NORTHWESTERN UNIVERSITY

Essays on Incentives in Organizations

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Managerial Economics and Strategy

By

Ruozhou Yan

EVANSTON, ILLINOIS

June 2022

Abstract

This dissertation studies the limitations of incentive design in organizations and how they lead to inefficient outcomes.

Chapter 1 studies how a coalition coordinates its members when they freely join and leave. It characterizes the conditions under which such coordination prevents the coalition from forming in the first place. In the model, a coordinator recommends actions to the members after they form the coalition. Members can disobey the recommendation by leaving the coalition, but doing so requires them to forego some synergies. In equilibrium, the coordinator's recommendation keeps members from leaving the coalition, but it may induce actions that make some members worse off than had the coalition not formed at all. In this case, these members protect themselves by refusing to join the coalition in the first place. Building on this result, the chapter also studies how the coalition can implement richer decision-making processes—for instance, by delegating decision-making authority or requiring consensus decision-making—to convince members to join.

Chapter 2 studies how incentives are affected by an agent's ability to worsen the quality of the performance measure that the principal uses. In the model, an agent sets up ways to manipulate performance before incentives are put in place. The principal designs a flatter incentive scheme when the agent manipulates more. This lowers the agent's rent

extraction and deters him from manipulating too much. Better alignment of the performance measure with the principal's objective reduces this deterrence effect and results in more manipulation. Chapter 2 then explores a two-period setting, where setting up to manipulate the second-period incentive reduces the agent's first-period performance. To encourage first-period effort and discourage the set-up, the principal optimally "overloads" the first-period incentive.

Chapter 3 studies the lemons problem in a delegation setting. A principal with private information about the returns to effort of a project chooses whether to delegate it to an agent. The principal optimally delegates only low-return projects. Therefore, the act of delegating becomes a negative signal to the agent. The agent is thus less willing to exert effort, so the principal optimally delegates even fewer projects. If the principal can commit to a set of projects to delegate, then she optimally delegates projects with a medium return.

Acknowledgment

I would like to give special thanks to my committee chair, Professor Daniel Barron, for his continuous encouragement and unwavering support. My growth as a researcher and as a person has greatly benefited from the numerous conversations that we have had throughout the years.

I would like to also express my gratitude to my other advisors, Professors Michael Powell, Niko Matouschek and Bruno Strulovici, who have provided invaluable time and feedback for my research. Discussions with Professors Tim Feddersen, Nicola Persico, Luis Rayo and Mark Satterthwaite have also been enormously helpful.

Many colleagues and friends have been for helping me shape ideas: Zuheir Desai, Luca Bittarello, Román Acosta, Sanket Patil, Henrique Castro-Pires and Andres Espitia. I also thank them for moral support when I went through difficult times.

My parents, Chengzhi Yan and Jing Zou, have always shielded me from extra pressure and extended support for my decisions. My girlfriend Yue Yin has provided me with good food, accommodation, consolation, and company. I could not have finished this dissertation without their love.

Table of Contents

List of Figures

Introduction

Incentives are ubiquitous in organizations, and their design affect the sustainability and functioning of organizations. The studies in this dissertation focus on the limitations of incentives in organizations and how these limitations lead to inefficiencies.

People make decisions with the intent of affecting the incentive design in future. Chapters 1 and 2 highlight the distortions created by a person's ability to manipulate future incentives. In the context of coalition formation, Chapter 1 studies why people choose to inefficiently stay away from a coalition to stop the coalition from coordinating their actions. In the context of performance evaluation in organizations, Chapter 2 studies how people take actions that allow them to manipulate the performance measures in their future incentive schemes.

One overarching result of these settings is that improving the ability to provide incentives can perversely affect prior decisions, exacerbating inefficiency. Chapter 1 shows that an increase in the coalition's ability to coordinate a member can deter the member from joining the coalition. Chapter 2 shows that making a performance measure less sensitive to manipulation can actually lead to more manipulation, harming the incentive designer. Policies that are designed to improve incentives may be counterproductive if such strategic responses are not taken in to account.

Incentive design may also be limited by the instruments available to the designer. In the context of performance evaluation, inefficiency partly stems from the imperfect alignment between the performance measure and the incentive designer's interest. Chapter 3 studies the case where the incentive takes the form of a choice between centralizing and delegating projects. When the principal has private information about the returns to effort of these projects, a lemons problem arises, leading to excessive centralization relative to first-best.

CHAPTER 1

Incentives in Coalitions

When people join forces, they often build coalitions to coordinate their efforts. Activists build social movements to effect change; firms form standard-setting organizations to regulate industries; states establish regional unions and international organizations to harmonize policies. However, coordination must be tactful as people may not like how they are coordinated and can leave the coalition if they wish. Coalition-building is therefore a balancing act between improving coordination and keeping members on board.

This chapter studies how a coalition coordinates its members. It shows that the need to keep members from leaving limits the coalition's power to coordinate their actions. Consequently, maximal coordination may require inducing actions that make some members worse off than they would be had they stayed out of the coalition. This happens when the coalition has much more power over these members than over others, and such a coalition thus fails to form in the first place. Building on this result, this chapter further studies how the coalition can implement richer decision-making processes to discipline how power is used and thus to convince everyone to join.

One example of the difficulty of building a coalition is the organization of the March on Washington in 1963. Civil rights, labor, and religious organizations formed a coalition to have a unified voice despite their distinct ideologies and strategies. The organizers of the March, A. Philip Randolph and Bayard Rustin, took up the coordinator role on many issues where participants disagreed and compromises had to be made. For instance, the March was held at the Lincoln Memorial rather than at the Capitol; white participants were invited, but communists were not (Barber, 2002).

Securing compromises was difficult when disagreements emerged unexpectedly. On the night before the March, several civil rights leaders strongly objected when they found out that John Lewis would harshly criticize the federal government and the pending civil rights bill in his speech. The organizers could not censor him since he could always break away and make the speech elsewhere. Meanwhile, John Lewis also saw value in solidarity with other civil rights leaders and decided not to "ruin [the March]" but to "stay together" (Lewis and D'Orso, 1999, p. 223). Compromises were reached at the last minute: the other leaders acquiesced as John Lewis made explicit the reserved support for the bill but kept much of the angry tone.

Other black leaders refused from the very beginning to be put in such a situation. Malcolm X, for instance, stayed away from the March. In his autobiography, he clearly expressed rejection of all the coordination made by march organizers, mocking the order and civility as a "farce" and the participation by white people as an "integrated picnic" (X and Haley, 1965, p. 281).

As this episode illustrates, coordination is shaped by how much coalition members are willing to compromise before they decide to walk away. However, the ability to walk away does not always convince them to join; they stay out if they dislike how the coalition coordinates once they join. In fact, the decision to stay out is driven by a crucial distinction between staying out and leaving after joining: by joining the coalition, a member changes how the coalition coordinates *other* members' actions, thereby changing the outside option that he faces if he were to leave the coalition later. In other words, Malcolm X stayed out because his participation would allow others to behave "worse" from his perspective.

This chapter formalizes this argument and derives the condition under which some members refuse to join the coalition. Section 1.1 sets up a two-member baseline model. Members (both "he") decide whether to form a coalition, which creates synergies for them (e.g., the value of solidarity). Members take any action they want if the coalition is not formed. If the coalition is formed, a third player, the coordinator ("she"), recommends a pair of actions for them to take. Members can nevertheless ignore the coordinator and take any action by leaving the coalition and forgoing the synergies.

Members' actions are about an issue of disagreement (e.g., their speeches). On such an issue, a member cares about coordination of their actions (e.g., how well aligned the speeches are with each other), as well as the consistency of both actions with his ideal action (e.g., how critical the speeches are towards the federal government). As members disagree on the ideal way to be coordinated, better coordination is achieved only by moving some member's action farther away from his ideal action.

Section 1.2 analyzes how members are coordinated after they form the coalition and how such coordination affects their incentives to join in the first place. A measure of the coordinator's power over a member is derived as how far she can distort a member's action away from his best response to the other member's action. Her power over both members shapes how actions are coordinated, particularly how far each member's action is from his ideal action. This distance measures the compromise that the coordinator secures from a member.

The main result of this model is that a coalition can form if and only if it allows the coordinator to secure more compromise from *both* members, and this happens if and only if her power over both members is sufficiently balanced. This result builds on the observation that a member joins the coalition only if the coordinator can secure more compromise from *the other* member. A member's outside option—to subsequently leave the coalition and forgo the synergy—determines his payoff from joining the coalition. Therefore, he joins the coalition only if doing so changes his outside option for the better, i.e., if the coalition secures more compromise from the other member.

Securing more compromises from both members requires the coordinator *not* to have much more power over one member than over the other. This result follows from the observation that compromises are strategic substitutes: exerting power over a member to secure more compromise from him always reduces the compromise from the other member. Suppose the coordinator exerts too much power over a member—whom we refer to as the weak member—to secure a large compromise. In that case, the other member makes little compromise, even less than his compromise had the coalition not formed. The worsening of outside option thus deters the first member from joining.

The second part of the chapter studies remedies that can convince members to join when they would not do so in the baseline setting. The decision-making process in the baseline model is enriched in different ways to analyze different decision-making protocols that can allow the coalition to form. Section 1.3 begins by showing that if the coordinator can commit not to make any recommendation, then the coalition always forms. However, doing so does not improve coordination as members optimally take actions as if the coalition has not formed. The chapter then explores remedies that induce a different pair of actions to coordinate members in a Pareto efficient way. Section 1.3.2 begins by showing that any such efficient remedy must limit the power exerted over the weak member so that the coordination does not make him worse off. However, it must not limit this power too much, for otherwise, an imbalance is created in the opposite direction, keeping the other member away.

This feature is reflected in the efficient remedies subsequently analyzed. Section 1.3.2 proceeds to study delegated coordination, in which the coordinator allows someone else to make the recommendation. Doing so allows the coalition to form provided that the weak member's interest is sufficiently but not excessively represented. In particular, the weak member himself should never coordinate on behalf of the coalition.

Section 1.3.3 then shows that giving the weak member a veto over the coordinator's recommendation also allows the coalition to form provided that his veto can induce an outcome sufficiently good but not too good for himself. It further shows that consensus—giving both members a veto—can augment any inefficient remedy that the coalition can adopt. Doing so constitutes a Pareto improvement upon this inefficient remedy, yielding more coordination to everyone's benefit.

Section 1.4 generalizes the previous analyses to $N \geq 3$ members, and derive the condition for the N-member coalition to form. The analyses show that members can be partitioned into two groups, and the coalition is formed if the "aggregate power" over the two groups is in balance. In this case, each member secures more compromises from other members "on average." This generalization also reveals a new channel through which members change the balance of power: their ideal actions affect how they are partitioned into those two groups and thus the balance of "aggregate power." Regarding remedies, the analyses show that the efficient remedies identified for the two-member coalition continue to be efficient, and they confirm the intuition that the weak group—the group of members under too much "aggregate power"—must be sufficiently but not excessively protected for these remedies to work. Finally, consensus continues to Pareto improve any inefficient remedy.

Applications. Besides social movements, other coalitions face similar misalignment of interest of their members and must coordinate after their formation. In standard-setting organizations (SSOs), firms make compromises when developing standards for new technologies (Bonatti and Rantakari, 2016). In the European Union, member states make compromises when harmonizing policies (Reh, 2012). Coalition-building also arises in other types of organizations when buy-in from multiple parties is necessary but cannot be imposed. As an example, the job of product or project managers in modern technology firms often involves building a coalition: they need to convince multiple functional teams to do things so that a product or a project can succeed,[1] but they cannot rely on monetary incentives or formal authority (Gemmill and Wilemon, 1972; Austin, 2017).

Section 1.2.4 applies this model to identify two channels through which different factors affect the balance of power and therefore the formation and coordination of such coalitions. The first channel concerns the imbalance of synergies, which is affected by asymmetry in members' sizes, resources, and external opportunities. The imbalance of synergies explains, for example, why developing country NGOs are reluctant to partner with NGOs from developed countries. The second channel concerns the mismatch of priorities, which

[1] In particular, these teams need to convinced to do things differently than they individually prefer. For instance, a project that improves user experience may require changes that neither the design team nor the engineering team finds appealing; a project that improves product recommendations may require tweaking metrics and algorithms that have been independently optimized by the organic content team and the advertisement team.

is affected by the heterogeneity in members' identities, beliefs, and concerns. For instance, a political coalition may appear "hijacked" by its radical wing because, on a "fringe" issue, it has more power over its moderate members than over its radical members who care immensely.

The need to address such power imbalances justifies many decision-making protocols in coalitions. Excluding issues from a coalition's scope of coordination is one such remedy that allows the coalition to form, and it is reflected in the divisions of competences of the EU and the enumerated powers of the US federal government. Delegating coordination to a group that sufficiently represents the interest of weak members is also a remedy; this result justifies the member representation requirement that a growing number of SSOs adopt for their working groups, as well as the general trend of prioritizing weak members in decision-making of social movement coalitions.

Finally, the observation that consensus decision-making Pareto improves other remedies explains its prevalent use in all types of coalitions. The analysis suggests that consensus decision-making can help bring potential members into the coalition, justifying the extra costs of building consensus. It also suggests a way to reduce these costs: by not seeking approval from members whose veto does not discipline the coordinator's decision-making.

Related Literature. This chapter makes two main contributions. First, it develops a theory of coordination and participation and reveals their connection when the coordinator has limited power over the members, which contributes to the literature on the allocation of authority and coordination. Second, it studies decision-making protocols that discipline coordination to induce participation and provides a formal understanding of the strategic value and effective use of these practices in real-world coalitions, which contributes to the literature on SSOs and social movements.

This chapter explores when centralized authority improves coordination, and therefore contributes to the literature on the allocation of authority (Dessein, 2002; Harris and Raviv, 2005; Swank and Visser, 2015; Deimen and Szalay, 2019) and its effect on coordination (Alonso et al., 2008; Rantakari, 2008; Dessein et al., 2016; Li and Weng, 2017). A major departure from the literature is that, in a model with complete information, members are allowed to ignore centralized coordination at the cost of leaving the coalition. This generates a novel constraint on the coordinator's power to induce actions and shapes centralized coordination. Similarly, Marino et al. (2010) allow disobedience in an employment relationship where the monetary contract and dismissal play a similar role. However, as in the rest of the literature, they let the "center" (the coordinator) allocate authority. In contrast, this model lets members jointly determine whether coordination is centralized, which means that coordination must give *everyone* the incentive to participate; this tension is at the crux of my analysis.[2]

In equilibrium, members' actions in a coalition affect their incentives to join a coalition. My analysis thus contributes to the vast literature on coalition formation (for recent examples, see Acemoglu et al., 2012; Barberà et al., 2015; Morelli and Park, 2016; Gallo and Inarra, 2018; see Ray and Vohra, 2015 for an excellent survey). The major departure

[2] This model thus blurs the line between centralization and delegation due to the distinction between formal and real authority as highlighted in the literature (Aghion and Tirole, 1997; Baker et al., 1999; Alonso and Matouschek, 2007; Li et al., 2017). The centralized coordination in my model can be understood as both members delegating the real authority over their actions to the coordinator (i.e., the center); as they still retain formal authority, they can ignore the coordinator's recommendation. The remedies must ensure that the coordinator uses its real authority in a balanced way so that both members are willing to delegate.

of this paper is its explicit modeling of how members can deviate and leave a coalition *after* its formation, which highlights the asymmetry in a member's options before and after coalition formation as driving a member's decision to join a coalition.

The main result of this model shows how *ex post* decisions affect *ex ante* incentives and is thus connected to research on similar mechanisms, in particular the literature on hold-up (Grossman and Hart, 1986; Hart and Moore, 1990). Similar to these models, the lack of commitment can lead to a distortion in a member's *ex ante* decision (i.e., whether to join the coalition). The special feature of my model is that, when the coalition is formed, a member's outside value decreases in the coordinator's power over him and increases in her power over the other member. This feature is why the balance of power over *both* members determines if a member's *ex ante* decision is distorted.

This chapter is also related to the literature on the institutional design of SSOs (Farrell and Simcoe, 2012; Simcoe, 2012, 2014; Bonatti and Rantakari, 2016), which mainly explores the trade-off between inefficient delay and suboptimal choice of standards under consensus decision-making. The current model abstracts from delays in SSOs and instead focuses on members' voluntary participation decisions in anticipation of the SSO's choice of standards (i.e., coordination). It explains why consensus decision-making is used in such settings despite its costs.

Finally, this chapter also contributes to the interdisciplinary literature on coalitions in social movements. This literature explores how coalition formation and success are affected by various factors such as resources, ideologies, and identities (Van Dyke and Amos, 2017; Curtis and Zurcher, 1973; Bystydzienski and Schacht, 2001; Arnold, 1995; Ganz, 2000;

Levi and Murphy, 2006). The current model provides a unifying framework for understanding how these factors work: they affect power balance on issues where members disagree. This framework highlights the strategic benefit of democratic practices—such as consensus—in facilitating coalition-building. This framework also reconciles opposing views on the relationship between centralization and solidarity (see Zald and Ash, 1966 and Gamson, 1975 for supportive arguments; see Polletta, 2002 and Della Porta and Diani, 2006 for opposing ones). It suggests that centralization fosters solidarity only if movement participants agree to be coordinated, and limiting the coordination ability of the center can facilitate such agreement.

1.1. Baseline Model

This section sets up a two-member baseline model and discusses the modeling ingredients that go into this stylized game.

Two members ($i \in \{1, 2\}$) must decide whether to form a coalition, which creates synergy $\omega_i > 0$ for each of them if they form and sustain the coalition till the end of the game. Let member i's *participation decision* be $x_i \in \{0, 1\}$. If members do not form the coalition, i.e., $x_1 x_2 = 0$, they simultaneously each takes an action $a_i \in \mathbb{R}$ and the game ends. If members form the coalition, i.e., $x_1 x_2 = 1$, a third player—the coordinator (C)—recommends to the members a pair of actions $(a_1^R, a_2^R) \in \mathbb{R}^2$. This recommendation does not automatically translate into actions, since each member i is still free to take any action $a_i \in \mathbb{R}$. However, if some member i wants to take an action $a_i \neq a_i^R$, he must leave the coalition at the same time and consequently *both* members lose the synergies that the

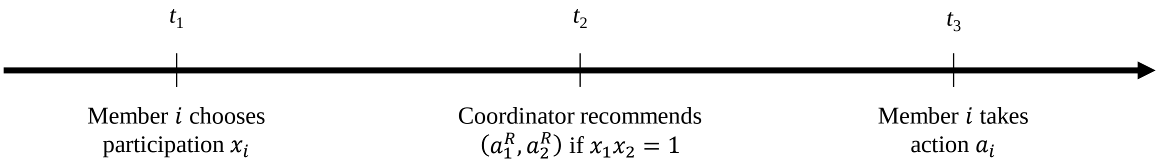
t_1
t_2
t_3
Member i chooses participation x_i
Coordinator recommends (a_1^R, a_2^R) if $x_1 x_2 = 1$
Member i takes action a_i

Figure 1.1.1. Baseline model timeline

coalition can create. There is no private information and decisions are publicly observed. The timing of the game is summarized in Figure 1.1.1.

Members' actions are about an issue of disagreement. On that issue, each member i has an ideal action $\theta_i \in \mathbb{R}$. For example, the issue could be about speeches; θ_i represents how harsh member i likes speeches to be and a_i represents how harsh his own speech is. Assume that $\theta_1 \neq \theta_2$ so that members actually disagree, and in this case assume without loss of generality that $\theta_1 < \theta_2$. Member i values *ideological consistency* and is better off when $|a_1 - \theta_i|$ and $|a_2 - \theta_i|$ decrease. He also values *coordination* and is better off when $|a_1 - a_2|$ decreases.

Each member i also benefits from the synergy ω_i if the game ends with *both* member staying in the coalition. Let $s \in \{0,1\}$ indicate whether members form *and* stay in the coalition, i.e.,

$$s = \begin{cases} 1 & x_1 = 1 \ \& \ x_2 = 1 \ \& \ a_1 = a_1^R \ \& \ a_2 = a_2^R \\ \\ 0 & \text{otherwise} \end{cases}$$

The synergies accrues to members if and only if $s = 1$.

Member i's overall payoff thus takes the following form

$$u_i(a_i, a_j, s) = s \cdot \omega_i - \left[\kappa_i (a_1 - a_2)^2 + (1 - \kappa_i) \sum_{j \in \{1,2\}} (a_j - \theta_i)^2 \right] \tag{1.1.1}$$

where the importance of coordination relative to the total ideological consistency is parameterized by $\kappa_i \in (0,1)$; the two extreme values are ruled out so that members always care about coordination but their preferences are still misaligned.

The coordinator only concerns with the coordination of the members' actions despite having no action of her own. Her payoff takes the following form

$$u_C(a_1, a_2, s) = s - (a_1 - a_2)^2$$

which is a special case of Equation (1.1.1) where $\kappa_C = 1$ and $\omega_C = 1$.[3]

The coordinator's strategy is her recommendation as a pair of actions $(a_1^R, a_2^R) \in \mathbb{R}^2$. Member i's strategy consists of his participation decision $x_i \in \{0,1\}$ and his action $a_i \in \mathbb{R}$ implicitly as a function of (x_1, x_2, a_1^R, a_2^R). The solution concept is Pure-Strategy Subgame-Perfect Nash Equilibrium. Any equilibrium is characterized by the payoff relevant triple (a_1, a_2, s), an equilibrium *outcome*.

