induce both players to work, one of them must receive at least $\frac{1}{2}(1-\delta)B$.

We now describe the optimal governance structure and equilibrium strategies. For simplicity, we describe only the organization that gives Player 1 a payoff of $\frac{1}{2}(1-\delta)B$.

Proposition 7 (**Optimal Organization**). *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.*

Part 1 of Proposition 7 shows that, as in the case of stationary optimal organizations, the optimal governance structure is fully centralized, with Player 1 being the designated enforcer in every period. The advantage of full centralization is that it eliminates the need to consider Player 2's incentive constraint to work. In other words, specialization in enforcement, which implies a designated managerial role given to the same player for all periods, is efficient for incentive provision.

The structure of the optimal equilibrium play in Part 2 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 the similarity, the reason for backloading is somewhat different. The logic in this model 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. 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 equi-

librium play unchanged. It can be checked that this modification remains an equilibrium. It increases the joint payoff of the players, and it reduces the payoff of Player 1.

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, even if there is no imperfect information[22]. The price of efficiency in the earlier periods of the relationship is that Player 2 needs to overwork 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 declines.

From an economic perspective, the equilibrium described in Proposition 7 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 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.

5 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

[22]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), 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)).

payoff asymmetry to incentivize the monitor. At some level, the asymmetry in payoff appears to suggest that all the power is in the hands of the manager. But if this were true, the manager wouldn't be induced to work. The subordinate also has power over the manager via his right to exit. If the manager does not work, then the subordinate will take his outside option in the future. However, the threat of exit alone may not be sufficient to induce the manager to work, especially when the total surplus is low. By allocating power to the manager, more of the surplus goes to the manager. If the manager is sufficiently powerful, the subordinate's threat of exiting becomes credible and induces the manager to work. That is, by giving more power to the manager, the subordinate gains power over him. This is the paradox of power.

It is perhaps helpful to compare this point to Alchian and Demsetz's (1972) famous statement that there is no difference in power between firms and markets. As Alchian and Demsetz emphasize, power emanates from the option to exit (and withhold future business) and to sue. In our model, an autonomous organization may design a payoff structure that facilitates the use of power through exit.

Our model shows that, in the absence of external enforcement, self-executing contracts can be used to support relational incentives. Thus, in a sense, 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.

A Proofs

Proof of Proposition 1. Denote the players' expected equilibrium payoffs as u_i for $i \in \{1, 2\}$. Without loss of generality, we 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 given by $(1 - \delta)\left(c - \frac{B}{2}\right) \leq \delta u_2$. 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 she 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 $g_t = 1$. □

Proof of Proposition 2. Without loss of generality, we assume that $u_1 \leq u_2$. Then, by Lemma 1, Player 1 shirks whenever she has the whip. Suppose a stationary organization is power-blind. Power-blindness implies that player 1 always shirks. Because Player 1 always chooses $e_1 = 0$, and $B/2 < c$, Player 2 will never choose $e_2 = 1$. Similarly, because $(B + b)/2 < 2c$, he will never choose $e_2 = 2$. It follows that $e_2 = 0$.

Next, suppose that a stationary autonomous organization is identity-blind. Recall that Player 1 does not put in effort when she has the whip (because $u_1 \leq u_2$). Now, suppose that she does not have the whip. If she is forced to put in $e_1 = 1$, her participation constraint is given by

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

where u_c is Player 1's continuation payoff. Because $u_1 \leq u_2$ 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 $e_1 = 2$, her 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 1 shirks, identity-blindness implies that Player 2 also shirks. □

Proof of Lemma 2. Consider a stationary equilibrium that maximizes the discounted sum of payoffs. Let (u_1, u_2) be the expected payoff from the stationary allocation of the whip.

Without loss of generality, assume that $u_1 \geq u_2$. Then, by Lemma 1, Player 2 does not put in effort when he has the whip, i.e., $e_2^2 = 0$.

When Player 1 has the whip and chooses effort $e_1^1 > 0$, we must have $e_1^1 = 1$ and $e_2^1 = 2$. To see this, notice that if $e_2^1 < 2$, then Player 1's payoff is lower than $B - c$ when she has the whip, and her payoff is non-positive when Player 2 has the whip (as he chooses $e_2^2 = 0$). Therefore, $u_1 < B - c$, implying that because $\delta < \delta^{FB}$, the future loss from deviating to $e_1^1 = 0$ will be smaller than the short-term gain. This shows that when $e_1^1 > 0$, we must have $e_2^1 = 2$. Furthermore, when $e_1^1 = 2$ and $e_2^1 = 2$, the same argument as above shows that $u_1 < B - c$ and Player 1 gains by choosing $e_1^1 = 0$ because $\delta < \delta^{FB}$. Thus, when $e_1^1 > 0$, we must have $e_1^1 = 1$ and $e_2^1 = 2$.