If the coalition creates immense synergies so that the disagreement becomes trivial in comparison, then members' participation decisions would not be affected by their disagreement. We make the following assumption to rule out trivial disagreements; in addition, it guarantees that the coordinator cannot coordinate members perfectly (i.e., she cannot induce $a_1 = a_2$ in equilibrium) so that she has a unique optimal recommendation, which greatly simplifies my analysis.

Assumption (Sufficient disagreement)**.** $\theta_2 - \theta_1 > \frac{\sqrt{\omega_1}}{1-\kappa_1} + \frac{\sqrt{\omega_2}}{1-\kappa_2}$.

[3] It is without loss of generality to assume $\omega_C = 1$ when $\kappa_C = 1$. As becomes clear later in the analysis, the coordinator's ranking of the outcomes under consideration remains constant irrespective of the magnitude of $\omega_C > 0$.

The derivation of this assumption is shown in Appendix C.1.1.

In Stages 1 and 3, two equilibria may exist as members move simultaneously. In one equilibrium, both members stay out of or leave the coalition. Such an equilibrium exists because one member's decision to do so makes the other's decision irrelevant to the game's outcome.[4] In the other equilibrium, members form or sustain the coalition, and this equilibrium is selected throughout the paper.[5]

Discussion. The modeling elements reflect many observations from the organization of the March on Washington. For instance, the model assumes that the coordinator cannot commit to a specific recommendation before the coalition forms. An important reason is that although the coordinator and coalition members can foresee issues down the road, specifying coordination plan for all disagreements that can emerge is hardly feasible.[6] Despite all the coordination on the message of the March, e.g., the selection of slogans and songs (X and Haley, 1965), tension still erupted one night before the March because of John Lewis' speech.

The March also illustrates the synergies that a coalition can create for its members. One may interpret synergies as better use of shared infrastructure and expertise, greater publicity, value of solidarity or legitimacy from broad participation; they disappear if

[4] To be rigorous, such an equilibrium exists *generically* in Stage 3, as the coordinator may recommend the same action that a member would take if they were to leave the coalition, so both leaving the coalition is not an equilibrium. This caveat does not affect the following analysis.

[5] In Stage 1, this equilibrium is also selected by weakly dominant strategies. In Stage 3 when the coalition is formed, selecting this equilibrium can be justified by an argument similar to forward induction: because both members leaving the coalition induces an equilibrium outcome identical to the one induced by *not* forming the coalition, members must have chosen to form the coalition in anticipation of a different equilibrium outcome.

[6] This assumption echoes March and Simon's observation that, in an organization, "the set of activities to be performed is not given in advance, except in a most general way" (March and Simon, 1958, p. 26).

members part ways. A case in point is the live TV coverage and the wide audience that the March attracted. Had the March been split into many small ones, its messages (e.g., "I Have a Dream") would not have had the same reach and impact.

The March also demonstrates why members must leave the coalition if they choose to ignore the recommendation. Members often need to take the recommended actions if they wish to stay in the coalition, just as John Lewis had to edit his speech in order to take the podium. There is perhaps a deeper reason for the link between a member's disobedience and his exit. As long as members are free to leave, a coordinator's maximal penalty is to force a member out in order to deny him the synergy. It is also optimal for the coordinator to use this penalty to provide maximal incentive for members to follow its recommendation. This point is elaborated in Appendix A, which develops a richer model that endogenizes the disobeying member's choice to leave the coalition, the exclusion of one pair of actions (a_1^R, a_2^R) from penalty and the forcing out of any disobeying member. The coordinator uses them jointly to provide optimal incentive for members.

The coordinator is modeled to be neutral on the issue where members disagree, since a coordinator or an organizer is usually *not* mandated to take a position on such issues. As Gamson (1961) suggested, "the stability of a coalition requires *tacit neutrality* of the coalition on matters which go beyond the immediate prerogative." This intuition is partly confirmed in Section 1.3.2.2 where we effectively relax this assumption and show that the coalition never forms if either member gets to make decision for the coalition.

Notice that members are modeled as acting simultaneously and unable to reverse their actions if the coalition breaks up. In other words, members cannot adjust their actions after observing each other's actions. The results of this paper build on this feature, and

Section 1.2.3 offers a discussion on why this is a straightforward way to capture a salient consideration in members' participation decisions.

Also notice that synergies are measured in the same unit as the payoffs from the issue of disagreement; therefore ω_i should be understood as "normalized synergy." See Section 1.2.4 for how this distinction allows different applications of the model, strengthening its explanatory power. In addition, the intuition of the results do *not* depend on the additivity of the synergy with respect to the issue payoff. See the end of Section 1.2.3 for a discussion.

While transfers are assumed to be infeasible for simplicity, the ability to make contingent transfers *per se* does not change the intuition of the model; it only adds to its complexity, as the optimal transfer (from whichever player) depends on the members' utilities from money relative to the issue. The ability to *commit* to such a transfer before members' participation decisions, on the other hand, changes the nature of the problem. See the end of Section 1.2.3 for a discussion.

1.2. Coalition Formation

In this section, we characterize the equilibrium outcomes when members stay apart and when they form the coalition. To do so, we derive a measure of **power** as the coordinator's ability to induce actions distant from members' best responses to each other. By comparing members' preferences over the two outcomes, we derive the key result of the model: the coalition is formed and sustained if and only if the coordinator has sufficiently balanced power over the members. We conclude this section with a discussion of two causes of imbalance of power—the imbalance of synergies and the mismatch of priorities—and how they can explain various factors that affect coalition formation and divide.

1.2.1. Disintegration

We first study **disintegration**, the equilibrium outcome that follows if members stay apart after Stage 1. As members do not receive any synergy, they only concern with each other's action, and they each has a unique best-response. Consequently, a unique equilibrium outcome emerges, which is formalized in the following lemma.

Lemma 1.1. *In any subgame where $x_1 x_2 = 0$, a unique equilibrium exists in which member i takes action*

$$a_i^D = \frac{(1-\kappa_i)\theta_i + \kappa_i(1-\kappa_j)\theta_j}{1-\kappa_i\kappa_j}$$

where $i \in \{1,2\}$ and $j \neq i$.

Proof. If $x_1 x_2 = 0$ then $s = 0$ irrespective of (a_1, a_2). Member i's optimal action $a_i^{BR}(a_j)$ therefore satisfies

$$\begin{aligned} a_i^{BR}(a_j) &\equiv \text{argmax } u_i(a_i, a_j, 0) \\ &= \kappa_i a_j + (1-\kappa_i)\theta_i \end{aligned} \tag{1.2.1}$$

Let (a_1^D, a_2^D) be the pair of equilibrium actions; they must satisfy

$$\begin{cases} a_1^D = a_1^{BR}(a_2^D) \\ a_2^D = a_2^{BR}(a_1^D) \end{cases}$$

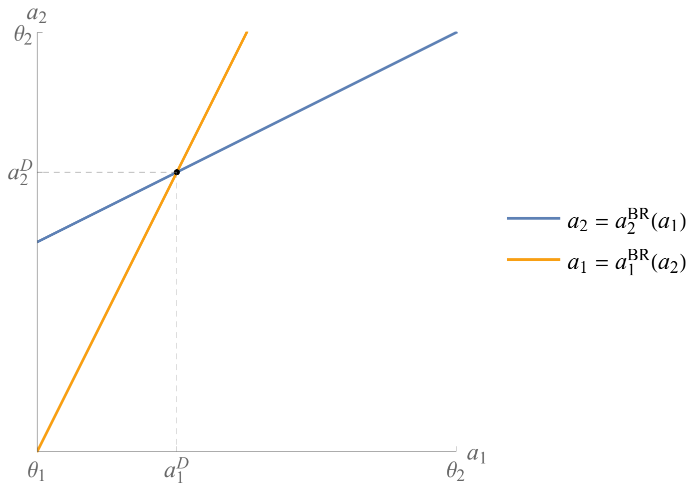

Figure 1.2.1. Disintegration

The result follows from the unique solution to this equation system. □

Notice that members' actions are strategic complements: as a_i increases, member j's best response also increases at the rate of $\kappa_j < 1$.[7] When the coalition is not formed, it guarantees the uniqueness of the equilibrium as illustrated in Figure 1.2.1. Members' ideal actions are at 0 and 1 respectively; their actions are on the two axes. Each upward-tilting line is a member's best-response curve to the other's action. As the slope of each curve is smaller than 1, there is a unique intersection point that represents the equilibrium outcome. As illustrated in the figure, $\theta_1 < a_1^D < a_2^D < \theta_2$: both members' actions are situated between their ideal actions, each between other member's action and his own ideal action.

Define $|\theta_i - a_i|$ as member i's **compromise** to member j. Notice that members make compromises in equilibrium even without the coalition since they concern with coordination.

[7] The feature that members' actions respond to each other at a positive rate smaller than 1 is preserved in more general settings where the losses are additive and convex, and is the reason why key features and results of this model are preserved in such settings.

Notice also that because members' actions are strategic complements, their compromises are *strategic substitutes*: when $\theta_1 < a_1 < a_2 < \theta_2$, a marginal change in a_i that increases $|\theta_i - a_i|$ must decrease $|\theta_j - a_j^{BR}(a_i)|$ at the rate of $\kappa_j < 1$. For instance, a marginal increase in a_1 from a_1^D results in an increase in $a_2^{BR}(a_1)$ from a_2^D. As shown later, this feature is preserved when the coalition is formed.

We denote player i's payoff from the disintegration outcome as $U_i^D \equiv u_i(a_1^D, a_2^D, 0)$. Members form a coalition and centralize decision-making in order to improve upon this payoff. We study how they do so in the next section.

1.2.2. Coordinated integration

This section analyzes **coordinated integration**, the equilibrium outcome that follows the coalition's formation. We show that because members do not follow every recommendation, the coordinator optimally recommends the one that induces actions closest to each other without pushing members out of the coalition; consequently, the synergies accrue to members in equilibrium.

The first observation is that the coordinator's optimal recommendation, which we denote as (a_1^C, a_2^C), must be followed by both members in equilibrium, which echoes Barnard (1938, p. 167)'s observation that "orders will not be issued that cannot or will not be obeyed." This is formalized in the following result.

Lemma 1.2. *If members form the coalition, the coordinator must optimally make a recommendation (a_1^C, a_2^C) that is followed by both members in equilibrium, and $|a_1^C - a_2^C| \leq |a_1^D - a_2^D|$.*

Proof. Consider the subgame in which the coordinator recommends (a_1^D, a_2^D). If an equilibrium exists in which some member does not follow it, then $s = 0$; proof of Lemma 1.1 then implies that the equilibrium actions must still be (a_1^D, a_2^D), a contradiction. Therefore $(a_1^D, a_2^D, 1)$ is the unique (pure-strategy) equilibrium outcome of this subgame.

Consequently, the coordinator's optimal recommendation must do weakly better, and therefore must lead to $s = 1$: both members follow the recommendation in equilibrium. The optimal recommendation (a_1^C, a_2^C) being chosen over (a_1^D, a_2^D) implies that $u_C(a_1^C, a_2^C, 1) \geq u_C(a_1^D, a_2^D, 1)$. Therefore $(a_1^C - a_2^C)^2 \leq (a_1^D - a_2^D)^2$, and hence the result. □

In other words, once the coalition forms, the coordinator can—and therefore will—make a recommendation that induces (weakly) better coordinated actions. Because the coordinator always induces member to stay in the coalition in equilibrium, the magnitude of her synergy $\omega_C \geq 0$ does not matter for her choice of recommendation.

We now characterize the set of *acceptable* recommendations—the ones that can induce an equilibrium in which both members stay in the coalition. In such an equilibrium, each member must be recommended an action sufficiently close to his best response to the other member's recommended action. Intuitively, this ensures that the member's synergy, which he receives if he follows the recommendation, is sufficient to compensate for his loss from *not* best responding to the other member.

Lemma 1.3. *In the subgame where some arbitrary action pair (a_1^R, a_2^R) is recommended, there exists an equilibrium in which both members follow the recommendation if and only* *if for all* $i \in \{1,2\}$ *and* $j \neq i$,

$$|a_i^R - a_i^{BR}(a_j^R)| \leq \sqrt{\omega_i} \tag{1.2.2}$$

PROOF. If member j follows the recommendation, it is optimal for member i to also follow the recommendation if and only if

$$u_i(a_i^R, a_j^R, 1) \geq u_i(a_i^{BR}(a_j^R), a_j^R, 0) \tag{1.2.3}$$

The right-hand side is member i's best deviation, i.e. what he could attain by best-responding to a_j^R, the action that member j would take. If neither member's best deviation is profitable, an equilibrium exists in which both follow the recommendation; the inverse is also true because if this condition does not hold for some member i, he has a profitable deviation.[8]

Notice that condition (1.2.3) is equivalent to Condition (1.2.2). Because of the quadratic loss function, $u_i(a_i^{BR}(a_j^R), a_j^R, 0) - u_i(a_i^R, a_j^R, 0) = [a_i^R - a_i^{BR}(a_j^R)]^2$. Therefore,

$$\begin{aligned} & u_i(a_i^R, a_j^R, 1) \geq u_i(a_i^{BR}(a_j^R), a_j^R, 0) \\ \Leftrightarrow\ & u_i(a_i^{BR}(a_j^R), a_j^R, 0) - u_i(a_i^R, a_j^R, 0) \leq \omega_i \\ \Leftrightarrow\ & |a_i^R - a_i^{BR}(a_j^R)| \leq \sqrt{\omega_i} \end{aligned}$$

□

[8] Member i does not have a best deviation if $a_i^R = a_i^{BR}(a_j^R)$; it is a limiting case and following the recommendation is still the optimal strategy given that $\omega_i \geq 0$. Therefore Condition (1.2.3) still holds in this case.

Condition (1.2.3) formalizes the idea of no profitable deviation: member i's payoff from staying and following the recommendation (which includes the synergy ω_i) needs to be no less than his payoff from leaving and best responding to member j's recommended action. We call the latter payoff member i's outside value given a_j^R and denote it as $U_i^O(a_j^R) \equiv u_i(a_i^{BR}(a_j^R), a_j^R, 0)$.

The equivalence between Condition (1.2.3) and Condition (1.2.2) implies that the latter (when satisfied for both members) fully characterizes the set of acceptable recommendations. Following such a recommendation, the equilibrium in which both members stay in the coalition is selected. Any other recommendation, on the other hand, admits disintegration as its unique equilibrium outcome. This is illustrated in Figure 1.2.2. The recommended actions to each member are on the axes; the dashed lines represent Condition (1.2.2) for both members, all as either horizontal or vertical shifts of the best-response curves (omitted in this figure). The enclosed shaded area represents the set of recommendations that members follow in equilibrium whereas all other recommendations break up the coalition.

Consequently, $\delta_i \equiv \sqrt{\omega_i}$ becomes a measure of the coordinator's ability to distort member i's action from $a_i^{BR}(a_j)$, and we refer to it as the coordinator's **power** over member i. Once the coalition forms, the coordinator can induce member 1 take any action a_1^R within distance δ_1 from $a_1^{BR}(a_2^R)$, provided that she also induces member 2 take the action a_2^R within distance δ_2 from $a_2^{BR}(a_1^R)$.

We say that the coordinator **exerts power** over a member when she distorts his action. Formally, let $e_i \in \mathbb{R}$ be the power exerted over member i, indicating both the magnitude

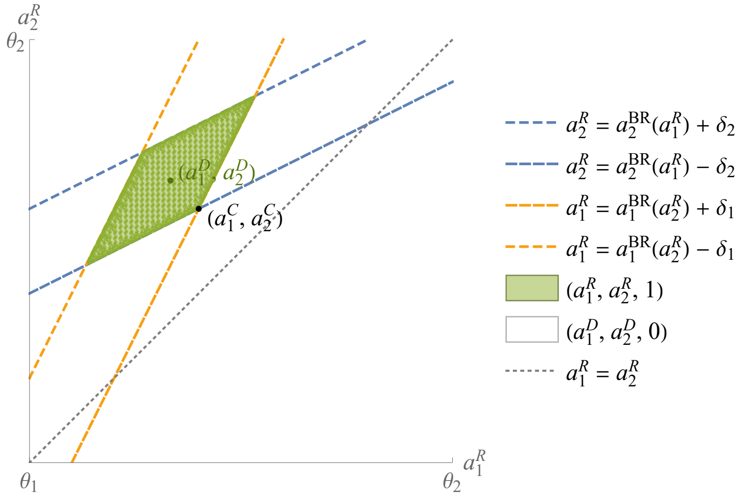

Figure 1.2.2. Centralization

and direction of how the coordinator distorts member i's action from $a_i^{BR}(a_j)$. Lemma 1.3 implies that once the coalition forms, the coordinator has a power of $\mathcal{E} \equiv [-\delta_1, \delta_1] \times [-\delta_2, \delta_2]$ over the members. We can therefore express any *acceptable* recommendation equivalently as $(e_1, e_2) \in \mathcal{E}$, i.e., as the power exerted over both members.

The coordinator's program of optimal recommendation can thus be formulated in the following way.

Program.

$$\max_{(e_1,e_2)\in\mathcal{E}} \tilde{u}_C(e_1, e_2) \equiv u_C(\tilde{a}_1, \tilde{a}_2, 1)$$

$$\textit{where} \begin{cases} \tilde{a}_1 = a_1^{BR}(\tilde{a}_2) + e_1 \\ \tilde{a}_2 = a_2^{BR}(\tilde{a}_1) + e_2 \end{cases}$$

The tilde in variables such as $\tilde{a}_1$ indicates that they take the exertion of power (e_1, e_2) as arguments. Because it is meaningful to talk about exerting power only if a recommendation is followed in equilibrium, $s = 1$ is always assumed and omitted in such an expression.

A unique solution for this program is ensured by the sufficient disagreement assumption, which guarantees that $a_1^R < a_2^R$ for all acceptable recommendations (see Appendix C.1.1), and therefore rules out the existence of a continuum of recommendations each achieving perfect coordination (i.e., $a_1 = a_2$), all optimal for the coordinator. In Figure 1.2.2, this assumption restricts the shaded area to be in the northwest of the dotted line of perfect coordination.

We now characterize (e_1^C, e_2^C) the coordinator's optimal recommendation. It requires the coordinator to exert power as much as it can over both members to move their actions towards each other.

Lemma 1.4. *$(e_1^C, e_2^C) = (\delta_1, -\delta_2)$ and $\theta_1 < a_1^C < a_2^C < \theta_2$.*

Proof. The two equations in Program 1.2.2 define the induced actions $(\tilde{a}_1, \tilde{a}_2)$ as functions of (e_1, e_2):

$$\begin{cases} \tilde{a}_1(e_1, e_2) = a_1^D + \frac{e_1 + \kappa_1 e_2}{1 - \kappa_1 \kappa_2} \\ \tilde{a}_2(e_1, e_2) = a_2^D + \frac{e_2 + \kappa_2 e_1}{1 - \kappa_1 \kappa_2} \end{cases} \tag{1.2.4}$$

which implies that both $\tilde{a}_1$ and $\tilde{a}_2$ increase in (e_1, e_2).

Suppose $e_1^C < \delta_1$. Because

$$\frac{\partial(\tilde{a}_2 - \tilde{a}_1)}{\partial e_1} = -\frac{1-\kappa_2}{1-\kappa_1\kappa_2} < 0$$

and $\tilde{a}_1 < \tilde{a}_2$ for all acceptable recommendation $(\tilde{a}_1, \tilde{a}_2)$, increasing e_1^C would strictly reduce $|\tilde{a}_2 - \tilde{a}_1|$. Therefore $e_1^C < \delta_1$ cannot be optimal. A similar argument can be made for $e_2^C > -\delta_2$.

The sufficient disagreement assumption then guarantees $\theta_1 < a_1^C < a_2^C < \theta_2$. It first guarantees that $a_1^R < a_2^R$ for all acceptable recommendations, and hence $a_1^C < a_2^C$. Now suppose that $a_1^C = \tilde{a}_1(\delta_1, -\delta_2) \leq \theta_1$, which implies that $\tilde{a}_1(0, -\delta_2) < \theta_1$, i.e., $a_1^{BR}(\tilde{a}_2(0, -\delta_2)) < \theta_1$. This implies that $\tilde{a}_2(0, -\delta_2) < a_1^{BR}(\tilde{a}_2(0, -\delta_2)) < \theta_1$, i.e., $\tilde{a}_2(0, -\delta_2) < \tilde{a}_1(0, -\delta_2)$, which contradicts the sufficient disagreement assumption as $(\tilde{a}_1(0, -\delta_2), \tilde{a}_2(0, -\delta_2))$ is an acceptable recommendation. Therefore $\theta_1 < a_1^C$, and similarly $a_2^C < \theta_2$. Hence the result. □

The coordinator's optimal recommendation is thus a corner solution, and is found at the Southeast corner of the shaded area in Figure 1.2.2. The coordinator fully exerts power because doing so always reduces the distance between the two actions. For instance, the coordinator exerts power to increase member 1's action even though member 2's best response also increases, because the increase in member 2's action is at the rate of κ_2, smaller than 1. Similarly, exerting power to decrease member 2's action also reduces $|\tilde{a}_2 - \tilde{a}_1|$.

1.2.3. Participation decisions

Members' participation decisions at the beginning of the game depend on their preferences over the outcomes: they form the coalition only if they both prefer *coordinated integration* to *disintegration*. In two steps, we show how the balance in the coordinator's power over the members determines their preferences. We then derive the key result of this model: the coalition is formed if and only if the coordinator has sufficiently balanced power over the members.

Because a members receives his respective outside value under both outcomes, his preference over the two outcomes is reduced to a comparison over the other member's action in those outcomes. This observation is formalized in the following lemma.

Lemma 1.5. *For all $i \in \{1, 2\}$ and $j \neq i$, $u_i(a_i^C, a_j^C, 1) \geq U_i^D$ if and only if $\left|a_j^C - \theta_j\right| \geq \left|a_j^D - \theta_j\right|$.*

Proof. Under disintegration, a member always best responds to the other member. Therefore member i's payoff is his outside value given a_j^D, i.e.,

$$U_i^D = u_i(a_i^{BR}(a_j^D), a_j^D, 0) \equiv U_i^O(a_j^D).$$

Under coordinated integration, because $|a_i^C - a_i^{BR}(a_j^C)| \equiv |d_i^C| = \delta_i \equiv \sqrt{\omega_i}$ from Lemma 1.4, the proof of Lemma 1.3 can be reversed to show that

$$
\begin{aligned}
&|a_i^C - a_i^{BR}(a_j^C)| = \sqrt{\omega_i} \\
\Leftrightarrow\ & u_i(a_i^{BR}(a_j^C), a_j^C, 0) - u_i(a_i^C, a_j^C, 0) = \omega_i \\
\Leftrightarrow\ & u_i(a_i^C, a_j^C, 1) = u_i(a_i^{BR}(a_j^C), a_j^C, 0) \equiv U_i^O(a_j^C)
\end{aligned}
$$

Therefore $u_i(a_i^C, a_j^C, 1) \geq U_i^D$ is equivalently $U_i^O(a_j^C) \geq U_i^O(a_j^D)$. Since $U_i^O(a_j) = -(1-\kappa_i^2)(a_j - \theta_i)^2$, member i's outside value decreases in the distance between a_j and θ_i. The result is then implied by $\theta_1 < a_j^D < \theta_2$ and $\theta_1 < a_j^C < \theta_2$ for all $j \in \{1,2\}$ as previously derived. □

Lemma 1.5 highlights an important feature of the model: a member's eventual benefit from the coalition comes not from the synergy that he receives, but from an increase in the other member's compromise. Under coordinated integration, it is optimal for the coordinator to distort a member's action as much as it can; the synergy that he receives is thus completely offset by the distortion in his action. Consequently, the member's outside value increases if and only if the other member makes more compromise, reducing the trade-off that he faces when best responding to the other member's action.

The next lemma shows that more compromise is secured from a member if and only if the coordinator has sufficient power over him compared to over the other member.

Lemma 1.6. *For all $i \in \{1,2\}$ and $j \neq i$, $\left|a_j^C - \theta_j\right| \geq \left|a_j^D - \theta_j\right|$ if and only if $\delta_j \geq \kappa_j \delta_i$.*

Proof. Plug $(e_1, e_2) = (\delta_1, -\delta_2)$ in Equations 1.2.4 to get

$$\begin{cases} a_1^C = a_1^D + \frac{\delta_1 - \kappa_1 \delta_2}{1 - \kappa_1 \kappa_2} \\ a_2^C = a_2^D - \frac{\delta_2 - \kappa_2 \delta_1}{1 - \kappa_1 \kappa_2} \end{cases}$$

The result is then implied by $\theta_1 < a_j^D < \theta_2$ and $\theta_1 < a_j^C < \theta_2$ for all $j \in \{1, 2\}$ as previously derived. □

Recall that compromises are strategic complements. Therefore the more power is exerted over a member to secure more compromise from him (δ_i), the less compromise can be secured from the other member ($-\kappa_j \delta_i$). Consequently, forming a coalition helps secure more compromise from member j if and only if the coordinator's power over him dominates her power over member i, i.e., $\delta_j \geq \kappa_j \delta_i$.[9]

Applying this condition to both members, we derive the key result of the model: members are willing to form the coalition if and only if the coordinator has sufficiently balanced power over them both. In other words, the coalition is formed if and only if δ_1 and δ_2 are sufficiently close in magnitude. Otherwise, as the coordinator has excessively more power over one member—whom we call the **weak** member—than over the other—whom we call the **strong** member—she optimally secures less compromise from the strong member once the coalition is formed, and as a result the weak member refuses to let the coalition form in the first place. The need for balanced power is formalized in the following result.[10]

[9] A related result is that $u_i(a_i^C, a_j^C, 1)$ decreases in ω_i (and increases in ω_j). In other words, if a member derives higher synergy from the coalition, he benefits less *a posteriori*. Higher synergy implies more power over him, and thus less compromise from the other member. Yet a member does not wish for extremely low synergy either as the coalition cannot be formed in this case.

[10] Staying apart is always an equilibrium of the game irrespective of members' preferences because a member's participation decision becomes non-pivotal if the other member refuses to form the coalition.

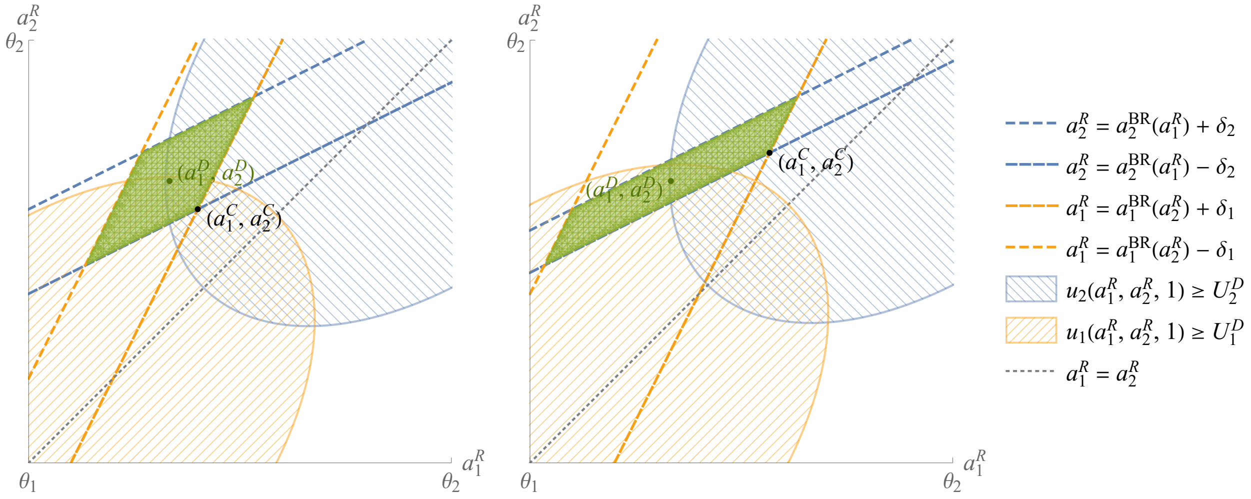

Figure 1.2.3. Balance of power and coalition formation

Proposition 1.1. *The coalition is formed and the coordinated integration outcome prevails if and only if*

$$\kappa_2 \leq \frac{\delta_2}{\delta_1} \leq \kappa_1^{-1}$$

otherwise, the disintegration outcome prevails.

Proof. This result is obtained by applying Lemmas 1.5 and 1.6 over both members and requiring $u_i(a_i^C, a_j^C, 1) \geq U_i^D$ for all $i \in \{1, 2\}$ and $j \neq i$. □

The left panel of Figure 1.2.3 illustrates the case of balanced power: a_1^C is higher than a_1^D and a_2^C is lower than a_2^D, and consequently (a_1^C, a_2^C) falls within the intersection of both striped regions, each representing the recommendations guaranteeing a member at least his disintegration payoff. Otherwise, with sufficient imbalance of power, say $\kappa_2\delta_1 > \delta_2$, the coordinator's optimal recommendation cannot secure more compromise from member 2 than under disintegration. This case is illustrated in the right panel: $a_2^C > a_2^D$ implies

The equilibrium refinement selects coalition formation whenever both member are willing to do so. The same refinement also results from selecting each member's weakly dominant strategy.

that (a_1^C, a_2^C) must make member 1 worse off than under disintegration, so he refuses to join the coalition.[11]

Discussion. This series of results highlight the change in each other's actions that members eventually face—because of the coordination carried out once they form the coalition—as driving their participation decisions. This change in action is not reversed if coordination fails because (a) members take actions simultaneously and cannot best respond to unilateral deviation, and (b) members do not choose actions again if they break up the coalition.[12]

In reality, to what extent a member can reverse his action when coordination fails depends on the context. For example, at the March on Washington, many civil rights leaders delivered their speeches before John Lewis delivered his; their actions were therefore sunk if John Lewis decided at the last moment to take his speech elsewhere. Even when actions are not sunk, reputational concerns and mere unpreparedness can make people stick to their plans when others deviate. All these factors make the change in each other's actions a salient consideration when members decide whether to join. This model captures this consideration in a straightforward way to understand how it shapes the condition for the coalition to form.[13]

[11] Notice that whether a member is the only one making more compromise depends crucially on the preference of the other member. This observation may explain the difficulty for coalition members to distinguish between "selling out a principle" and "making smaller compromises to win larger ones" (Rustin, 2003, p. 126).

[12] As later shown in Section 1.3.3, if members can choose actions again after breaking up the coalition, they optimally take (a_1^D, a_2^D). As a result, the coordinator's optimal recommendation must guarantee each member his disintegration payoff, and members cannot be made worse off joining the coalition.

[13] A related question is whether the coordinator prefers to recommend members to take actions sequentially if she can specify the order of actions. The answer is no for two reasons. First, the coordinator has (weakly) smaller power over the members by inducing members to act sequentially than inducing them to act simultaneously. The coordinator has the same power over the second-mover, but she has smaller power over the first-mover because now when he deviates, he is "forward-looking" and must take

The results do *not* depend on the additivity of the synergy with respect to the issue payoff. For instance, the synergy can be alternatively conceptualized as a multiplier, possibly asymmetric across the members. A member's payoff is then the issue payoff (the quadratic losses plus a large positive constant to make the payoff positive) times the multiplier. Then the same intuition applies: the larger the multiplier, the more the member's action can be distorted by the coordinator without making him leave the coalition. Let the multipliers be sufficiently small so that perfect coordination is infeasible. In this case, any member's benefit from this multiplier is offset by the distortion in his action by the coordinator's recommendation. So the net effect of joining the coalition only depends on how the other member's action changes, which in turn depends on the relative ability of the coordinator to distort the two actions, and is therefore determined by the relative magnitudes of the multipliers.

The intuition of the results does *not* depend on the infeasibility of transfers either. Transfers only add to the complexity of the mechanism. For instance, if the coordinator can specify transfer contingent on members' actions when making her recommendation, such

into account the second-mover's re-optimized action. The first-mover therefore benefits more from his deviation than in the simultaneous-move case. Therefore the coordinator can distort the first-mover's action less. Second, the there is less coordination when members move sequentially. This is due to the first-mover's forward-looking decision of taking an action more biased (compared to his simultaneous-move action) towards his own ideal action to force the second-mover to take a similarly biased action. As the second-mover's reaction is less than 1-to-1, such a bias results in larger distance between the two actions.

transfers become part of the synergies that members receive. How the coordinator optimally allocates her "budget" across the members obviously depends on how much synergies members receive besides those transfers and how much they value money relative to the issue. Therefore the effect of such transfers is ambiguous.[14, 15]

1.2.4. Factors affecting power balance

This model can be applied to better understand how various factors contribute to coalition formation or divide.[16] They affect power balance of a coalition through two channels, which become apparent if we reverse an implicit normalization in members' payoff functions. Assume that $\omega_i = \dfrac{\beta_i}{\rho_i}$ and rewrite member i's payoff as

$$u_i = s \cdot \beta_i - \rho_i[\kappa_i(a_1 - a_2)^2 + (1 - \kappa_i) \sum_{j \in \{1,2\}} (a_j - \theta_i)^2]$$

where $\beta_i > 0$ is the actual synergy that member i receives and $\rho_i > 0$ parametrizes the importance of the issue relative to the synergy.

The imbalance of power could firstly be a result of imbalance of synergies (β's). The member who derives larger synergy from the coalition has more to lose, and is consequently more willing to make a compromise.

[14] In the special case where members' synergies all come from the coordinator's transfers, the member who cares much more about money than the other member does will be the one who stays out of the coalition. It is cheaper to induce deviation in his action than in the other member's action, so coordination optimally allocates more power over him, moving both actions closer to the other member's ideal action.

[15] If *before* members make participation decisions the coordinator can *commit* to transfers conditional on their participation, then she optimally uses such transfers to balance her power over members. With sufficiently large budget, the coordinator has no problem inducing coalition formation.

[16] Notice that this model can be applied to understand the split of a formed coalition as an equilibrium phenomenon as well if the coalition is interpreted as being formed at the beginning of the game and $x_i = 0$ is interpreted as member i's decision to leave the formed coalition.

The large synergy that a weak member derives may be the result of large benefits that other members bring to the table thanks to their size or abundant resource. This observation is consistent with "power differentials" highlighted in the social movement literature.[17] For instance, organizations in the developing world rely on support from their coalition partners from the developed world much more than the other way around, and as a result, centralized decision-making would "reinforce power imbalances among organizational participants" (Bandy and Smith, 2005, p. 11) and could result in the departure of the members who "are systematically ignored, have their priorities devalued, or are marginalized" (Wood, 2005, p. 99).[18]

Similar observations have been made with SSOs as well.[19] The interests of smaller firms and of consumers are found to be inadequately protected by SSOs (Baron et al., 2019, pp. 168-171), and recent conflicts surrounding the change of patent policy in IEEE-SA (which affects the development of Wi-Fi) and the publication of a DRM standard by W3C (which affects how videos are streamed to a web browser) give credence to such claims (Cohen, 2021; Doctorow, 2017). Both cases involve less powerful members who complain for being systemically ignored by the SSOs and threaten to withdraw participation as a result.[20]

[17] See Brooks (2005) for a case of anti-child labor campaign resulting in arrangements favoring the interest of the US labor rather than the livelihoods of Bangeladeshi families. For an intranational example, see Staggenborg (1986) for tensions among constituent organizations within a coalition of pro-choice movement in the US.

[18] Such dynamic is also present in movements at local levels, manifesting as "a fear of 'take-over' of the coalition by one or more powerful organizations or factions" (Staggenborg, 1986, p. 384). In the realm of intergovernmental cooperation, the marginalization of small or poor countries and interest groups is prevalent; see (Stiglitz, 2006, Ch. 10) for a discussion.

[19] This model applies because of the flexibility allowed in standards: the recommendation (a_1^R, a_2^R) can be interpreted as delimiting the range of actions that a standard allows.

[20] An actual resignation by a member organization happened in the latter case.

Large synergy may also be the result of a lack of good external opportunity of a member, for whom leaving the coalition becomes very costly. A product manager has more power over a team who needs a good project, and is more likely to get the cold shoulder from a team that has other good projects to work on.

The imbalance of power may also be caused by a mismatch of members' priorities (ρ's). The less important a member perceives the issue to be, the more power the coordinator has over him, and the more willing he is to make a compromise. In contrast, a member who perceives the issue as his priority would be less willing to make a compromise.

This mismatch of members' priorities helps explain why political coalitions and social movements tend to be “hijacked” by their *radical* factions: as the redicals envisage more “comprehensive versions of change” (Zald and McCarthy, 2017), they concern with issues that others see as ancillary. Consequently, the coordinator secures more compromise from other factions, resulting in her recommendation better aligned with the radicals' ideal action. Similarly, a product manager has more power over a team when compromises on this project do not seriously affect its performance elsewhere; compromises are hard to secure if the team's overall performance is at stake.

Identities, principles or beliefs may create similar mismatch of priorities if they are not shared by all coalition partners. The absence of a “grand coalition” at the March on Washington is a case in point. The divide between the March leaders and Malcolm X can be explained by their mismatched priorities over many issues. On one hand, Malcolm X's strong resistance to the (white) establishment did not resonate with the March leaders, who were confident in shaping the March by themselves (Lewis and D'Orso, 1999, p. 205). They did not reject endorsement by the Kennedy administration and by white participants despite being accused by Malcolm X as “nuzzling up to the white man” (X and Haley, 1965, Ch. 15). On the other hand, having a civil and orderly protest was very important to the March leaders, who intended to draw large crowds and to protect the pending civil right legislation (Branch, 1990, p. 871); this concern was not shared by Malcolm X who dismissed the lack of confrontation as a “farce.” To form a “grand coalition” while tackling power imbalance on those issues, therefore, may have been too difficult.[21]

1.3. Remedies

In this section, we enrich the baseline model to study decision-making protocols that can allow the coalition to form despite an imbalance of power. Formally, we keep members' decisions and the coalition's synergies unchanged, but we allow the coordinator to *commit* to a different protocol for making coordination decision before members decide whether to join. We focus on three types of remedies: (a) **uncoordinated integration**, which bars centralized coordination, (b) **delegated coordination**, which changes the decision-maker's preference, and (c) **member approval**, which subjects the recommendation to additional constraints.

Just forming the coalition—as *uncoordinated integration* always does in equilibrium—cannot fully recover the efficiency loss due to disintegration: there is still scope for improving coordination to everyone's benefit. On the other hand, we show that *delegated coordination* and *member approval* do not always induce the coalition to form, but when

[21] Similar divide was also observed in the protest movements in Hong Kong (largely prior to 2019). It had been difficult for the movement coordinator to work with the militant factions (Cai, 2017). Cases were recorded where the militant protesters ignored the coordinator's plea for restraint, including one incidence where they attempted to remove the coordinator from power (Ng, 2014).

they do, they improve coordination in a Pareto efficient way. In particular, we show that the coalition can adopt consensus decision-making—a specific case of member approval—to augment *any* remedy that coordinates inefficiently; doing so achieves efficiency gain that benefits everyone.

The conditions under which these two remedies induce the coalition to form share common features, because the two remedies need to induce the same set of recommendations on the Pareto frontier in terms of members' payoffs. The trade-off between members' payoffs along this frontier implies that these remedies must balance the their payoffs to induce *both* members to join. Specifically, they must sufficiently protect the weak member's interest to induce his participation, but not excessively so to avoid driving away the other member.

For analyses in this section, we assume that $\delta_2 < \kappa_2 \delta_1$, i.e., member 1 is *weak* and would refuse to join the coalition in the baseline setting. All the subsequent results are formulated accordingly.

1.3.1. Uncoordinated integration

If coordination causes members to stay apart, then the coordinator can stop herself from doing so. The European Union, for example, chooses not to harmonize policy in certain sensitive areas (such as education and culture). We call this practice *uncoordinated integration,* as it bars centralized coordination on an issue and instead allows each member to take whichever action he prefers.

Formally, we change the baseline model by eliminating the recommendation and letting $s = x_1 \cdot x_2$. In other words, members receive their synergies once the coalition is formed no matter what actions they take. We interpret this as the coordinator committing at the beginning of the game to making no recommendation on the issue of disagreement.[22] In equilibrium, members take (a_1^D, a_2^D) but are better off than under disintegration because of the sauvaged synergies. They thus always agree to form the coalition. The coordinator also has incentive to implement uncoordinated integration as she is also better off than under disintegration.

Uncoordinated integration explains the "big-tent" coalitions that have emerged in the recent social movements. In the 2019-2020 protest movement in Hong Kong, for example, uncoordinated integration helped overcome the divide on the use of violence that plagued previous protest movements. A popular slogan of "兄弟爬山，各自努力", which literally translates as "two brothers climb a mountain, each making his own effort" (Kuhn, 2019), was widely accepted and allowed the militant faction to join hands with nonviolent protesters on the street. This commitment to uncoordinated integration followed from the virtual mobilization and spontaneous (and sometimes anonymous) participation by activists—features that greatly inconvenience centralized coordination, and are also shared by other movements across the world (Kidd and McIntosh, 2016; Tufekci, 2017). The commitment to uncoordinated integration often followed from the virtual mobilization and spontaneous (and sometimes anonymous) participation by activists—features

[22] Partial uncoordinated integration—the coordinator making recommendation on some member i's action but let member j take whichever action—always prevents coalition from forming. The situation is equivalent to the coordinator having no power over member j but some power over i, resulting in the maximal imbalance. Despite moving first, the coordinator is not a Stackelberg leader and enjoys no advantage: a member's ability to leave the coalition prevents the coordinator from committing him *not* to react to the other member's action.

that greatly inconvenience centralized coordination and are shared by movements across the world (Kidd and McIntosh, 2016; Tufekci, 2017).

Uncoordinated integration can also be interpreted as adjusting the coalition's scope of coordination when it can coordinate on multiple issues. A case in point is the Lisbon Treaty which distinguishes between policy areas ("competences") where the European Union can and cannot legislate. In social movement coalitions, sometimes a "hybrid organizational form" (Polletta, 2013) is in place to help keep certain issues outside centralized coordination so that "multiple strategies [can be] adopted where [members have] differences" (Gottfried and Weiss, 1994; Arnold, 1995, p. 36).

1.3.2. Delegated coordination

Although uncoordinated integration allows a coalition to form, it prevents any improvement in coordination. In contrast, the other two remedies that we study—*delegated coordination* and *member approval*—both coordinate members efficiently, as they always induce a recommendation that is Pareto efficient given that members can still leave to act freely.[23]

Such efficient remedies must share common features because they need to induce the same set of recommendations along members' Pareto frontier. Therefore, we first study these *constrained efficient* recommendations. We show that, compared to $(e_1^C, e_2^C) = (\delta_1, -\delta_2)$, these recommendations limit how power is exerted over the weak member. They leave some slack so that the weak member's payoff is increased neither too little nor too much.

[23] A natural concern is that an efficient remedy should induce a distribution of recommendations. The concavity of the players' utility functions precludes this possibility.

This feature reflects the need to balance both members' payoffs in order to induce their participation and shapes the two efficient remedies that we later analyze.