The discussion above implies that we can restrict attention to two classes of equilibrium. In the first class, $e^1 = (1, 2)$ and $e^2 = (e, 0)$, where $e = 0, 1, 2$. In the second class, $e^1 = (0, e)$ and $e^2 = (e', 0)$, where $e, e' = 0, 1, 2$. In the second class, the relevant constraint is the participation constraint. Because the pair $e^1 = (0, 1)$ and $e^2 = (1, 0)$ have the highest joint payoff within this class, this case is the easiest to satisfy the participation constraint. However, this action profile is identity-blind, and thus not enforceable (see Proposition 2). Therefore, the whole second class of equilibria is eliminated.

□

Proof of Proposition 3. Let (u_1^i, u_2^i) be the payoff pair when Player i has the whip and (u_1, u_2) be the expected payoff from the stationary allocation of the whip.

(i) Analysis of the action profile $e^1 = (1, 2)$ and $e^2 = (0, 0)$. We can calculate the expected payoffs of both players as follows:

$$u_1^1 = (1 - \delta)\left(B + \frac{b}{2} - c\right) + \delta u_1,\ u_2^1 = (1 - \delta)\left(B + \frac{b}{2} - 2c\right) + \delta u_2,$$

$$u_1^2 = \delta u_1,\ u_2^2 = \delta u_2,$$

$$u_1 = pu_1^1 + (1 - p)u_1^2, u_2 = pu_2^1 + (1 - p)u_2^2.$$

Solving these equations gives $u_1 = p(B + \frac{b}{2} - c)$ and $u_2 = p(B + \frac{b}{2} - 2c)$. (Note that these

expressions can be obtained directly because in each period, with probability p Player 1 gets $B + \frac{b}{2} - c$ and Player 2 gets $B + \frac{b}{2} - 2c$, while both players earn zero with probability $1 - p$).

To sustain (u_1, u_2) as equilibrium payoffs, there are both incentive constraints and participation constraints:

$$\text{(IC)} \quad u_1^1 \geq (1-\delta)\frac{B+b}{2},$$

$$\text{(PCs)} \quad u_1^1 \geq 0, u_1^2 \geq 0, u_2^1 \geq 0, \text{ and } u_2^2 \geq 0.$$

Notice that the two participation constraints of Player 2 are automatically satisfied because $B + \frac{1}{2}b \geq 2c$, and that $u_1^1 \geq 0$ implies that $u_1^2 \geq 0$ holds. Thus, if Player 1's IC constraint holds, her participation constraints also hold. The IC constraint can be written as:

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

Solving this constraint shows that for this action profile to be an equilibrium, we need the following:

$$p = \frac{(1-\delta)(2c-B)}{\delta(2B+b-2c)} =: p_1(\delta).$$

Notice $p_1(\delta)$ decreases in δ because $(1-\delta)/\delta$ decreases in δ. When $\delta = (2c-B)/(B+b) =: \delta_1$, $p_1(\delta) = 1$. It follows that the action profile of $e^1 = (1,2)$ and $e^2 = (0,0)$ can be sustained by a stationary SPE if and only if $\delta \geq \delta_1$ and $p \geq p_1(\delta)$. This finishes the proof of the first part.

(ii) Analysis of the action profile $e^1 = (1,2)$ and $e^2 = (1,0)$. We can calculate the expected payoffs of both players as follows:

$$u_1^1 = (1-\delta)\left(B + \frac{b}{2} - c\right) + \delta u_1,\ u_2^1 = (1-\delta)\left(B + \frac{b}{2} - 2c\right) + \delta u_2,$$

$$u_1^2 = (1-\delta)\left(\frac{B}{2} - c\right) + \delta u_1,\ u_2^2 = (1-\delta)\frac{B}{2} + \delta u_2,$$

$$u_1 = pu_1^1 + (1-p)u_1^2, u_2 = pu_2^1 + (1-p)u_2^2.$$

Solving these equations gives that $u_1 = p(B + \frac{b}{2} - c) + (1-p)(\frac{B}{2} - c)$ and $u_2 = p(B + \frac{b}{2} - 2c) + (1-p)\frac{B}{2}$. (Note that these expressions can be obtained directly because in each period, with probability p Player 1 gets $B + \frac{b}{2} - c$ and Player 2 gets $B + \frac{b}{2} - 2c$, while with probability $1-p$ Player 1 gets $\frac{B}{2} - c$ and Player 2 gets $\frac{B}{2}$).