1.3.2.1. Constrained efficient recommendations. Formally, we define the set of constrained efficient recommendations—denoted as $\mathcal{E}^*$ and expressed in terms of the power exerted over the members—as satisfying the following conditions:

Definition (Constrained efficient recommendation). $(e_1, e_2) \in \mathcal{E}^*$ if

(Participation constraint) $\tilde{u}_i(e_1, e_2) \geq U_i^D$, for all $i \in \{1, 2\}$,

(Acceptance constraint) $(e_1, e_2) \in \mathcal{E}$, and

(Pareto efficiency) $\nexists (e_1', e_2') \in \mathcal{E}$ such that $\tilde{u}_i(e_1', e_2') \geqslant \tilde{u}_i(e_1, e_2)$ for all $i \in \{1, 2\}$ with $\tilde{u}_i(e_1', e_2') > \tilde{u}_i(e_1, e_2)$ for some i.[24]

The following lemma fully characterizes $\mathcal{E}^*$.

Lemma 1.7. $\mathcal{E}^* = \{(e_1, -\delta_2) | e_1 \in [\kappa_1 \delta_2, \bar{e}_1]\}$ *where* $\bar{e}_1 \in (\kappa_1 \delta_2, \delta_1)$ *is uniquely determined by* $\tilde{u}_1(\bar{e}_1, -\delta_2) = U_1^D$.

Proof. See Appendix C.1.3. □

First notice that the coordinator has strict incentive to induce any such recommendation rather than disintegration given the directions of the power exerted over both members. Power is fully exerted over member 2 (the strong member) to lower his action, and $e_1 \in [\kappa_1 \delta_2, \bar{e}_1]$ specifies the range of power that still needs to be exerted over member 1 (the weak member).

[24] We derive the same set of recommendations if this condition is specified with a weaker concept of Pareto efficiency and with regard to all three players' payoffs, i.e., $\nexists (e_1', e_2') \in \mathcal{E}$ such that $\tilde{u}_i(e_1', e_2') > \tilde{u}_i(e_1, e_2)$ for all $i \in \{1, 2, L\}$. This implies an even stronger desirability of $\mathcal{E}^*$.

To understand why there is only slack in e_1, note that a *strong* Pareto improvement can always be found if there is slack in the power exerted over both members, i.e., if $e_1 < \delta_1$ and $e_2 > -\delta_2$. In this case, exerting more power over both members can better coordinate their actions without reducing anyone's total ideological consistency.[25] Consequently, as there can be slack in the power exerted over at most one member, it has to be the weak member (i.e., $e_1 < \delta_1$), who needs to be made better off than in the baseline model. More specifically, the upper bound $\bar{e}_1$ on the power exerted over him guarantees that he receives at least his disintegration payoff, and is thus induced to join the coalition.

The lower bound $\kappa_1\delta_2$, which caps the weak member's payoff, reflects the need to also guarantee the strong member sufficient payoff. Notice that we can reformulate the Pareto efficiency condition as

$$\text{(Pareto efficiency)} \quad \forall (e_1', e_2') \in \mathcal{E}, \text{ if } \tilde{u}_i(e_1', e_2') > \tilde{u}_i(e_1, e_2) \text{ then}$$
$$\tilde{u}_j(e_1', e_2') < \tilde{u}_j(e_1, e_2) \text{ where } i, j \in \{1, 2\} \text{ and } j \neq i.$$

In other words, along the Pareto frontier, if some recommendation makes the weak member too much better off, then it must make the strong member too much worse off. Therefore to induce the strong member to join the coalition, any recommendation in $\mathcal{E}^*$ must cap the weak member's payoff by restricting $e_1 \geq \kappa_1\delta_2$.

Figure 1.3.1 illustrates the range of constrained efficient recommendations, shown as the thick line segment along the boundary of acceptable recommendations and within the intersection of the striped regions. The two ends represent the two bounds $(\kappa_1\delta_2, -\delta_2)$ and $(\bar{e}_1, -\delta_2)$ respectively.

[25] The Pareto improvement that we use here is *very* strong. It improves the payoff of anyone who concerns with coordination and ideological consistency—no matter where that player's ideal action locates in $\mathbb{R}$—and such an improvement is always strict. See Lemma C.1 in Appendix C.1.3.

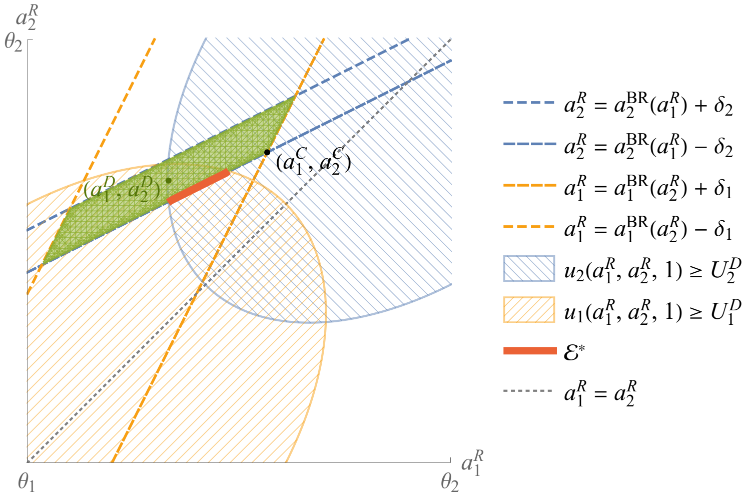

Figure 1.3.1. Constrained efficient recommendations

1.3.2.2. Delegated coordination. The previous analysis implies that any efficient remedy must protect the weak member without over-protecting him. We now show how this feature shapes *delegated coordination.*

A coalition often delegates decisions to different decision-making bodies; one way to do so is to set up groups and committees to coordinate on many issues. SSOs, for example, often have a formal process of mobilizing members to form a working group for every new standard they wish to develop. We call this practice *delegated coordination,* and study how the preference of the delegate needs to be shaped to induce the coalition to form.

We show that the coalition can form—while coordinating efficiently—if a balance of members' interests is found in the delegate's preference: the weak member's interest must be sufficiently but not excessively represented. And as a result, simply delegating to the weak member himself never works.

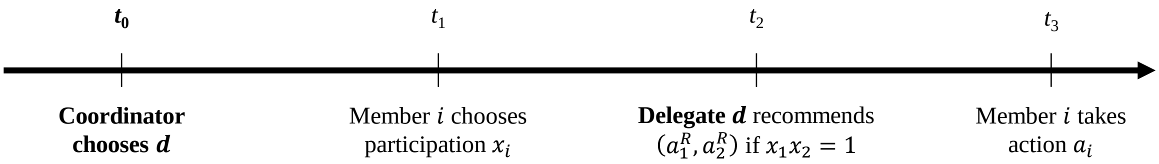
t_0
t_1
t_2
t_3
Coordinator chooses d
Member i chooses participation x_i
Delegate d recommends (a_1^R, a_2^R) if $x_1x_2 = 1$
Member i takes action a_i

Figure 1.3.2. Delegated coordination timeline

Formally, we enrich the baseline model by letting the coordinator publicly choose a delegate (d, she) at the beginning of the game. Subsequently, if both members agree to form the coalition, this delegate—rather than the coordinator herself—makes the recommendation. The timing of the new game is as shown in Figure 1.3.2.

We assume that any delegate concerns with coordination and ideological consistency just as any other player; her preference is thus fully characterized by (θ_d, κ_d):[26]

$$u_d(a_1, a_2, s) = s - [\kappa_d(a_1 - a_2)^2 + (1 - \kappa_d) \sum_{i \in \{1,2\}} (a_i - \theta_d)^2] \tag{1.3.1}$$

My first observation concerns *efficiency*: if a delegate induces the coalition to form, then her equilibrium recommendation must be found in $\mathcal{E}^*$. We show this result by contrdction, using the same *strong* Pareto improvement to find another recommendation in $\mathcal{E}^*$ that yields everyone (including the delegate) a strictly higher payoff if her recommendation is not constrained efficient.

Naturally, the coordinator optimally chooses a delegate among those who makes a constrained optimal recommendation if such a delegate exists. Such a delegate must concern with the ideological consistency between members' actions and her own ideal action ($\kappa_d < 1$), and the more she concerns, the less her ideal action should be biased towards

[26] As discussed in Section 1.1, the magnitude of her synergy is irrelevant as long as it is positive.

the weak member. In the case that member 1 is the weak member and $\theta_1 < \theta_2$, such a delegate's θ_d decreases in κ_d in the *strong set order*.[27]

Motivated by the phenomena of working groups and committees, we focus on a restricted domain of delegates. Let a delegate represent an arbitrary convex combination of the other three players' preferences. The coordinator's choice of delegate thus becomes a choice of $(\gamma_1, \gamma_2, \gamma_C)$ where $\gamma_1 + \gamma_2 + \gamma_C = 1$ and $\gamma_i \geq 0$ is the weight of member i or of the coordinator in the delegate's preference. We interpret these weights as resulting from the coordinator's choice of the composition of a committee, which consequently shapes the committee's preference. In particular, γ_C can be understood as the weight of "third-party/neutral experts" in the committee.[28] This preference is still fully characterized by the pair of parameters (θ_d, κ_d), derived as

$$\begin{cases} \theta_d = \frac{\sum_{i\in\{1,2,C\}} \gamma_i(1-\kappa_i)\theta_i}{\sum_{i\in\{1,2,C\}} \gamma_i(1-\kappa_i)} \\ \kappa_d = \sum_{i\in\{1,2,C\}} \gamma_i \kappa_i \end{cases}$$

The weights that induce the coalition to form reflects the common feature previously derived: the coordinator must put sufficient but not excessive weight on the weak member's preference, as formalized in the following proposition.

[27] If the domain of delegates is unrestricted, i.e., θ_d and κ_d can take arbitrary value in $\mathbb{R} \times [0,1]$, then the set of delegates whose recommendation falls within $\mathcal{E}^*$ is always nonempty and fully characterized by a negative relationship between the two parameters (or a positive relationship if member 2 is weak instead). It maps each θ_d to an interval on $[0,1]$ and each κ_d to an interval on $\mathbb{R}$, and both mappings are decreasing in the *strong set order* (Milgrom and Shannon, 1994; Topkis, 1998).

[28] Quantitative results on the weights $(\gamma_1, \gamma_2, \gamma_C)$ are not discussed in this paper for two reasons. (1) The feasibility of any such weights is constrained by the granularity with which weights can be adjusted. (2) The interpretation of such weights is sensitive to the assumptions made to formally derive the aggregation of players' preferences. Both are outside the scope of this paper. For a theory of aggregating preferences in a setting similar to a committee, see the literature on probabilistic voting, e.g., Lindbeck and Weibull (1987).

Proposition 1.2. *There always exist some weights* $(\gamma_1, \gamma_2, \gamma_C)$ *that induce the coalition to form, and such weights must satisfy* $0 < \gamma_1 < 1$.

Proof. See Appendix C.1.4. □

This result has two implications. Firstly, the weak member must be (sufficiently) represented in the committee. He is indispensable because the neutral coordinator's recommendation (as in the baseline model) does not guarantee him sufficient payoff, and the additional representation of the strong member does not help either. If the committee makes a recommendation different from (a_1^C, a_2^C) for the benefit of member 2 (the strong member), then the recommendation must reduce coordination to increase both actions. Therefore, member 1 (the weak member) can only be made worse off.

Secondly, simply delegating to the weak member cannot allow the coalition to form: the weak member's own recommendation would benefit himself too much to yield the strong member sufficient payoff.[29]

This result explains the difficulties that SSOs face with their working groups. When mobilizing members to form working groups, their composition is easily affected by practical matters such as manpower and expertise. Yet if weak members' continuing involvement is important (e.g., because of legitimacy), SSOs need to ensure that their interest be sufficiently represented. Ignoring this factor risks driving them out, as demonstrated in the recent controversies where the lack of representation in working groups has been a major point of criticism. In contrast, some other SSOs have made requirements on representation in their working groups (Baron et al., 2019, pp. 119-121).

[29] This result still holds even if the weak member cannot commit to penalizing himself when he deviates from his own recommendation.

We can further interpret the committee in my model as any decision-making body in a coalition.[30] In this sense, this result resonates with the general trend of social movement coalitions prioritizing weak members in decision-making. Researchers have identified "cause-affirmation process" as one such practice: the claims of less powerful activists are prioritized during coalitional decision-making so that they need not fear of being "co-opted into something else" (Beamish and Luebbers, 2009). They have also shown that coalitions are shown to succeed when the "smaller, less resource-rich ally" is reassured that its role in the coalition "entail[s] influence over the larger organization's salient choices" (Pullum, 2018, p. 237).

1.3.3. Member approval

A coalition sometimes disciplines its decision-making by subjecting it to the members' approval. It may resort to an official vote or to a tacit acceptance, as implemented by the "written procedure" and the "silence procedure" that the Council of the European Union uses (General Secretariat of the Council, 2009, Article 12); it may involve no gathering and no vote at all but a collection of stakeholders' endorsements, as in a familiar Japanese managerial practice termed *ringisei* or 稟議制 (Vogel, 1975) that are widely used for projects initiated from the bottom of an organization's hierarchy. Regardless of the form it takes, such a practice requires members to express approval (or disapproval) of a recommendation before its adoption. We call such a practice *member approval*, and study

[30] It can be interpreted as the coordinator herself if we let the members choose her preference/composition at the beginning of the game; how they do so is nonetheless outside the scope of this paper.

how to optimally use the necessity of gaining approval to discipline the coordinator's recommendation.

How the recommendation is disciplined depends on two factors: whose approval is needed for the recommendation to be adopted, and what happens if the recommendation is rejected. We show that it is optimal to always require the weak member's approval for adopting a recommendation; we also show that doing so induces the coalition to form—and to coordinate efficiently—if and only if rejecting the recommendation guarantees the weak member a sufficient but not excessive payoff.

We apply this result to show that *consensus decision-making*—requiring approval from both members for adopting the recommendation—constitutes a Pareto improvement when it is used to augment any inefficient remedy that the coalition can use. We argue that this feature justifies the prevalent use of consensus in many coalitional environments.

Formally, we enrich the baseline model with an adoption procedure. The timing of the new game is as shown in Figure 1.3.3. The coordinator publicly chooses an adoption rule $r(m_1, m_2) : \{0,1\}^2 \to \{0,1\}$ at the beginning of the game, specifying how members' messages later in the game affect the adoption of the coordinator's recommendation. If the coalition forms, the coordinator publicly makes its preliminary recommendation $(a_1^P, a_2^P) \in \mathbb{R}^2$ and then members simultaneously each send a public message $m_i \in \{0,1\}$.[31] If $r(m_1, m_2) = 1$, the preliminary recommendation is adopted, i.e., $(a_1^R, a_2^R) = (a_1^P, a_2^P)$. If $r(m_1, m_2) = 0$, the preliminary recommendation is rejected and replaced by another recommendation $(a_1^\emptyset, a_2^\emptyset)$ drawn from a distribution of

[31] Members cannot use their messages to bargain with the coordinator as they cannot commit to sending contingent messages based on the coordinator's preliminary recommendation.

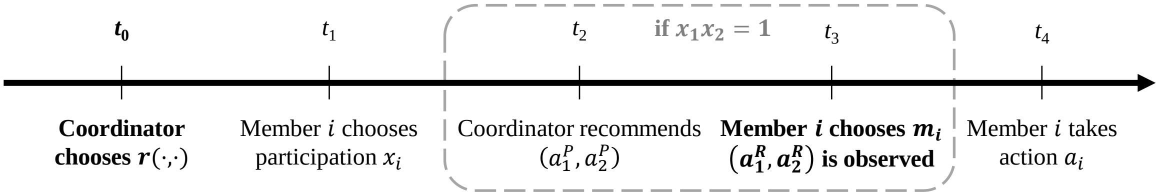
t_0
t_1
t_2
if $x_1x_2 = 1$
t_3
t_4
Coordinator chooses $r(\cdot,\cdot)$
Member i chooses participation x_i
Coordinator recommends (a_1^P, a_2^P)
Member i chooses m_i
(a_1^R, a_2^R) is observed
Member i takes action a_i

Figure 1.3.3. Member approval timeline

fallback recommendations $\alpha^{\emptyset} \in \Delta(\mathbb{R}^2)$ (which we assume to be exogenous for now), i.e., $(a_1^R, a_2^R) = (a_1^{\emptyset}, a_2^{\emptyset})$.[32] Subsequently, members take their actions with the common knowledge of (a_1^R, a_2^R).

Notice that, in equilibrium, members follow (a_1^R, a_2^R) if and only if it is an acceptable recommendation irrespective of how it is generated.

In order to interpret $m_i = 1$ as member i giving his approval to (a_1^P, a_2^P), we restrict the domain of adoption rules. The message $m_i = 1$ cannot change the result from adoption to rejection, as formalized in the following condition.[33]

Condition (Monotonicity)**.** For any $j \neq i$ and $m_j \in \{0, 1\}$, $r(m_i = 1, m_j) = 1$ if $r(m_i = 0, m_j) = 1$.

Six adoption rules satisfy this condition. Two are degenerate; they respectively adopt or reject (a_1^P, a_2^P) regardless of members' messages. The other four rules all satisfy $r(1, 1) = 1$, and we can say that member i is given a *veto* against the coordinator's recommendation

[32] It is possible that $\alpha^{\emptyset}$ is degenerate so that the fallback recommendation is deterministic, i.e., $\alpha^{\emptyset}(a_1^{\emptyset}, a_2^{\emptyset}) = 1$.

[33] Given that we can relabel members' messages and show equivalence between approval rules, this condition only eliminates those rules that approve or reject (a_1^P, a_2^P) depending on whether members send matching or mismatched messages. If such a rule allows the coalition to form, then under this rule the coordinator recommends (a_1^C, a_2^C) in equilibrium and it is approved half the time. Later we show that such randomization must constitutes an inefficient remedy, and can be improved upon once $\alpha^{\emptyset}$ is endogenized.

if $r(m_i = 0, m_j = 1) = 0$. Consequently, these four rules respectively give both members, or either member, or no member a veto.

Under any of the six rules, multiple equilibria exist. As with members' participation decisions, we select the equilibrium using members' weakly dominant strategies: a member gives approval if and only if $r = 1$ yields him no less payoff than $r = 0$. This strategy is formalized in $\tilde{m}_i$, member i's message function whose value depends on $(e_1^P, e_2^P) \in \mathcal{E}$ the preliminary recommendation and $U_i^\emptyset \in \mathbb{R}$ his (expected) payoff from the fallback recommendations:[34]

$$\tilde{m}_i(e_1^P, e_2^P, U_i^\emptyset) = \begin{cases} 1 & \tilde{u}_i(e_1^P, e_2^P) \geq U_i^\emptyset \\ 0 & \tilde{u}_i(e_1^P, e_2^P) < U_i^\emptyset \end{cases}$$

The coordinator's recommendation decision is simplified by this refinement. Once the adoption rule r is chosen and the coalition formed, the coordinator's optimal recommendation is always found among the acceptable ones that get adopted under r.[35] Her program thus becomes

Program.

$$\max_{(e_1^P, e_2^P) \in \mathcal{E}} \tilde{u}_C(e_1^P, e_2^P)$$

$$\textit{subject to } r(\tilde{m}_1(e_1^P, e_2^P, U_1^\emptyset), \tilde{m}_2(e_1^P, e_2^P, U_2^\emptyset)) = 1$$

This simplification follows from two observations. (1) It is not optimal to fall back on $\alpha^\emptyset$, as the coordinator can always construct an acceptable recommendation $(\bar{a}_1, \bar{a}_2)$ to

[34] The message function only needs to be defined over acceptable recommendations since the coordinator cannot optimally make an unacceptable recommendation.

[35] In the case that the preliminary recommendation is automatically rejected, the set of acceptable recommendations that gets adopted is empty; the coordinator's choice has no consequence and any recommendation is optimal.

recommend to members.[36] It always gets adopted under r as it makes every player better off than falling back on $\alpha^{\emptyset}$, strictly so if $\alpha^{\emptyset}$ is non-degenerate, or degerate but not constrained efficient. (2) It is not optimal either to make an unacceptable recommendation that gets adopted and then induces disintegration, as the coordinator can always recommend (a_1^D, a_2^D), which also gets adopted but makes every player better off.

We further simplify the analysis by focusing on the four non-degerate rules. It is without loss of generality to do so because the coordinator always finds her optimal adoption rule among them. The automatic adoption of (a_1^P, a_2^P) prevents the coalition from forming; the automatic rejection is weakly worse than giving both members a veto, as the coordinator can at least recommend $(\bar{a}_1, \bar{a}_2)$ under the latter rule, which gets adopted and yields the coordinator weakly more than her fallback payoff.

This simplification implies that *member approval* must be an efficient remedy if it induces the coalition to form. If a non-degenerate rule induces a recommendation that allows the coalition to form, it must be a constrained efficient recommendation; otherwise we can again use the same *strong* Pareto improvement to find a recommendation in $\mathcal{E}^*$ that yields the coordinator a strictly higher payoff while still securing necessary approval from the members.

The common feature of an efficient remedy thus must be reflected in the condition for *member approval* to induce the coalition to form. Specifically, the coordinator should always give the weak member a veto, and consequently, his veto provides the appropriate discipline to the coordinator's recommendation if and only if the fallback recommendations

[36] See Lemma C.6 in Appendix C.1.5 for how $(\bar{a}_1, \bar{a}_2)$ is constructed.

guarantee him a sufficiently high but not excessive payoff. We formalize this result in the following proposition.

Proposition 1.3. *It is optimal for the coordinator to give the weak member a veto; doing so induces the coalition to form if and only if* $U_1^{\emptyset} \in [U_1^D, \tilde{u}_1(\kappa_1\delta_2, -\delta_2)]$.

Proof. See Appendix C.1.5. □

It is optimal to give the weak member a veto because doing so is necessary for the coalition to form. Relative to the baseline model, each member's veto adds a constraint to the coordinator's program.[37] If a constraint is binding, it changes the coordinator's recommendation to that member's benefit. Since the weak member needs to be made better off for the coalition to form, his veto is necessary.

Whether to also give the strong member a veto, on the other hand, does not matter.[38] If the existing constraint does not allow the coalition to form, then the additional constraint—which can only make the strong member better off—does not either. If the existing constraint is already binding and induces a constrained efficient recommendation, then adding the new constraint cannot induce a different recommendation. Otherwise, this new recommendation must yield the strong member a strictly higher payoff without reducing the weak member's payoff, contradicting the Pareto efficiency of the original recommendation.

[37] Under the rule that gives no veto, the coordinator only needs to secure either member's approval; she need not change her recommendation from the baseline setting, since (a_1^C, a_2^C) is Pareto efficient and cannot be dominated by $\alpha^{\emptyset}$; consequently, at least one member gives approval.

[38] Recall that the coordinator can always recommend $(\bar{a}_1, \bar{a}_2)$, so her program always has a solution.

For the weak member's veto to induce the coalition to form, the fallback recommendations must guarantee him a sufficiently high but not excessive payoff. As the binding constraint implies, the coordinator's equilibrium recommendation yields the weak member exactly his fallback payoff to gain his approval. Therefore his fallback payoff must be neither too much nor too little so that the coordinator's recommendation induces sufficient payoffs for *both* members.

1.3.3.1. Consensus. We now further enrich the model by endogenizing $\alpha^{\emptyset}$. Let the rejection of (e_1^P, e_2^P) lead to an arbitrary subgame that ends with the members simultaneously taking an action while facing the same penalty of losing synergies. Notice that the (mixed) equilibrium outcome of this subgame can always be equivalently induced by some $\alpha^{\emptyset}$ in the previous model.[39]

As a special case, let this subgame be an arbitrary remedy that induces the coalition to form, and let the adoption rule be one of **consensus**, i.e., giving both members a veto. We call this game the *augmented* remedy. It is straightforward to show that the weak member's payoff from the original remedy must be neither too much nor too little as required in Proposition 1.3; consequently, the augmented remedy must also induce the coalition to form.

[39] The (mixed) equilibrium outcome of this subgame is a distribution $\Gamma \in \Delta(\mathbb{R}^2 \times \{0,1\})$; in its support, any outcome with $s = 1$ must be associated with an *acceptable* recommendation, and any outcome with $s = 0$ must be associated with (a_1^D, a_2^D). Consequently, Γ can be induced by some $\alpha^{\emptyset}$ in the previous model by constructing $\alpha^{\emptyset}$ in the following way: pick an arbitrary unacceptable recommendation $(\underline{a}_1, \underline{a}_2)$, and let

$$\begin{cases} \alpha^{\emptyset}(a_1, a_2) = \Gamma(a_1, a_2, 1) & \text{for all acceptable } (a_1, a_2) \\ \alpha^{\emptyset}(\underline{a}_1, \underline{a}_2) = \int_{\text{supp}(\Gamma)} \Gamma(a_1, a_2, 0) \mathrm{d}(a_1, a_2, s) & \end{cases}$$

Moreover, if the original remedy is inefficient, then the augmented remedy induces an outcome that Pareto improves on the original outcome.[40] Notice that the original remedy is *off the equilibrium path* and only serves as a disciplinary device; on the equilibrium path, the coordinator makes a constrained efficient recommendation with better coordination that benefits everyone (strictly so for the strong member and the coordinator herself).

This observation suggests that, irrespective of what inefficient remedy a coalition has, the coalition could improve it by adopting consensus decision-making. For instance, if the coordinator can implement *uncoordinated integration* on an issue where power imbalance would otherwise divide the coalition, she can do even better by committing to taking her recommendation to a vote and dropping the issue all together if her recommendation lacks unanimous support.

This insight given an explanation for why consensus decision-making is widely observed in various coalitional environments. In social movements, consensus decision-making is contrasted with bureaucratized centralization and majoritarian voting and praised for ensuring that "decisions [...] be approvable to all participants" (Della Porta, 2009; Polletta, 2013). Consensus is also called for in other types of coalitions that concern with participation and legitimacy. In the United States, for example, a SSO can develop national standards only if it shows "evidence of consensus" in its decision-making process, e.g., a voting result indicating that no single interest group has objection (American National

[40] Because the coordinator's optimal recommendation must gain both members' approval, they receive at least their fallback payoffs, i.e., equilibrium payoffs from the inefficient remedy. The previous discussion on the degenerate rule of automatic rejection also implies that the coordinator receives strictly more than his fallback payoff.

Standards Institute, 2021).[41] The Council of the European Union also requires unanimity for decisions in many policy areas (European Union, 2016).

The obvious drawbacks of consensus are the extra costs needed to build consensus (Obach, 1999, pp. 63-64) and the consequent delay in decision-making (Polletta, 2013). However, as the discipline is provided by the weak member's veto, an approach to reduce those costs is to identify weak members and only seek their approval. This resonates with "respect for minorities" as one of the core values associated with consensus decision-making (Della Porta, 2009).

A different type of cost comes from the action of veto itself. Members may need to justify their veto,[42] or to fend off other players who try to persuade them. This cost decreases the weak member's fallback payoff (off the equilibrium path), and is translated into a weaker discipline on the coordinator's decision-making. Coordination may actually improve if the coalition can still form with such increased cost.[43]

1.4. N-member coalition

In this section we generalize the model to $N \geq 3$ members, and we show that the key insight from the two-member model continues to hold for the formation of the N-member coalition. We show that members can be partitioned into two groups by the average of their ideal actions, and the coalition is formed if and only if the "aggregate power"

[41] This example follows from the benchmark requirement for the case of safety-related standards.

[42] For SSOs that develop national standards in the US, their members are required to "include specific wording or actions that would resolve the objection" for their negative votes (American National Standards Institute, 2021).

[43] This cost also suggests that *disintegration* is not a robust fallback subgame even though it should implement the coordinator's ex ante optimal recommendation. Any small cost of the weak member to trigger disintegration decreases his fallback payoff below the necessary level to discipline the coordinator, and consequently the coalition cannot form.

over both groups is in balance.[44] This generalization also reveals a new channel through which members change power balance: their ideal actions affect how they are partitioned into those two groups. We also show that all the remedies identified for a two-member coalition generalize to the N-member coalition.

1.4.1. An N-member generalization

Generalizing the model to N members complicates the analysis in least two aspects. First, there are $\sum_{k=2}^{N}\binom{N}{k}$ possible coalitions, each creating for its members a distinct set of synergies. The alternatives for a member to joining a coalition include joining a different one. Even after a coalition is formed, the coordinator may induce *ex post* exit by certain members if it increases her power over the others so she can coordinate better. We assume that synergies exist only for the N-member coalition. In this case, the only alternative to forming the coalition is disintegration, and the coordinator optimally makes a recommendation that every member follows.

Second, each member's outside value now depends on actions by $N-1$ other members; the aggregate effect is not necessarily characterized in a simple way. However, as long as each member is concerned with how well aligned his action is with the others on average, the balance of power is still crucial. We study a generalization that isolates this concern and we show that the condition for a coalition to form is fully characterized by the balance of power.

[44] Also important is the question of whether this coalition or some other smaller one would win out given a coalition formation process or whether this coalition satisfies some desirable property; the answer to such a question relies on the result from this section, but a formal treatment is outside the scope of this paper.

Let $\mathcal{N} \equiv \{1, 2, \cdots, N\}$ be the set of potential members of the coaltion. They simultaneously decide whether to form the N-member coalition. If all members agree to form the coalition, then the coodinator publicly makes a recommendation $\mathbf{a}^R \equiv (a_1^R, a_2^R, \cdots, a_N^R)$. Irrespective of whether the coalition is formed or not, members simultaneously each takes an action a_i. The synergies $(\omega_1, \omega_2, \cdots, \omega_N)$ accrue to each member if everyone follows the recommendation.

For tractability, we assume that members have identical concern with coordination relative to ideological consistency, i.e., $\kappa_1 = \kappa_2 = \cdots = \kappa_N = \kappa$. Member i's preference is

$$u_i(\mathbf{a}, s) = s \cdot \omega_i - \kappa(a_i - a_{-i})^2 - (1 - \kappa)[(a_i - \theta_i)^2 + (N - 1)(a_{-i} - \theta_i)^2]$$

where $\mathbf{a} \equiv (a_1, a_2, \cdots, a_N)$, $a_{-i} \equiv \frac{\sum_{j \neq i} a_j}{N-1}$ is the average of other members' actions and $s \equiv \prod_{i \in \mathcal{N}} x_i \cdot \mathbb{1}[a_i = a_i^R]$.

Notice that member i's best-response to $\mathbf{a}_{-i} \equiv (a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_N)$ is a function of a_{-i} only, i.e., $a_i^{BR}(\mathbf{a}_{-i}) \equiv \kappa a_{-i} + (1 - \kappa)\theta_i$. This greatly simplifies the analysis. In particular, as in the two-member model, we can define $\tilde{a}_i(e_1, e_2, \cdots, e_N), \forall i \in \mathcal{N}$ as the unique solution to the following system:

$$\tilde{a}_i = a_i^{BR}(\tilde{\mathbf{a}}_{-i}) + e_i, \forall i \in \mathcal{N}$$

where e_i indicates the direction and magnitude of the power exerted over member i.

The coordinator's preference is specified as

$$u_C(\mathbf{a}, s) = s \cdot \omega_C - \frac{1}{N} \sum_{i \in \mathcal{N}} (a_i - a_{-i})^2 = s \cdot \omega_C - \frac{N^2}{(N-1)^2} \mathrm{Var}(\mathbf{a})$$

The coordinator thus wants to minimize the distance between each member's action and the average action.

1.4.2. Optimal recommendation and coalition formation

As in the two-member model, the coordinator's optimal recommendation must be followed in equilibrium, and therefore it is equivalently expressed as the optimal exertion of power, i.e., $\mathbf{e}^C \equiv (e_1^C, e_2^C, \cdots, e_N^C)$ from the set $\mathcal{E} \equiv \bigtimes_{i \in \mathcal{N}} [-\delta_i, \delta_i]$.

Let $\bar{\theta} \equiv \frac{1}{N} \sum_{i \in \mathcal{N}} \theta_i$ be the average ideal action and $\bar{\delta} \equiv \frac{1}{N} \sum_{i \in \mathcal{N}} \delta_i$ be the average power. The appropriate generalization of the sufficient disagreement assumption is the following condition.[45]

Assumption (Generalized sufficient disagreement)**.**

$$(1-\kappa) \left| \theta_i - \bar{\theta} \right| > \frac{N-2}{N} \delta_i + \bar{\delta}, \forall i \in \mathcal{N}$$

This condition ensures that however power is exerted over the members, $a_i = a_{-i}$ is not feasible for any member i. It then implies that the coordinator optimally exerts all her power over the members and hold them all to their respective outside value. The direction of her exertion of power depends on the member's group affiliation as characterized by the following result.

[45] This condition has $N-1$ degrees of freedom and thus collapses into an single inequality when $N = 2$. It rules out the non-generic case of some $\theta_i = \bar{\theta}$. See Appendix C.1.1 for its derivation.

Lemma 1.8.

$$\begin{cases} e_i^C = \delta_i & \theta_i < \bar{\theta} \\ e_i^C = -\delta_i & \theta_i > \bar{\theta} \end{cases}$$

Proof. See Appendix C.1.2. □

Notice that $\theta_i \lesseqgtr \bar{\theta}$ is equivalent to $a_i^D \lesseqgtr \bar{a}_{-i}^D$, and this result follows from the monotonicity of the coordinator's payoff in the power exerted over each member, i.e., $\frac{\partial \mathrm{Var}(\tilde{\mathbf{a}})}{\partial e_i} < 0$ if $\tilde{a}_i < a_{-i}$ and $\frac{\partial \mathrm{Var}(\tilde{\mathbf{a}})}{\partial e_i} > 0$ if $\tilde{a}_i > a_{-i}$, which implies the optimality of the corner solution.

As each member i is held to his outside value given $\mathbf{a}_{-i}^C$, coordination does not decrease member i's outside value if and only if a_{-i}^C is closer than a_{-i}^D to θ_i. The following result gives the condition for coordination to induce such an outcome.

Proposition 1.4. *The N-member coalition is formed if and only if*

$$-(1-\kappa) \cdot \min\{\delta_j | \theta_j > \bar{\theta}\} \leq \sum_{i:\theta_i < \bar{\theta}} \delta_i - \sum_{j:\theta_j > \bar{\theta}} \delta_j \leq (1-\kappa) \cdot \min\{\delta_i | \theta_i < \bar{\theta}\}$$

Proof. See Appendix C.1.2. □

In other words, members are partitioned by $\bar{\theta}$ into two groups, and the coalition exerts power to move actions towards the average action. Consequently, the coalition can be formed if and only if the "aggregate power" over both groups is sufficiently balanced. In this case, from each member's perspective, the average action of all other members moves (weakly) closer to his own ideal action, improving his outside value.

On the other hand, when the coalition fails to form, the relative magnitude of the aggregate power indicates which group has members refusing to join. For instance, $\sum_{i:\theta_i<\bar{\theta}} \delta_i > \sum_{j:\theta_j>\bar{\theta}} \delta_j$ indicates that $\{i|\theta_i < \bar{\theta}\}$ is the **weak group**, i.e., some member i with $\theta_i < \bar{\theta}$ prevents the coalition from being formed.

This condition also highlights how members' ideal actions affect the balance of power. Unlike in the two-member model where each member is his own group, ideal actions matter when $N \geq 3$ because they determine how members are partitioned into these two groups.

An interesting observation is that extreme ideology (i.e., ideal action) can either make or break a coalition: roughly speaking, the coalition is more likely to form if extreme ideology is held by a weak member. Suppose that the coordinator has power of similar magnitude over each member; because a member's extreme ideology affects $\bar{\theta}$ so much, a large set of "central" members must be grouped with others on the opposite end from him. The coalition's aggregate power over them could be so great that some member in this group is better off staying out. On the other hand, this intuition also implies that when the coalition has excessive power over a member, balance could still be maintained if he has an extreme ideology and constitutes (perhaps with some other members) a small group. Extreme ideology can thus mitigates the disadvantage of a member's relative weak position in a coalition.

1.4.3. Remedies

It is straightforward that *uncoordinated integration*, i.e., committing to making no centralized coordination, always induces the coalition to form. To analyze the efficient remedies, on the other hand, we first derive a partial characterization of $\mathcal{E}^* \subseteq \mathcal{E}$, the set of constrained efficient recommendations in the N-member game; as in the two-member model, this characterization determines the common features of efficient remedies.

We show in Appendix C.1.3 that an strong Pareto improvement can be achieved by raising some e_i where $\tilde{a}_i < \tilde{a}_{-i}$ and lowering some e_j where $\tilde{a}_j > \tilde{a}_{-j}$. This result indicates that any constrained efficient recommendation must exert full power over all members of one group, and it must be the strong group given that some members of the weak group need to be made better off than under disintegration to be willing to participate. As in the two-member model, the strong Pareto improvement result also implies that if *delegated coordination* or *member approval* induces the coalition to form, it must induce some $\mathbf{e}^* \in \mathcal{E}^*$.

The pitfall of over-protection continues to be a prominent feature of efficient remedies, as there is a threshold for power exerted over the weak group in aggregate below which some member from the strong group starts to prefer staying out of the coalition.

The results for *delegated coordination* in the N-member model are highly analogous to those of the two-member model. Sufficient representation from the weak group is necessary, but excessive representation of the weak group makes all the members from the strong group worse off than under disintegration. As a special case, delegating to either group can never allow the coalition to form.

The result for *consensus* is also analogous. We show that consensus decision-making continues to be Pareto efficient, and constitute a Pareto improvement on any inefficient remedy just as in the two-member model. We also show that, generically, vetos of members of the strong group are not all consequential to providing discipline over the coordinator's decision-making.

1.5. Conclusion

People get together to better coordinate themselves; but when there is sufficient disagreement, coordination inevitably involves trading off different interests. This is why *balance* becomes of paramount importance to get people on board. My model highlights a coalition's limited power over its members' actions, and how its imbalance causes divide. To overcome the divide, the coalition needs to discipline how power is exerted over its members, and one way to do so is to enrich the decision-making process to protect its weak member's interest.

We suggest some interesting avenues for future research. The first one concerns repeated decision-making. An enduring coalition coordinates on many issues over time, and the gain from coordination on one issue can serve as synergy on another. This gives rise to the problem of agenda setting: how does a coalition sequence and bundle issues to maximize coordination while sustaining itself? Another one concerns the eliciting and exchange of knowledge. Coalition partners often come from diverse backgrounds, but decisions are not necessarily made with good appreciation of their priorities and strengths. Much benefit of consensus-building and deliberative democracy is about discovering coalition partners' concerns and positions, and about facilitating the emerging of new ideas and solutions. There is value in better understanding those aspects of decision-making and how coalitions should organize accordingly.

CHAPTER 2

Performance Manipulation and Incentive Design

Though widely used to provide incentives, performance measures are often manipulated. Police chiefs push officers to under-report crimes when prevention does not sufficiently reduce the numbers. Executives manufacture earnings when operations alone cannot produce the desired financial result. Yet people do not just manipulate; they make decisions that facilitate manipulation in the future. Uncooperative police officers are removed from their precincts. Mergers and restructuring are planned to create future earnings. Even reputable firms like GE have been accused of building up “cookie jar reserves” so that the executives could reach into them in bad times. Arthur Levitt, then SEC Chairman, discussed how these “gimmicks” work in his “Numbers Game” speech (Levitt, 1998) on earnings management: unrealistic assumptions are used to create large estimates for future expenses incurred because of a restructuring or a merger, or for items such as sales returns, loan losses or warranty costs; these estimates can then be reversed and “miraculously reborn as income” when “future earnings fall short.”

Those applications differ from the existing literature on multitasking and gaming (Kerr, 1975; Milgrom, 1988; Holmstrom and Milgrom, 1991; Baker, 1992) in a crucial way: manipulation are not mere reactions to incentives; they are *set-ups* made in anticipation of future incentives. Such set-ups are a concern whenever performance measures are well established. In the motivating examples above, corporate executives and police chiefs understand what numbers they need to massage and what they need to do in order to manipulate future performance.

In an environment where an agent can set up to manipulate future performance, incentives respond to the chosen set-up. This chapter uses a principal-agent model to capture these strategic considerations. Specifically, a principal (P, "she") designs a contract based on an imperfect performance measure to induce effort from an agent (A, "he") protected by limited liability. The agent can take two types of actions to increase the probability of high performance: one benefits the principal, i.e., "effort"; the other harms the principal, i.e., "manipulation". The key twist is that manipulation takes place *before* the principal designs the contract. This timing makes it a *set-up* because the agent chooses manipulation in anticipation of the future contract.

As in models where the agent chooses both effort and manipulation *after* the principal designs her contract (see Holmstrom and Milgrom, 1991; Baker, 1992; Beyer et al., 2014 for multitasking models, and Dye, 1988; Crocker and Slemrod, 2008; Sun, 2014 for other manipulation models), the agent manipulates to extract more rent from the principal. Specifically, given a fixed contract that rewards high performance, higher manipulation increases the probability of high performance and therefore increases the agent's payoff. As a result, the principal optimally offers a contract less sensitive to performance than it would be absent the agent's ability to manipulate.

When the agent manipulates before the contract is chosen, however, he is deterred from exerting too much manipulation because of its negative impact on future incentives. Too much manipulation makes the principal face a performance measure badly aligned with her objective, i.e., the agent's effort, resulting in low-powered incentives. The agent then extracts less rent. Therefore, the agent does not manipulate too much in equilibrium.

The deterrence effect decreases with the quality of the performance measure, i.e., how sensitive the performance measure is to effort relative to manipulation. The less a performance measure is affected by manipulation, the more the agent can manipulate without inducing low-powered incentives. This observation implies that the principal can be harmed by an improvement of quality in the performance measure. Indeed, if manipulation is costless, then better-quality performance measures always decrease the principal's payoff. This result stands in contrast to the classic results in the multitasking literature (Baker et al., 1994; Fehr and Schmidt, 2004; see Gibbons (1998) for a discussion).

This result calls into question the value of improving performance measures. While increasing the cost of manipulation unambiguously increases the principal's payoff, policies that improve the alignment between the performance measure and the principal's objective can harm the principal. Section 2.2.1 shows that the use of benchmark or discretion improves this alignment and leads to worse outcomes for the principal. In contrast to the "informativeness principle" (Holmstrom, 1979; Shavell, 1979), this result suggests that "better" information can lead to worse outcomes when agents act strategically before the principal designs incentives.

This chapter also studies a two-period dynamic model in which set-ups in period 1 improve the performance in period 2 while worsening the performance in period 1. This model is motivated by the observation that making restructuring provisions lowers the current year's earnings.[1] The principal's optimal incentive in period 1 is steeper than it would be without manipulation. There are two reasons for this result. Firstly, the agent is incentivized to reduce manipulation when the first-period performance is better rewarded. Secondly, as the first-period performance measure has been manipulated, inducing effort is actually less costly for the principal, as it reduces the rent extracted by the agent per unit of effort. The principal therefore wants to induce more effort. These two effects also lead to the "front-loading" of incentives, i.e., the first-period incentive is sharper than the second-period incentive.