To sustain (u_1, u_2) as equilibrium payoffs, there are both incentive constraints and participation constraints:

$$\text{(IC)} \quad u_1^1 \geq (1-\delta)\frac{B+b}{2},$$

$$\text{(PCs)} \quad u_1^1 \geq 0, u_1^2 \geq 0, u_2^1 \geq 0, \text{ and } u_2^2 \geq 0.$$

Notice that the two participation constraints of Player 2 are automatically satisfied because $B + \frac{1}{2}b \geq 2c$, and again the IC constraint implies Player 1's participation constraints. The IC constraint can be written as:

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

Solving this constraint shows that for this action profile to be an equilibrium, we need the following:

$$p \geq \frac{2c - B}{\delta(B+b)} =: p_2(\delta).$$

Notice $p_2(\delta)$ decreases in δ because $1/\delta$ decreases in δ. When $\delta = \delta_1 = (2c - B)/(B + b)$, $p_2(\delta) = 1$. It follows that the action profile of $e^1 = (1, 2)$ and $e^2 = (1, 0)$ can be sustained by a stationary SPE if and only if $\delta \geq \delta_2$ and $p \geq p_2(\delta)$. This finishes the proof of the second part. □

Proof of Corollary 1. Without loss of generality, suppose $u_1 \geq u_2$. Lemma 1 implies that payoffs in which Player 2 works or overworks when he is the manager cannot be sustained. Thus, if $p < 1$, the best profile that can be sustained is $e^1 = (1, 2)$ and $e^2 = (1, 0)$, which is Case 2 in Proposition 3. Proposition 3 thus implies that, if $p < 1$, the maximum joint payoff the organization can produce is $p(B + b - c) + B - c$ if $p \geq p_2(\delta)$, $p(2B + b - 2c)$ if $p \in [p_1(\delta), p_2(\delta))$, and zero if $p < p_1(\delta)$. Because the maximum joint payoff is increasing in p (and strictly increasing if $p \geq p_2(\delta)$), the optimal organization

must have $p = 1$. □

Proof of Proposition 4. Set $e^1 = (1,2)$ and $e^2 = (1,1)$, and let (u_1^i, u_2^i), $i = 1,2$, denote the associated payoff profiles. Consider the action profile that randomizes between e^1 and e^2 with probabilities α and $1-\alpha$. Let (u_1, u_2) be the expected payoffs for this mixed action profile. In each period, Player 1 gets $B + \frac{b}{2} - c$ with probability α and $B - c$ with probability $1-\alpha$, thus, her expected payoff equals $B - c + \alpha\frac{b}{2}$. Similarly, Player 2's expected payoff equals $B - c + \alpha(\frac{b}{2} - c)$. The payoff pairs $\{(u_1^i, u_2^i)\}_{i=1,2}$ can be written as

$$u_1^1 = (1-\delta)\left(B + \frac{b}{2} - c\right) + \delta u_1, u_2^1 = (1-\delta)\left(B + \frac{b}{2} - 2c\right) + \delta u_2$$

$$u_1^2 = (1-\delta)(B-c) + \delta u_1, u_2^2 = (1-\delta)(B-c) + \delta u_2,$$

where $u_1 = B - c + \alpha\frac{b}{2}$ and $u_2 = B - c + \alpha(\frac{b}{2} - c)$.

Because $u_2^2 > u_2^1$, we need to check only one of Player 2's participation constraint: $u_2^1 \geq 0$, which holds for any δ and α because of Assumption 2. Similarly, because $u_1^1 > u_1^2$, we need to check only one of Player 1's incentive-compatibility constraints: $u_1^2 \geq (1-\delta)\frac{B}{2} \Rightarrow \delta u_1 \geq (1-\delta)\left(c - \frac{B}{2}\right)$. Solving Player 1's incentive constraint implies

$$\alpha \geq \frac{(1-\delta)\left(c - \frac{B}{2}\right) - \delta(B-c)}{\delta\frac{b}{2}} =: \alpha_1(\delta).$$

Because $(1-\delta)/\delta$ decreases in δ and $2c > B$, $\alpha_1(\delta)$ decreases in δ. Note that $\alpha_1(\delta_1) = 1$. Because $\alpha_1(\delta)$ decreases in δ, we know that if $\delta \in [\delta_1, \delta^{FB})$, $0 \leq \alpha_1(\delta) \leq 1$.