The chapter is organized as follows. Section 2.1 analyzes the model of costless manipulation and shows that a performance measure of higher quality induces more manipulation and harms the principal. Section 2.2 shows that policies that increase the cost of set-up is beneficial to the principal whereas policies that improve the quality of performance measure may be counterproductive. Section 2.3 then extends the model to two periods and shows that the principal optimally front-loads incentive in the first period if set-up harms the first-period performance and benefits the second-period performance.

Related Literature

The agent's ability to change measured performance without exerting productive effort is often called "manipulation." Manipulation is modeled in the literature in different ways. This chapter conceptualizes manipulation as the agent improving the contracting variable at the expense of the principal. The direct harm to the principal is the defining feature of

[1] In the example of policing, similar effect on the current-period performance can be justified by the talent loss and sapped morale that accompny the dismissal of honest officers.

manipulation in contrast to (productive) effort. This harm befits the motivating examples of crime statistics manipulation and earnings management, and is in contrast to the security design literature (Koufopoulos et al., 2019; Lauzier, 2021) which usually assumes that an agent (enterpreneur) can manipulate or "window-dress" the contracting variable (return of a project) by borrowing and then paying it back out of his own pocket after his performance is observed and the contract executed. In such a setting, the principal (financier of the project) directly benefits from the agent's manipulation, though the agent's ability to manipulate distorts the incentive design of the principal and indirectly harms her as in the current model.

The current model builds on "multitasking" models that illustrate the difficulty caused by a single-dimensional performance measure capturing activities of multiple dimensions. The seminal contributions in this literature by Holmstrom and Milgrom (1991, 1994) and Baker (1992) highlight that the principal optimally provides lower-powered incentive when her objective and the performance measure become worse aligned. Applied to the setting where the misalignment stems from the agent's manipulation (Beyer et al., 2014), this feature implies that the principal provides lower-powered incentives when compared to how she would design the contract absent the agent's ability to manipulate. Similar insights are also derived in other manipulation models (Dye, 1988; Crocker and Slemrod, 2008; Sun, 2014). A collorary is that, because an increase in the cost of manipulation leads the principal to optimally increase the contract's sensitivity to performance, making her better off. Both features are also reflected in this model. Unlike the existing literature, the agent in the current model acts strategically before the principal's incentive design, so his manipulation is deterred by a rent extraction consideration that trades off the probability of a bonus with the size of that bonus. Consequently, the current model shows that improving the quality of the performance measure can have the perverse effect of causing more manipulation and harming the principal.

The fact that the agent acts before the principal's incentive design connects this chapter to the literature on pre-contracting information design (Gul, 2001; Lau, 2008; Hermalin and Katz, 2009; Condorelli and Szentes, 2020). Closest to this chapter is Garrett et al. (2021), which studies a setting where the agent can choose the cost of his effort in order to manipulate his incentives under limited liability. That paper shows that the agent optimally chooses binary outputs and a particular class of cost functions. The model in this chapter fixes the cost function; the agent instead chooses the informativeness of the performance measure about effort, with lower informativeness (i.e., more manipulation) reducing the principal's payoff.

2.1. Baseline model

In this section, we set up the baseline model of a single period and costless manipulation. We derive the result that, as long as the performance measure remains imperfect, an improvement in its quality strictly harms the principal. The functional forms used in this model are chosen for simplicity of exposition; in Appendix C.2, we prove the results under more general conditions.[2] We also use subscripts to denote partial derivatives.

[2] We assume that high performance is realized with probability $\Pr[\kappa = H] = p(e, \phi) + q(m, \phi)$ where $p_e \geq 0, p_\phi \geq 0, p_{e\phi} > 0, q_m \geq 0, q_\phi \leq 0$ and $q_{m\phi} < 0$ and they are all bounded away from infinity. We also assume that p is multiplicatively separable in (e, ϕ). We say that a performance measure approaches perfect alignment when $q_m \to 0, \forall m \in [0, 1]$ and perfect misalignment when $p_e \to 0, \forall e \in [0, 1]$. The principal's payoff is $U^P = u(e) - h(m) - t$ and the agent's payoff is $U^A = t - c(e)$ where u, h and c are weakly positive and strictly increasing; $c_e(0) = 0$ and $c_{ee} > 0$.

A principal retains an agent who takes two actions, which we refer to as effort and manipulation. The levels of effort and manipulations are $(e, m) \in [0,1]^2$. Both effort and manipulation contribute to a performance measure $\kappa \in \{L, H\}$, whose realization is determined by $\Pr[\kappa = H] = \phi e + (1-\phi)m$ where $\phi \in (0,1)$, the **quality** of the performance measure, parametrizes the sensitivity of high performance to effort and inversely to manipulation. We refer to the performance measure as approaching perfect alignment with the principal's objective if $\phi \to 1$ and perfect misalignment if $\phi \to 0$.

The agent's choices of e and m are observed by the principal, but only the realized performance is contractible. The principal's contract offer to the agent is thus a contingent transfer $t(\kappa) : \{H, L\} \to \mathbb{R}_+$. We assume that the agent's outside value is weakly negative so that his participation constraint is not binding in equilibrium.

At the beginning of the game, the principal chooses from two performance measures: $\ddot{\kappa}$ and $\dot{\kappa}$, with the associated parameter $1 > \ddot{\phi} > \dot{\phi} > 0$. Without loss of generality, we assume that $\dot{\kappa}$ is a garbling of $\ddot{\kappa}$.[3] Therefore $\ddot{\kappa}$ is better aligned than $\dot{\kappa}$ to the principal's objective.

The formal timing of the game is the following (every choice is publicly observed).

(1) P chooses between $\{\ddot{\kappa}, \dot{\kappa}\}$.

(2) A chooses $m \in [0,1]$.

(3) P chooses $t(\kappa) : \{H, L\} \to \mathbb{R}_+$.

(4) A chooses $e \in [0,1]$.

(5) Performance κ is realized and the contract is executed.

[3] In Section 2.2.1 We discuss the possibility of using two performance measures together; it is equivalent to using an improved performance measure of which both measures are a garbling.

The principal's payoff is $U^P = e - \lambda m - t$ and the agent's payoff is $U^A = t - \gamma \frac{e^2}{2}$, where his effort cost is assumed to be quadratic. Throughout the chapter we also assume that (a) all the optimization problems are strictly concave, and (b) the solution (e, m) is always interior, which requires effort cost to be sufficiently high and $\ddot{\phi}$ sufficiently low.[4]

In the following sections, we use backward induction to show the main results of the chapter. We ignore the principal's first-stage problem for now as her optimal choice will follow from the comparative statics result of the agent's second-stage problem. The agent's problem is to find the optimal level of manipulation given that it affects the principal's optimal contract, which in turn affects the level of effort induced from the agent and, by extension, his rent extraction.

In Section 2.1.1, we first show that the principal's optimal contract pays bonus for high performance and zero for low performance, and that the optimal bonus level is lower when the agent's manipulation is higher. The intuition is that higher manipulation makes inducing effort more costly, so the principal lowers the bonus and induces less effort. Consequently, it is not optimal for the agent to manipulate too much. This deterrence effect is absent in models where the agent acts strategically only after the principal designs the contract. The agent's optimal manipulation therefore balances the probability of bonus payment against the size of the bonus.

Section 2.1.2 shows that the agent's optimal manipulation makes the principal induce *the same* level of effort irrespective of the quality of the performance measure. If the performance measure becomes more sensitive to effort, then the agent optimally increases manipulation so that the principal's marginal (and average) cost of inducing effort stays

[4] The necessary and sufficient conditions are $\gamma \geq \frac{1}{3}$ and $\ddot{\phi} \leq \frac{3\gamma}{1+3\gamma}$ under the current specification.

unchanged.[5] Section 2.1.3 then shows that the principal is harmed by an improvement in the quality of the performance measure. Even though such an improvement does not affect equilibrium effort induced from the agent, it decreases how much the principal's optimal bonus responds to the agent's manipulation level. Consequently, the agent manipulates more in equilibrium without affecting the cost of effort inducement and therefore his rent extraction. This decrease in deterrence thus results in more harm to the principal. As a result, the principal optimally chooses the performance measure of a lower quality at the first stage.

2.1.1. Principal's optimal effort inducement

We first derive the effort induced by the optimal contract that the principal offers the agent. As the contract is offered after the principal has observed agent's choice of manipulation level, we keep $m \in [0,1]$ fixed in this section. We refer to the principal's expected payoff under his optimal contract as her *interim* value.

Lemma 2.1. *The principal's optimal effort inducement and interim value both decrease in manipulation m and increase in performance measure alignment ϕ.*

Proof. See Appendix C.2.1.1. □

We show the key intuitions of this result with the quadratic-linear functional form.

[5] This feature is guaranteed by the multiplicative separability between ϕ and e and the additive separability between e and m in the principal's payoff.

As with a canonical limited liability setting, the principal optimally sets $t(\kappa = L) = 0$. With abuse of notation, let $t = t(\kappa = H)$ and we refer to it as the **bonus** for high performance.

The agent's *ex post* program is thus

$$\max_{e \in [0,1]} \; t \cdot \Pr[\kappa = H] - \gamma \frac{e^2}{2}.$$

Recall that the optimal effort is interior; it solves

$$\phi t - \gamma e = 0. \tag{2.1.1}$$

Define $t^P(e) = \frac{\gamma}{\phi} e$ as the bonus that induces effort level e. The principal's optimal contract then solves

$$\max_{e \in [0,1]} e - \lambda m - t^P(e, \phi) \cdot \Pr[\kappa = H] \quad .$$

The resulting optimal effort, $e^P(m, \phi)$, is determined by the first-order condition:

$$1 - t^P \cdot \frac{\partial \Pr[\kappa = H]}{\partial e} - t^P_e(e, \phi) \cdot \Pr[\kappa = H] = 0,$$

or

$$\underbrace{1 - \gamma e^P}_{\text{marginal surplus}} = \underbrace{\frac{\gamma}{\phi}[\phi e^P + (1 - \phi) m]}_{\text{marginal rent}}. \tag{2.1.2}$$

Define $\rho \equiv \frac{\phi}{1-\phi}$ as the marginal ratio of performance measure sensitivities. We can rewrite Equation (2.1.2) as

$$e^P(m,\phi) = \frac{1}{2\gamma} - \frac{m}{2\rho}.$$

Therefore,

(1) $e^P_m < 0$. More manipulation reduces effort by increasing the marginal rent paid for manipulation.

(2) $e^P_\rho > 0$ (or equivalently $e^P_\phi > 0$.) Better aligned performance measure increases effort.

The second result is not *a priori* obvious. It follows from the observation that better alignment reduces the rent paid for manipulation but does not affect the rent paid for effort. Although better alignment reduces the size of the bonus that induces a fixed level of effort ($t^P_{e\phi} \leq 0$), it also increases the probability of paying the bonus. However, these two effects cancel. Better alignment shrinks the marginal bonus needed to induce any effort at the rate of *marginal* sensitivity to effort, while inflating the probability of bonus payment due to effort at the rate of *average* sensitivity to effort. Those two rates coincide and the two effects cancel as can be seen in (2.1.2) for $m = 0$. Consequently, for any level of effort, the rent paid for effort is unaffected by performance measure alignment.[6]

Since $t^P(e^P(m,\phi),\phi)$ is the principal's optimal contract, her *interim* value is $V^P(m,\phi) \equiv e^P(m,\phi) - \lambda m - t^P(e^P(m,\phi),\phi) \cdot \Pr[\kappa = H]$. This expression simplifies to

[6] And by extension, for any level of effort inducement, the (total) motivational rent from effort is unaffected either. One way to see this is through an integration over effort level while noticing that motivational rent from inducing zero effort is always zero. This result is used later in Proposition 2.3. A more general result can be established. See Lemma C.9 and Corollary C.3 in Appendix C.2.1.

$$V^P(m, \phi) = -\lambda m + \frac{1}{4\gamma} - \frac{m}{2\rho} + \frac{\gamma m^2}{4\rho^2}.$$

Therefore,

(1) $V_m^P < 0$. More manipulation reduces the principal's *interim* value. This is straight-forward from the envelope theorem; more manipulation results in more harm to the principal and more rent extraction by the agent.

(2) $V_\rho^P > 0$ (or equivalently $V_\phi^P > 0$.) Better alignment increases the principal's *interim* value. This also follows the envelope theorem, since better alignment affects the marginal rent only through manipulation.

The size of the optimal bonus decreases with manipulation due to reduced effort:

$$\frac{\partial t^P(e^P(m, \phi), \phi)}{\partial m} = t_e^P \cdot e_m^P < 0.$$

Notice that $t^P(e^P(0, \phi), \phi)$ is the optimal bonus when manipulation is infeasible. This observation leads to the following result, later used in the two-period model in Section 2.3:

Proposition 2.1. *The optimal bonus is lower than the bonus without manipulation.*

Proof. $t^P(e^P(0, \phi), \phi) > t^P(e^P(m, \phi), \phi)$ where $m > 0$ follows from $\frac{\partial t^P(e^P(m,\phi),\phi)}{\partial m} = t_e^P \cdot e_m^P < 0$. □

2.1.2. Agent's optimal manipulation

In this section, we study Stage 2 of the game and show that irrespective of alignment ϕ, the agent's optimal manipulation always results in a fixed level of effort.

Proposition 2.2. *Irrespective of ϕ, the agent's optimal manipulation makes the principal induce the same level of effort.*

PROOF. See Appendix C.2.1.2. □

Here is a rough intuition for this independence. Manipulation allows the agent to set the informativeness of the performance measure, and hence the principal's cost of—and his own rent from—inducing some fixed level of effort. Importantly, he can use manipulation to replicate the informativeness of another performance measure. For example, $\Pr[\kappa = H] = \frac{1}{2}e + \frac{1}{2}m$ and $\Pr[\kappa = H] = \frac{2}{3}e + \frac{1}{3}(2m)$ is identically informative of effort, and thus generate the same cost and rent for inducing any level of effort. Moreover, there is a unique optimal cost of effort inducement that maximizes the agent's rent extraction. As a result, there is a unique equilibrium effort associated with such cost, independent of the alignment.

Formally, we take a detour to prove this result in two steps. First, we show that a marginal increase in manipulation has two effects that must cancel at the optimal manipulation level. These two effects help us later understand the case where manipulation is costly. Second, we show that when both effects are expressed in terms of the effort induced by the principal, the solution must be independent of alignment ϕ.

To derive the two effects, notice that the agent's optimal manipulation solves

$$\max_{m\in[0,1]} V^A(m,\phi) \equiv t^P \cdot \Pr[\kappa = H] - \gamma\frac{(e^P)^2}{2},$$

where $e^P \equiv e^P(m,\phi)$ and $t^P \equiv t^P(e^P,\phi)$.[7] Recall that the optimal manipulation is interior; it solves

$$t^P[\phi e^P_m + (1-\phi)] + t^P_e e^P_m[\phi e^P + (1-\phi)m] - \gamma e^P e^P_m = 0. \tag{2.1.3}$$

This equation can be simplified using the first-order condition for effort, (2.1.1). Equation (2.1.3) thus becomes

$$\underbrace{t^P(1-\phi)}_{\text{direct effect}} + \underbrace{t^P_e e^P_m \cdot [\phi e^P + (1-\phi)m]}_{\text{strategic effect}} = 0.$$

Using the first-order conditions of the principal's *interim* program, we further simplify to:

$$\underbrace{t^P(1-\phi)}_{\text{direct effect}} + \underbrace{e^P_m \cdot (1-\gamma e^P)}_{\text{strategic effect}} = 0. \tag{2.1.4}$$

This condition shows that a marginal increase in manipulation increases the probability of a bonus (the direct effect), but decreases effort and hence the size of the bonus (the strategic effect).

We now show that these two effects can be characterized in terms of the effort induced by the principal.

[7] It can be shown that V^A is supermodular in (m,ϕ) so that the agent's manipulation level increases in performance measure alignment. However, we still delve into this condition and show how contracting itself has a deterrence effect on the agent's manipulation decision; this effect gives an intuition for how manipulation changes with performance measure alignment.

Recall that $\rho \equiv \frac{\phi}{1-\phi}$, $t^P = \frac{\gamma}{\phi}e$ and $e_m^P = -\frac{1}{2\rho}$. The first-order condition (2.1.4) is therefore equivalent to

$$\gamma e^P - \frac{(1-\gamma e^P)}{2} = 0. \tag{2.1.5}$$

This condition has a unique solution $e^* = \frac{1}{3\gamma}$ that is independent of ϕ.

As a corollary, we can characterize how the optimal manipulation changes with ϕ.

Corollary 2.1. *The agent's optimal manipulation increases in performance measure alignment ϕ.*

Proof. The previous result implies that $e^P(m^A(\phi), \phi) = e^*$; take its total derivative w.r.t. ϕ, we get

$$m_\phi^A = -\frac{e_\phi^P}{e_m^P} > 0.$$

□

Intuitively, when the performance measure is well-aligned, the agent optimally chooses a high level of manipulation to keep the optimal effort—and hence rent extraction—constant.

2.1.3. Principal prefers worse-aligned performance measures

In this section, we show that the principal's payoff is decreasing in ϕ.

The principal's *ex ante* payoff is given by

$$\bar{V}^P(\phi) \equiv e^P(m^A(\phi), \phi) - \lambda m^A(\phi) - t^P \cdot \Pr[\kappa = H].$$

where $t^P \equiv t^P(e^P(m^A(\phi), \phi), \phi)$. This simplifies to

$$\bar{V}^P = \frac{1 - 3\lambda\rho}{9\gamma}.$$

Recall that $\rho \equiv \frac{\phi}{1-\phi}$; we thus have the following result.

Proposition 2.3. *The principal's ex ante payoff decreases in ϕ.*

PROOF. See Appendix C.2.1.3. □

There are two reasons that better alignment harms the principal. The first reason follows from Proposition 2.2, which says that better alignment *does not* affect equilibrium effort. This implies that alignment *does not* affect the benefit from effort, the cost of effort, or the agent's rent paid for effort. The second reason is that alignment *does not* affect the agent's rent paid for manipulation, which is shown in Lemma C.11 in Appendix C.2.1.3 and we give a brief intuition here.

Recall that the *interim* first-order condition (2.1.2) says that the principal's marginal benefit from effort must equal the marginal bonus payment. Suppose that alignment improves. As the optimal manipulation must not change effort, neither do the marginal benefit and cost of effort change, nor the marginal rent paid for effort. Therefore the optimal manipulation must not change the marginal rent paid for manipulation either. In other words, while better alignment shrinks bonus, the optimal manipulation must inflate the probability of that bonus in a way that offsets the previous effect.[8]

[8] This result follows the multiplicative separability assumption; a more general result can be established. See Lemma C.11 in Appendix C.2.1.

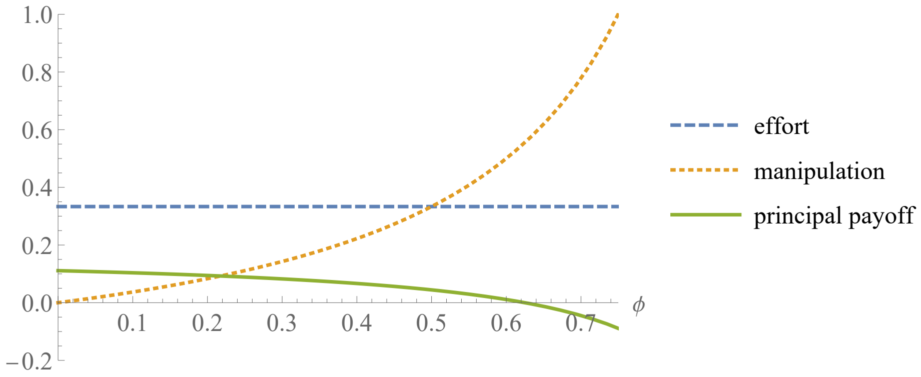

Figure 2.1.1. Change in performance measure alignment

In summary, any change in performance measure alignment would not affect either the surplus from effort or the bonus payment. The change in principal's value only comes from the direct harm caused by manipulation. Since better alignment increases manipulation, the principal's payoff decreases.

Consequently, the principal optimally chooses the *worse* aligned performance measure.

Corollary 2.2. *The principal optimally chooses* $\dot{\kappa}$ *over* $\ddot{\kappa}$*.*

PROOF. It follows from the previous proposition. □

Figure 2.1.1 illustrates these results with $\gamma = 1, \lambda = 0.2$, and ϕ is on the horizontal axis.

The same intuition explains why the agent's *ex ante* payoff is unaffected by the change in the performance measure alignment.

Corollary 2.3. *The agent's ex ante payoff is independent of* ϕ*.*

PROOF. It follows from the previous proposition. □

2.2. Policy intervention

In this section, we discuss two types of policy intervention. The first type concerns an improvement of performance evaluation so that high performance becomes either more likely due to effort or less likely due to manipulation. For example, new technology could enhance the precision of a performance measure, and so could the principal's accumulated experience in identifying manipulated outcome. This can be thought of as an increase in ϕ in our model.

Section 2.2.1 shows that an improvement of performance evaluation can also be interpreted as introducing additional performance signals. That is, the use of additional signals leads to a worse outcome in our model. This result stands in contrast to the "informativeness principle" (Holmstrom, 1979; Shavell, 1979), and it suggests that adding performance measures might have the unintended consequence of encouraging manipulation.

The other type of policy intervention concerns increasing the cost of manipulation, which we can interpret as, for example, requiring the agent justify certain types of decisions. Section 2.2.2 analyzes this situation by enriching the baseline model with a cost of manipulation. It shows that the principal is unambiguously better off when manipulation becomes more costly.

2.2.1. Additional signals

In this section, we enrich the model to give a foundation for $\dot{\kappa}$ and $\ddot{\kappa}$. We interpret the choice of $\ddot{\kappa}$ over $\dot{\kappa}$ as the use of additional signals to improve the quality of performance measure.

Let $\omega \in \{\mathcal{E}, \mathcal{M}\}$ be a state of the world that determines if performance measure κ reflects effort or manipulation:

$$\Pr[\kappa = H | \omega = \mathcal{E}] = e$$

$$\Pr[\kappa = H | \omega = \mathcal{M}] = m.$$

Note that the marginal probability of high performance is $\phi e + (1 - \phi)m$ when $\Pr[\omega = \mathcal{E}] = \phi$. Let $\sigma \in \{E, M\}$ be a binary signal correlated with ω. Define $r_{\sigma|\omega} \equiv \Pr[\sigma|\omega]$ and assume that $0 \leq r_{E|\mathcal{M}} < r_{E|\mathcal{E}} \leq 1$ (which implies $1 \geq r_{M|\mathcal{M}} > r_{M|\mathcal{E}} \geq 0$) so that $\sigma = E$ is indicative of $\omega = \mathcal{E}$ and vice versa.

Instead of choosing between $\dot{\kappa}$ and $\ddot{\kappa}$ at the beginning of the game, the principal now chooses whether σ would be contractible. We show that the problem is equivalent to the baseline model. We first show that, with both signals, the optimal contract pays a bonus only when $\kappa = H$ and $\sigma = E$. We refer to this outcome as high performance under the effort signal.

Lemma 2.2. *When σ is contractible, the principal's optimal contract rewards the agent only when $\kappa = H$ and $\sigma = E$.*

Proof. See Appendix C.2.2.1. □

Intuitively, the optimal contract must minimize wasteful bonus payment by concentrating it on the signals that most strongly indicate effort, which means (a) that no bonus is paid when performance is low and that (b) no bonus is paid when the additional signal indicates a relatively high probability of manipulation causing the high performance.

Now, construct a composite performance measure:

$$\hat{\kappa} = \begin{cases} H & \kappa = H, \sigma = E \\ \\ L & \text{otherwise.} \end{cases}$$

Note that

$$\Pr[\hat{\kappa} = H] = \phi r_{E|\mathcal{E}} e + (1 - \phi) r_{E|\mathcal{M}} m$$

where the coefficients $\phi r_{E|\mathcal{E}}$ and $(1 - \phi) r_{E|\mathcal{M}}$ do not add up to 1. The following lemma shows that the comparative statistics in Proposition 2.3 nevertheless hold.

Lemma 2.3. *Any outcome of a given performance measure that yields high performance with probability $p(e) + q(m)$ can be replicated by another performance measure that yields high performance with probability $\alpha p(e) + \alpha q(m)$ where $\alpha > 0$.*

Proof. See Appendix C.2.2.1. □

In fact, any linear transformation of the probability of high performance realization has no effect on the possible outcomes that the principal and agent can have. This result depends on the additive separability between p and q, since a linear transformation has the same effect on the marginal and the average sensitivities to effort and manipulation.

We can therefore write

$$\Pr[\hat{\kappa} = H] = r_{E|\mathcal{E}}[\phi e + (1 - \phi) \frac{r_{E|\mathcal{M}}}{r_{E|\mathcal{E}}} m]$$

which has the identical outcomes as using a performance measure yielding high performance with probability $\phi e + (1 - \phi) \frac{r_{E|\mathcal{M}}}{r_{E|\mathcal{E}}} m$, an improvement of κ given that $\frac{r_{E|\mathcal{M}}}{r_{E|\mathcal{E}}} < 1$.

Proposition 2.3 then implies that the principal finds it undesirable to introduce this additional signal into the contract.

Proposition 2.4. *The principal optimally chooses not to introduce the additional performance measure.*

Proof. It follows from Proposition 2.3 and Lemma 2.3. □

Applications: discretion and benchmarking. The additional signal σ may come from the principal's own assessment (assumed to be truthful in our discussion) or from an "objective" benchmark. In our motivating example of earnings management, one manipulation trick is to initiate a merger or restructuring to make provisions that can be injected into earnings in future years. In this case, the board might use some observable features of such a merger or restructuring as a noisy signal of manipulation. This additional signal functions as σ in our model, and Proposition 2.4 thus suggests that its use is likely to exacerbate the manipulation problem. The reason is that the additional signal makes the equilibrium contract less sensitive to manipulation, so the executive worries less about the size of the bonus when making large provisions.

Benchmarking, or relative performance evaluation, is another form of additional signal. The conventional wisdom is that incorporating another agent's performance can help tease out correlated noises in performance measures. In our setting, the other agents' performance could contain information regarding how likely our agent's high performance is due to manipulation rather than effort. This is the case, for instance, when an industry experiences headwinds due to macroeconomic hardship (so effort is unlikely to yield hefty profit) but one firm continues to have exceptional earnings. Using relative performance evaluation is therefore an improvement in the performance measure, which can exacerbate manipulation.[9]

2.2.2. Costly manipulation

Manipulation can be costly, and policies or regulations are often designed to increase the cost of manipulation. To understand the effect of those instruments, we enrich the baseline model to allow costly manipulation.

We assume that the manipulation cost takes the same functional form as the effort cost, i.e. $\mu\frac{m^2}{2}$ (where $\mu > 0$ is bounded away from zero or infinity), and we refer to an increase in μ as an increase in manipulation cost. In this case, an increase in μ has an unambiguous positive effect, as formalized in the following proposition.

Proposition 2.5. *The principal's ex ante payoff increases in manipulation cost.*

Proof. See Appendix C.2.2.2. □

For any fixed m, the optimal contract satisfies the same conditions as in the baseline model. The cost of manipulation, therefore, affects outcomes only through the agent's decision at Stage 2. With a cost of manipulation, the agent's optimal manipulation solves

$$\gamma e^P - \frac{(1-\gamma e^P)}{2} - \rho\mu m = 0 \tag{2.2.1}$$

[9] Note that when manipulation is present, the other agent's low performance may suggest that the focal agent's high performance is due to manipulation rather than effort. This is the case, for instance, if ω is common among the two agents (the macroeconomic shock in our hypothetical case) and the other agent's choice is such that his high performance is more likely due to effort. If on the contrary the other agent's high performance is more likely due to manipulation (e.g. the market is plagued with executives manipulating earnings), then the focal agent's high performance combined with the other agent's low performance is a strong signal of effort.

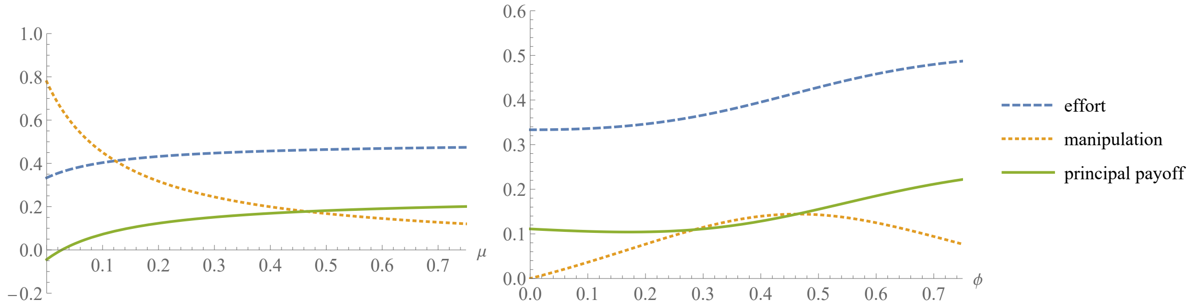

Figure 2.2.1. Costly manipulation

This condition implies that the optimal manipulation decreases with μ: the optimal manipulation is

$$m^A = \frac{1}{3\gamma/\rho + 4\mu\rho}.$$

As the cost increases, manipulation unambiguously decreases, and the principal's payoff unambiguously increases by the envelope theorem.

The left panel of Figure 2.2.1 illustrates this result with $\gamma = 1, \lambda = 0.2, \phi = 0.7$ and μ is on the horizontal axis.[10]

2.3. Two-period model

In many settings, the agent must make a trade-off between executing productive effort and setting up future manipulaiton. An executive might sacrifice this year's earnings to build provisions for the future when setting up a restructuring. A police chief pushes out disobedient officers, possibly losing talent and sapping morale at the same time. In these

[10] On a side note, the ambiguous effect of ϕ can also be inferred in Equation (2.2.1), and the right panel of Figure 2.2.1 illustrates this result with $\gamma = 1, \mu = 1, \lambda = 0.2$ and ϕ is on the horizontal axis. See Appendix C.2.2.2 for a discussion.

examples, setting up manipulation can adversely affect the current performance. The optimal contract should take this mechanism into account to deter manipulation.

To model this situation, we extend the baseline model to two periods. The set-up action in anticipation of the second period is taken concurrently with productive effort and negatively affects performance in the first period. We show that the principal benefits from making first-period incentives steeper than they would be if deterrence were not possible. This effect manifests as a "front-loaded" incentive, namely the incentive in the first period is steeper than in the second period.

This section considers the following game:

(1) Principal chooses a contract for the first period t_1.

(2) Agent chooses m and e_1, both are observed.

(3) The first period performance is realized.

(4) Principal designs a contract t_2.

(5) Agent chooses e_2.

(6) The second period performance is realized.

As in the baseline model, all choices are publicly observed, and only performance is contractible. The principal offers a short-term contract in each period.[11] We assume that the set-up for manipulation directly decreases performance; however, the results would be the same if manipulation instead increased effort costs, as shown in Appendix C.2.3.1.

For illustration we use the linear-quardratic specification

[11] Since manipulation level is non-contractible, any long-term contract made in the first period has no deterrence effect on the manipulation choice of the agent. We assume that this makes the long-term contract less preferred than the short-term contract.

$$\Pr[\kappa_1 = H] = \phi e_1 - \nu(1-\phi)m$$

$$\Pr[\kappa_2 = H] = \phi e_2 + (1-\phi)m$$

$$U^P = e_1 + e_2 - \lambda m - t_1 - t_2$$

$$U^A = t_1 + t_2 - \gamma\frac{e_1^2}{2} - \gamma\frac{e_2^2}{2}$$

where $\nu > 0$ parametrizes the "costliness" of manipulation relative to effort exertion. If $\nu = 0$, the analysis becomes effectively identical to that of the baseline model with an additional first-period effort inducement problem independent of the rest of the analysis.

2.3.1. Optimal manipulation

Notice that the second-period problem (steps 4–6) is identical to what has been analyzed in Sections 2.1.1 and 2.1.1. Our analyses therefore focus on only the first-period problem.

The agent's first-period program is

$$\max_{(e_1,m)\in[0,1]^2} U^A(e_1, m, t, \phi) \equiv \mathbb{E}[t|e_1] - \gamma\frac{e_1^2}{2} + V^A(m, \phi)$$

where with abuse of notation $t \equiv t_1$, V^A is as defined in Section 2.1.2, representing the agent's second-period surplus given manipulation m. The solutions $(\hat{e}_1, \hat{m})$ are determined by a pair of first-order conditions:

$$\begin{cases} \gamma\hat{e}_1 - t\phi = 0 \\ V^A_m(\hat{m}, \phi) - t(1-\phi)\nu = 0 \end{cases}$$

Due to V^A's concavity in m, the agent's optimal manipulation decreases in bonus, i.e. $\hat{m}_t < 0$.

Lemma 2.4. *Manipulation $\hat{m}$ decreases in first-period bonus t.*

Proof. See Appendix C.2.3. □

2.3.2. Overloading & front-loading incentives

In this section, we consider the contracting problem in the first period. Notice that the first-period contract affects the first-period effort as well as the amount of manipulation entering the second period. As manipulation reduces the agent's rent in the first period, the principal can deter manipulation by increasing the first-period bonus. We show that this mechanism causes the principal to front-load incentives.

To simplify notation, we let $t = t_1$, $e = e_1$ and suppress the argument ϕ in all expressions. Recall the agent's first-period program,

$$\max_{(e,m)\in[0,1]^2} U^A(e,m,t) \equiv \mathbb{E}[t|e] + V^A(m) - \gamma\frac{e^2}{2}.$$

Let $(\hat{e}, \hat{m})$ be the unique optimal choice this program.[12] The principal's program is thus

$$\max_{t\geq 0} \hat{U}^P(t) \equiv \hat{e}(t) - \mathbb{E}[t|\hat{e}(t)] + V^P(\hat{m}(t))$$

where V^P is as defined in Section 2.1.1, representing the principal's second-period surplus given manipulation m.

[12] This program is concave under the linear-quardratic specification. For more general specifications, we assume this concavity.

An *auxiliary* program can be formulated by fixing the agent's manipulation.[13] His first-period program is then

$$\max_{e\in[0,1]} U^A(e,m,t) \equiv \mathbb{E}[t|e] + V^A(m) - \gamma\frac{e^2}{2}$$

Let $\tilde{e}(m,t)$ be the unique optimal effort choice in this *auxiliary* program. The principal's program is then

$$\max_{t\geq 0} \tilde{U}^P(m,t) \equiv \tilde{e}(m,t) - \mathbb{E}[t|\tilde{e}(m,t)] + V^P(m)$$

Notice that the agent's effort decisions are identical across the two problems. In other words, $\hat{e} = \tilde{e}$ and $\hat{e}_t = \tilde{e}_t$.[14] An implication follows:

Proposition 2.6. *The principal sets the first-period bonus higher than it would be if the agent's manipulation were fixed at its equilibrium level.*

Proof. See Appendix C.2.3. □

In other words, $\hat{t} > \tilde{t}(\hat{m})$, where $\hat{m}$ is the agent's optimal manipulation choice in the original program, $\hat{t}$ is the optimal contract in the original program and $\tilde{t}$ is her optimal contract in the *auxiliary* program. The principal overloads incentives when contracting with agents who can manipulate in order to both inducing effort and detering manipulation. In contrast, in the *auxiliary* program where the agent's manipulation is fixed, higher

[13] This program is concave under the linear-quardratic specification. For more general specifications, we assume this concavity.

[14] More generally, $\hat{e}_t \geq \tilde{e}_t$ suffices for the following result; this inequality is guaranteed if m is instead assumed to negatively affect performance by increasing effort cost.

bonus does not benefit the principal by deterring manipulation, so the optimal bonus is smaller than in the original program.

Notice that the optimal bonus increases in manipulation in the *auxiliary* program. The intuition is that more manipulation lowers the probability of bonus payment without affecting effort. The principal thus induces more effort by setting a higher bonus when effort inducement is cheaper. This implies that the optimal bonus in the *auxiliary* program with positive manipulation is higher than that with no manipulation. In other words, $\hat{t} > \tilde{t}(\hat{m}) > \tilde{t}(0)$.

Proposition 2.7. *The optimal first-period bonus is higher than the bonus without manipulation.*

Proof. See Appendix C.2.3. □

Recall Proposition 2.1, which shows that the second-period bonus is lower than the bonus without manipulation. Hence the following corollary.

Corollary 2.4. *The first-period bonus is higher than the second-period bonus.*

Proof. It follows from Propositions 2.1 and 2.7. □

2.4. Conclusion

This chapter sheds light on how agents set up opportunities to manipulate future performance in order to game future incentives. It identifies the deterrence effect of the principal's incentive design, which can undermine policies that improve performance evaluation. This framework also shows that frontloading incentives can improve the principal's payoff by deterring an agent from manipulating.

CHAPTER 3

Incentives in Delegation

This chapter focuses on the case where contingent transfers are infeasible and formal authority resides at the top of a hierarchy. Delegation becomes an important organizational tool in such settings. For example, consider a manager who can decide whether to work on a project herself or delegate to a subordinate. If the subordinate is less incentivized to exert effort than the manager is, then the manager faces an "effort vs opportunity" trade-off and must weigh the cost of eliciting insufficient effort from the subordinate versus the opportunity cost of *not* working on another project.

The severity of this trade-off depends on the subordinate's effort. Her effort choice, in turn, depends on the project's return to effort. This return is often referred to as the "impact" of a project. For example, in a corporate hierarchy, high-impact projects are those that has high potential to improve the profit of the firm. Effort devoted to such a project can greatly benefit both the manager and the subordinate in terms of bonus or promotion. Similarly, if the "manager" is an academic researcher and the "subordinate" is his junior collaborator, then the return on effort of a project can be measured by how its outcome could develop a field or produce a good paper.

This chapter studies how information asymmetry exacerbates the "effort vs opportunity" trade-off and makes delegation less likely. It shows that, when the return on effort of a project is the private information of the manager, the manager optimally delegates fewer projects than when the return is commonly known; the delegated projects are also of lower returns.

Formerly, consider a privately informed manager ("she") who decides whether to delegate a project to a subordinate ("he"). For any level of effort exerted by the subordinate, the "effort vs opportunity" trade-off implies that manager optimally delegates projects with returns below a threshold. Like in the market of "lemons", the act of delegating therefore becomes a negative signal about the value value of effort. The bubordinate infers that only low-return projects are delegated and exerts even less effort. The manager responds by only delegating projects of even lower return. This feedback loop results in a delegation threshold that is lower than in the symmetric-information environment.

This model highlights how the information environment affects delegation decisions: a project may appear suitable for delegation to outsiders, yet a manager may find delegation suboptimal at the time of making the delegation decision, as she anticipates low effort because the subordinate would treat such a project as a "lemon". An implication is that the subordinate's belief about the distribution of projects matters. A manager is more likely to delegate a project in an environment where it is commonly known that high-return projects abound, since the threshold of delegation is higher in this case and a subordinate exerts higher effort in equilibrium.

This model also highlights a complementarity between credible communication and delegation. As the manager is harmed by the subordinate's imprecise inference, she benefits from credibly communicating her private information. Cheap-talk does not allow her to communicate any information, as she has the incentive to exaggerate the return to effort. On the other hand, if she can commit to a signaling strategy, e.g., à la Kamenica and

Gentzkow (2011), then she optimally reveal all information because her payoff is convex in the subordinate's posterior belief over the return of delegated projects. Such commitment is not needed if the manager can costlessly produce evidence about the project's return à la Grossman (1981). The case of incomplete evidence à la Dye-Jung-Kwon (Dye, 1985; Jung and Kwon, 1988) is also discussed.

When credible communication is infeasible, the manager can still benefit from the commitment to a delegation strategy, i.e., the set of projects to delegate. We show that the manager optimally delegates projects with medium return. Intuitively, she abandons delegating the very low-return projects to increase the subordinate's effort, and she herself works on the very high-return projects as the subordinate's low effort does not justify delegating such projects.

The structure of the chapter is as follows. Section 3.1 sets up the symmetric-information benchmark and characterizes the principal's delegation strategy in this setting. Section 3.2 sets up the private-information game and shows how the "effort-opportunity" tradeoff is exacerbated by adverse selection. Section 3.3 then discusses how commitment to a signaling strategy and voluntary disclosure can restore the principal's payoff to her symmetric-information benchmark. It also shows that commitment to a delegation strategy can improve outcomes when credible communication is infeasible.

Related Literature

This chapter builds on the economics literature in delegation, which largely focuses on delegated decision-making and assumes that the agent, rather than the principal, has private information (Dessein, 2002; Alonso and Matouschek, 2007; Alonso et al., 2008; Alonso and Matouschek, 2008; Li et al., 2017; Deimen and Szalay, 2019). This information asymmetry makes delegation valuable to the principal, as it allows the agent to use his private information. This chapter focuses on a qualitatively different type of information asymmetry: the principal has private information about the return to effort for a project. This asymmetry is what leads to the “lemons” problem that makes delegation less valuable to the principal.

This chapter is also related to papers that study how the allocation of authority can signal information (Garicano and Santos, 2004; Dessein, 2005). The closest is Dessein (2005), in which a privately informed enterpreneur can relinquish formal control to an investor. Formal control allows the investor to impose an action on the enterpreneur when the preferences of the two parties diverge. In equilibrium, the enterpreneur relinquishes control to signal that their preferences are congruent. In contrast, relinquishing control is a negative signal in our setting since it indicates that the project has low return.

This chapter is related to the literature of adverse selection (Akerlof, 1970) and the literature on how disclosure of evidence can remedy adverse selection problems (Grossman, 1981; Dye, 1985; Jung and Kwon, 1988). Empirical findings in the organizational behavior literature also supports features of this model. The result that managers delegate less when the issue at hand is more important is supported by Leana (1986) and Yukl and Fu (1999). Leana (1986) also reports the positive correlation between delegation and managers’ workload (which accounts for their opportunity cost of not delegating); Yukl and Fu (1999) also reports the positive correlation between delegation and subordinates’ competence and alignment of objective with manager (both captured by γ in the model).

3.1. Symmetric-information benchmark

This section sets up the symmetric-information benchmark of the delegation game and characterizes the equilibrium delegation strategy.

3.1.1. Model set-up

Consider the following principal-agent model. The principal (P, "she") draws a project from a distribution of projects. A project is characterized by its impact α, which equals the principal's marginal return to effort on the project. The principal also has a non-delegatable project of impact β. It is common knowledge that α follows a continuous distribution μ_0 over $[\alpha^L, \alpha^U] \subseteq (0, 1]$, with CDF $F_0(\cdot)$ and PDF $f_0(\cdot)$. The impact of the non-delegatable project $\beta > 0$ is also common knowledge.