Finally, to maximize the joint surplus the designer must choose the lowest possible α to minimize the frequency of overworking. Thus, $\alpha = \alpha(\delta)$ is the optimal randomization weight. □

Proof of Proposition 5. **Part 1**. As before, to streamline the analysis, we consider only the case in which $\mu_i(g_t)$ is degenerate; the case of non-degenerate $\mu_i(g_t)$ can be proven using the same line of arguments. Suppose $p = 0.5$. Without loss of generality, suppose that in equilibrium, $u_1 \geq u_2$. From Lemma 1, Player 2 shirks as manager. Proposition 3 implies that effort profile $e^1 = (1,2)$ cannot be supported with either $e^2 = (0,0)$ or $e^2 = (1,0)$

(choosing $e^2 = (2,0)$ is also not an option because it contradicts $u_1 \geq u_2$). If we choose either $e^1 = (1,1)$ or $e^1 = (2,2)$, we must have $e^2 = (0,0)$, otherwise we have $u_1 < u_2$. But then Assumption 3 implies that these profiles are also not sustainable. Setting $e^1 = (0,1)$ with $e^2 = (0,0)$ or $e^2 = (1,0)$ does not meet Player 2's participation constraint (setting $e^2 = (2,0)$ contradicts $u_1 \geq u_2$). The latter argument also applies to $e^1 = (0,2)$. Finally, profiles $e^1 = (1,0)$, $e^1 = (2,0)$ and $e^1 = (2,1)$ all contradict $u_1 \geq u_2$ (given that Player 2 must shirk as manager). The only e^1 that is not ruled out is $(0,0)$, which must imply $e^2 = (0,0)$. This action profile yields zero surplus and is trivially a Nash equilibrium. This equilibrium is inefficient because the equilibrium described in Proposition 4 delivers a strictly positive payoffs to both players.

Part 2. It is implied by Corollary 1.

Part 3. A decentralized autonomous organization can only implement $(0,0)$ (see Part 1), thus it cannot be optimal for $\delta \geq \delta_1$, as the equilibrium described in Proposition 4 delivers strictly positive payoffs. If the organization is non-autonomous, it can implement the first-best effort profile if $(1-\varrho)B \geq c$. In this case, the joint surplus is $2(1-\varrho)B - 2c$. To be an optimal organization, this surplus must be greater than that implied by Proposition 4, which is $2(B-c) + \alpha_1(\delta)(b-c)$. These two inequalities jointly imply $2\varrho B < \min\{2(B-c), \alpha_1(\delta)(c-b)\}$. □

Proof of Proposition 6. 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 $(u_1, f(u_1))$. In this direction, Proposition 4 has shown that if $u_1 \geq \frac{1-\delta}{\delta}(c - \frac{B}{2})$ and (u_1, u_2) belongs to the line segment considered, the payoff pair can be sustained by some SPE.

Consider the point on the line segment where $u_1 = \frac{1-\delta}{\delta}(c - \frac{B}{2})$ (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 action profile randomizes between $(1,2)$ and $(1,1)$ with probabilities $\alpha_1(\delta)$ and $1-\alpha_1(\delta)$. In particular, when profile $(1,1)$ realizes, Player 1's

payoff can be written as

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

which equals $\frac{1}{2}(1-\delta)B$. Then, because no player deviates given this contingency, the point with $u_1=\frac{1}{2}(1-\delta)B$ on the line segment between $(B-c,B-c)$ and $(B+\frac{1}{2}b-c,B+\frac{1}{2}b-2c)$ (this is point V in Figure 3, which is to the left of IV) 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 $\alpha_1(\delta)$ and $1-\alpha_1(\delta)$ from period 2 on. Player 1's IC constraint at $t=1$ is satisfied with equality ($\frac{1}{2}(1-\delta)B\geq\frac{1}{2}(1-\delta)B$) and Player 2's participation constraint at $t=1$ also trivially holds. This proves 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 7. 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 6 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. This completes the proof. □

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.

——— (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.

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.

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.

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," *AEJ Micro*, 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.