The principal decides between two responsibility assignments, $\{C, D\}$, i.e. centralization vs delegation. Under delegation, the agents gets to exert effort on the project, and the principal can exert effort on the non-delegatable project. Under centralization, the agent has no project, and the principal can exert effort on both projects.

Throughout the chapter, we assume that the principal's effort cost is small compared to the payoffs of either project; in equilibrium, she always exerts the upper limit of 1 unit of effort across the two projects. Her effort cost can therefore be ignored in the model. Let $e_P \in [0, 1]$ be the level of effort exerted on the α-impact delegatable project; $(1 - e_P)$ is thus exerted on the non-delegatable project. Let $e_A \in [0, 1]$ be the agent's effort exerted on the delegatable project.

The principal's payoff is thus

$$\begin{cases} \alpha e_P + \beta(1 - e_P) & g = C \\ \alpha e_A + \beta(1 - e_P) & g = D \end{cases}$$

whereas the agent's payoff is

$$\begin{cases} 0 & g = C \\ \gamma \alpha e_A - c(e_A) & g = D \end{cases}$$

where $c(e_A)$ is the agent's effort cost. We assume that $\gamma > 0$ is sufficiently small so that e_A is interior for all $\alpha \in [\alpha^L, \alpha^U]$. We also assume that $c'(\cdot) > 0$, $c''(\cdot) > 0$, $c'''(\cdot) \leq 0$, $c'(0) = 0$. For illustration, we assume $c(e_A) = \frac{e_A^2}{2}$ in this chapter, but the results are formally derived with a more general cost function $c(\cdot)$ such that $\lim_{e \to 0} (\ln c')' > 1$ and $\gamma > 0$ is sufficiently small. See Appendix C.3.1 for details.

Formally, the timing of the symmetric-information benchmark game is as follows.

Game B (Benchmark)

(1) The impact α is drawn and publicly observed.

(2) The principal chooses responsibility assignment $g \in \{C, D\}$.

(3) Both players choose effort level $e_P \in [0, 1]$ and $e_A \in [0, 1]$.

We analyze the pure-strategy Subgame-Perfect Nash Equilibrium of the game, which consists of the principal's responsibility assignment $g(\alpha) : [\alpha^L, \alpha^U] \to \{C, D\}$ and his effort choice $e_P(\alpha) : [\alpha^L, \alpha^U] \to [0, 1]$, as well as the agent's effort choice $e_A(\alpha) : [\alpha^L, \alpha^U] \to [0, 1]$ (on-path only if $g = D$).

Discussion

The identity of the player who exerts effort on the delegatable project does not affect the principal's payoff. This assumption reflects settings such as in a corporate hierarchy where a principal is assigned a project and takes credit irrespective of who among her team works on the project. If identity matters in the sense that principal and the agent each receive a fraction of the project's proceeds when she delegates, the results of the model continue to hold and the lemons problem is further exacerbated.

The parameter γ can be interpreted as either the agent's competence (in the sense that higher γ reduces the effort required obtain a given return from the project) or his alignment of objective with the principal (in the sense that higher γ implies a higher private benefit from the project). The results of the model hold so long as γ is small enough so that the agent exerts strictly less effort than the principal does on any project. See Appendix C.3.1 for proof of the existence of such an upper bound on γ.

3.1.2. Delegation threshold

We now describe the equilibrium of the symmetric-information benchmark. The principal delegates projects below a certain impact threshold and the agent exerts effort accordingly on the delegated project given its impact α: the higher the impact, the greater the effort. His effort level is an increasing function $h(\cdot)$ defined by $c'(h(x)) = x$.

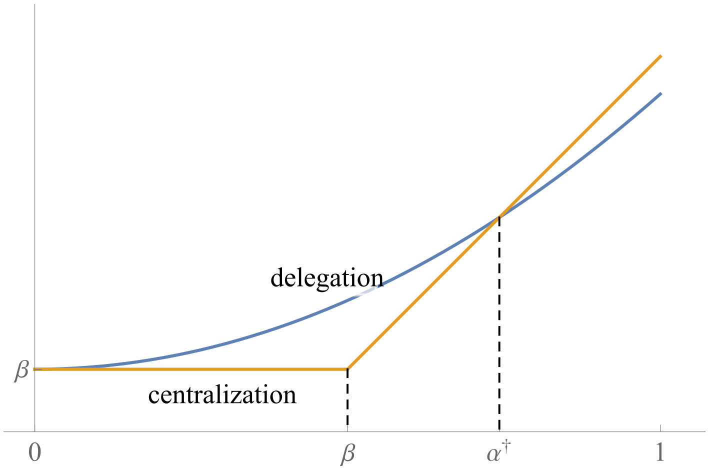

Figure 3.1.1. Principal's payoff from centralization vs delegation

Proposition 3.1. *There exists a threshold $\alpha^{\dagger} \in (\beta, \alpha^U]$ such that*

(1) *If $\alpha \leq \alpha^{\dagger}$, the principal delegates and exerts 1 unit of effort on the β-impact project. The agent exerts $h(\gamma\alpha)$ unit of effort on the α-impact project.*

(2) *Otherwise, the principal centralizes and exerts 1 unit of effort on the α-impact project.*

Proof. See Appendix C.3.1. □

The principal prefers to delegate projects with lower impact. When the project has sufficiently low impact ($\alpha \leq \beta$), delegation dominates since the principal herself would not work on it anyway. As the project's impact gets higher, it is still worthwhile to delegate to a point because the loss due to the agent's insufficient effort is compensated by her payoff gain of β from the other project.

This equilibrium result is illustrated in Figure 3.1.1. The principal's payoff from centralization u_C is a piecewise function which is continuous and convex in α and has a kink at β.

Her payoff from delegation, u_D, is increasing in α, illustrated in this figure for $c(e_A) = \frac{e_A^2}{2}$ and $\gamma = 0.22$. Notice that u_D is strictly convex and single-crosses u_C from below since its slope never exceeds 1 (see Appendix C.3.1 for the derivation of such properties). The unique interior intersection between u_D and u_C defines $\alpha^\dagger$.[1]

3.2. Delegating lemons

In this section, we consider a game in which the principal observes α privately before deciding whether to delegate. She cannot commit *ex ante* to an delegation rule, nor can she credibly communicate with the agent.[2]

We show that the principal delegates projects below a certain threshold of impact; compared to the symmetric-information benchmark, this threshold is necessarily lower. Therefore information asymmetry makes the principal delegate fewer projects, and these projects are of lower impact.

Formally, the game has the following timing.

Game L (Lemons).

(1) The impact α is drawn from μ_0 and privately observed by the principal.

(2) The principal chooses responsibility assignment $g \in \{C, D\}$.

(3) Both players choose effort level $e_P \in [0, 1]$ and $e_A \in [0, 1]$.

[1] If the slope of u_D exceeds 1 and u_D crosses u_C twice at two different thresholds $\alpha_1^\dagger, \alpha_2^\dagger$ where $\alpha_1^\dagger < \alpha_2^\dagger$, then the principal optimally delegates projects at both ends of the impact distribution. The agent exerts high enough effort for the high-impact projects (because of the curvature of his effort cost) so that the principal now loses sufficiently little from delegating these projects. Yet the general message of this model continues to hold, as it is straightforward to show that the threshold of delegation under information asymmetry is lower than $\alpha_1^\dagger$. So information asymmetry continues to make the principal delegate fewer projects, and these projects are of less impact.

[2] If she is allowed to cheap-talk, the equilibria of the game do not change. See discussion in Section 3.2.2.

We analyze the pure-strategy Perfect Bayesian Equilibrium of the game. An equilibrium of the game consists of the principal's responsibility assignment $g(\alpha) : [\alpha^L, \alpha^U] \to \{C, D\}$ and his effort choice $e_P(\alpha) : [\alpha^L, \alpha^U] \to [0, 1]$, as well as the agent's posterior belief of the delegated project's impact $\mu \in \Delta[\alpha^L, \alpha^U]$ and (with abuse of notation) his effort choice $e_A(\mu) : \Delta[\alpha^L, \alpha^U] \to [0, 1]$ (both on-path if $g = D$).[3] We do not distinguish among equilibria that differ only in beliefs or strategies in measure-zero events (including off the equilibrium path).

In Appendix C.3.2 we show that for the purpose of characterizing the agent's strategy and the principal's payoff, the posterior belief μ can be reduced to its expectation $\alpha_\mu \equiv \mathbb{E}_\mu[\alpha]$. Specifically, the principal's expected payoff from the delegated projects is $u_D(\alpha_\mu)$. Hence, with abuse of notation, we denote the agent's belief by $\alpha_\mu \in [\alpha^L, \alpha^U]$ and his effort choice by $e_A(\alpha_\mu) : [\alpha^L, \alpha^U] \to [0, 1]$.

To show intuition, for the rest of the chapter we focus on the case where $\alpha^L < \alpha^\dagger < \alpha^U$, i.e., some projects are delegated and some are not under the symmetric-information benchmark.[4] Similarly, we guarantee the uniqueness of equilibrium in this game by assuming $\mathcal{F}_0(\alpha) \equiv \int_{\alpha^L}^{\alpha} F_0(x)\mathrm{d}x$ is log-concave.[5] Notice that this assumption is implied by F_0 being log-concave, which in turn is implied by f_0 being log-concave.[6]

[3] The belief μ is defined over Borel-measurable sets.

[4] Without this assumption, Proposition 3.2 still holds with weak inequality.

[5] See Bagnoli and Bergstrom (2005) for a discussion on this property in probabilities.

[6] Without this assumption, Proposition 3.2 continues to hold with the qualifier "for every equilibrium". See proof of Proposition 3.2 in Appendix C.3.2 for details.

3.2.1. Delegation threshold

We first show that in any equlibrium of the lemons game where any delegation happens, the principal's delegation strategy must follow a threshold rule. We then derive the necessary and sufficient condition for such an equilibrium to exist.

Let $\mathcal{A}_D \equiv \{\alpha \in [\alpha^L, \alpha^U] : g(\alpha) = D\}$ be the set of projects that she delegates in equilibrium.

Lemma 3.1. *In any equilibrium of Game L where the set of delegated projects $\mathcal{A}_D$ has a non-zero measure, $\mathcal{A}_D$ takes the form of $[\alpha^L, \tilde{\alpha}]$ where $\tilde{\alpha} \geq \alpha^L$.*

Proof. See Appendix C.3.2. □

We call $\tilde{\alpha}$ the delegation threshold of such an equilibrium. The intuition for threshold delegation is similar to that in the benchmark game, with an important difference that now the agent exerts uniform effort across all delegated projects. This guarantees that for any effort that the agents exerts, the delegation threshold is unique. Because the loss from delegation is always higher for higher-impact projects, only low-impact projects are delegated.

If $\alpha^L < \alpha^\dagger < \alpha^U$, then the principal delegates some projects and centralizes others in equilibrium. The following analysis shows necessary and sufficient conditions for such an equilibrium to exist.

Notice that at the "threshold" project, the principal is indifferent between delegation and centralization, while the agent exerts her optimal effort given that the principal delegates projects below that threshold. Therefore we first construct a threshold payoff for the principal given the agent's posterior belief.

Definition. For each or each $\alpha \in [\alpha^L, \alpha^U]$, the threshold payoff is $\overline{u}_D(\alpha) = \alpha h(\gamma \mathbb{E}_{\mu_0}[\alpha' | \alpha' \leq \alpha]) + \beta$.

The payoff $\overline{u}_D(\alpha)$ is continuous, weakly below the delegation payoff from the benchmark game $u_D(\alpha)$, and strictly increasing in α. When it intersects u_C at some $\alpha^* \in (\alpha^L, \alpha^U)$, a "threshold delegation" equilibrium is identified.[7] This result is formalized in the following lemma.

Lemma 3.2. *An equilibrium exists for Game L in which the principal delegates according to a threshold $\alpha^* \in [\alpha^L, \alpha^U]$ if and only if $\overline{u}_D(\alpha^*) = u_C(\alpha^*)$. The agent's effort exertion is $h(\gamma \mathbb{E}_{\mu_0}[\alpha | \alpha \leq \alpha^*])$.*

Proof. See Appendix C.3.2. □

The construction of this equilibrium is illustrated in Figure 3.2.1. The prior distribution is uniform over the range $[\alpha^L, \alpha^U] = [0.1, 0.9]$. The dashed curve is the "threshold" payoff $\overline{u}_D$. It intersects the centralization payoff u_C at α^*, which is the delegation threshold in equilibrium. The principal's *ex post* payoff is the thick line with a kink at α^*. By construction, her payoff from delegation at α^* is $\overline{u}_D(\alpha^*)$, which equals her payoff from centralization $u_C(\alpha^*)$. To the left of α^*, delegation dominates centralization, and to the right of α^*, centralization dominates delegation. The agent's optimal effort is embedded in the slope of the straight line to the left of α^*.

[7] Notice that $\overline{u}_D$ and u_C may not intersect if $\alpha^L < \alpha^\dagger < \alpha^U$ is not satisfied. If $\overline{u}_D$ is weakly smaller than u_C on $[\alpha^L, \alpha^U]$, a "full centralization" equilibrium is identified. If $\overline{u}_D$ is weakly greater than u_C on $[\alpha^L, \alpha^U]$, a "full delegation" equilibrium is identified. See Lemmas C.15 and C.16 in Appendix C.3.2.

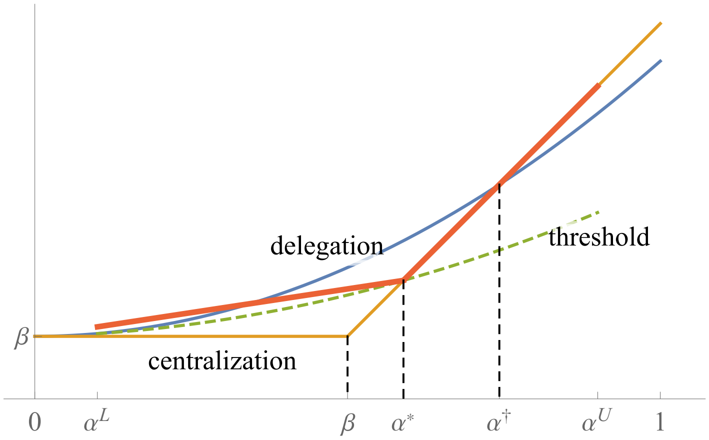

Figure 3.2.1. “Threshold delegation” equilibrium

3.2.2. “Insufficient” delegation

In this section, we establish the existence and uniqueness of the equilibrium in the lemons game. In doing so, we prove the key intuition of this analysis—the principal delegates less than she would in the benchmark game.

Proposition 3.2. *When $\alpha^L < \alpha^\dagger < \alpha^U$, there exists a unique equilibrium with delegation threshold $\alpha^* \in (\alpha^L, \alpha^U]$. Furthermore, $\alpha^* < \alpha^\dagger$.*

Proof. See Appendix C.3.2. □

The principal delegates fewer projects in the lemons game relative to the benchmark game, she must delegate strictly less in the lemons game.[8] Figure 3.2.1 illustrates this result. The dotted curve represents the delegation payoff with symmetric information (“benchmark”). The principal optimally centralizes the projects with the highest impact

[8] This intuition still holds with weak inequality if $\alpha^L < \alpha^\dagger < \alpha^U$ does not hold. If $\alpha^\dagger \leq \alpha^L$, i.e., the principal delegates all projects in the Game B, she necessarily (albeit trivially) delegates weakly less in Game L. If $\alpha^\dagger \geq \alpha^U$, i.e., the principal does not delegate at all in Game B, she does not delegate in Game L either.

among those projects originally delegated in the benchmark game. She does so because the agent's effort is too low for these projects. Delegation also has a negative signaling effect. As the agent infers that the average project delegated is of lower impact compared to the benchmark, he lowers his effort accordingly. This effort choice in turn makes the principal exclude even more projects from delegation. This feedback loop pushes the delegation threshold downwards; in equilibrium, the principal delegates only the very low impact projects and the agent exerts very little effort.

Notice that the principal is worse off compared to the benchmark game.

Corollary 3.1. *The principal's expected payoff in Game L is lower than in Game B.*

Proof. See Appendix C.3.2. □

As shown in the proof in Appendix C.3.2, the payoff loss has two sources. One source is that the agent exerts a uniform effort across all delegated projects rather than according to each project's impact. The other source is the fact that there is "insufficient" delegation in equilibrium.

This mechanism suggests that a project may appear suitable for delegation if assessed in hindsight when α is known, but a manager may find delegation suboptimal at the time of making the delegation decision, as she anticipates low effort if she delegates the project. In fact, whether a project is suitable for delegation depends on the distribution of all delegatable projects. An implication is that the principal is more likely to delegate the same project in an environment where high return projects that can be delegated abound.

Note that cheap-talk communication is not enough to remedy such inefficiency. Indeed, cheap-talk results in babbling: the principal has a monotone preference over the agent's effort, so if different signals induce different efforts then she would always send the signal that induces the most effort. In equilibrium, all signals must induce the same effort from the agent.

3.3. Remedies

Since information asymmetry leads to a lower (expected) payoff for the principal, she would want to share information with the agent. As previously argued, however, cheap-talk does not allow credible communication, as the principal cannot help inducing the agent to believe that the project has high impact and thus losing her credibility. We therefore turn to other mechanisms that enable the principal to credibly communicate her information, namely commitment to a signaling strategy and voluntary disclosure with evidence. We also discuss whether committing to a delegation strategy mitigates the inefficiency.

3.3.1. Credible communication

Suppose the principal can commit to a signaling strategy before observing α. The principal chooses a signal space S and a signaling strategy $s : [0,1] \to \Delta S$; after the draw α, $s(\alpha)$ is realized and the agent observes it before the principal chooses whether to delegate.[9] We show that the principal optimally commits to revealing the impacts of those projects which are delegated in the benchmark model, and that she subsequently delegates them.

[9] How the principal finds commitment power is outside the scope of the chapter; one possibility is that repeated interactions with agents may create reputational concern so that the principal finds it in her interest to save future credibility by not lying today. Indeed the full revelation strategy can be sustained by the long-run incentive as shown in Best and Quigley (2017).

The reason for not pooling any projects is that the principal's payoff is convex in α so any pooling always entails a loss.

Proposition 3.3. *If the principal can commit to a signaling strategy over the impacts of projects, she optimally commits to revealing α if $\alpha \in [\alpha^L, \alpha^\dagger]$ and she subsequently delegates these projects if one such project is realized.*

Proof. As in Kamenica and Gentzkow (2011), it is without loss of generality to identify s with the posterior belief that it induces. Because the principal's payoff is convex in the agent's posterior belief in α, it is never optimal for the principal to induce a non-degenerate posterior belief over delegated projects.

Since in equilibrium, the agent knows the impact of a delegated project, it then follows from the analysis in Section 3.1.2 that the principal wants to delegate projects with $\alpha \in [\alpha^L, \alpha^\dagger]$. Hence her optimal signaling strategy must fully reveal these impacts, and she subsequently prefers to delegate these projects. □

Given that full revelation is optimal, commitment to a signaling strategy can be replicated by verifiable information. Consider a game in which the principal has verifiable information about α à la Grossman (1981); she chooses whether reveal it to the agent when choosing whether to delegate. We show that the principal optimally discloses the impacts of those projects which are delegated in the benchmark model, and that she subsequently delegates them.[10]

[10] If instead the principal can generate verifiable information randomly à la Dye-Jung-Kwon (Dye, 1985; Jung and Kwon, 1988), partial disclosure is optimal. When the principal can generate verifiable information, she optimally disclose $\alpha \in [\alpha_1, \alpha^\dagger]$ where $\alpha_1 > \alpha^L$, and she delegates the project if $\alpha \leq \alpha^\dagger$. When she cannot generate verifiable information, she optimally delegate projects with $\alpha \leq \alpha_2$ where $\alpha_1 < \alpha_2 < \alpha^\dagger$. See Appendix B for a discussion.

Proposition 3.4. *If the principal can generate verifiable information about α, she optimally discloses α if $\alpha \in [\alpha^L, \alpha^\dagger]$ and she subsequently delegates these projects.*

Proof. As argued in Grossman (1981), the principal optimally discloses the impact of the delegated projects in any equilibrium because her preference is monotonic in the agent's posterior belief. It is straightforward to show that if the principal induces any non-degenerate posterior belief over delegated projects, then she always has a profitable deviation of revealing the highest-impact project among them. Therefore for any project to be subsequently delegated, the principal's equilibrium strategy must fully disclose its impact. It then follows from the analysis in Section 3.1.2 that the principal wants to delegate projects with $\alpha \in [\alpha^L, \alpha^\dagger]$, so the unique equilibrium strategy is to disclose their impacts at the same time. □

3.3.2. Commitment to a delegation strategy

We now consider a scenario where the principal can commit to a delegation strategy. Intuitively, the principal could be better off if she can commit to delegate an optimal pool of projects.

We model this game with commitment to delegation strategy in the following way:

Game C (Commitment)

(1) The principal publicly chooses delegation strategy $g(\alpha) : [0, 1] \to \{C, D\}$.

(2) The impact α is drawn and privately observed by the principal.

(3) Responsibility assignment g is realized and communicated to the agent.

(4) Both players choose effort level $e_P \in [0,1]$ and $e_A \in [0,1]$.

Compared to Game L, the principal now commits to the responsibility assignment for each α. In what follows, we treat strategies that differ on zero-probability events as the essentially same. We show that, the principal's equilibrium strategy is essentially the same as centralizing the projects with the highest and lowest impact and she delegating projects with an intermediate impact.

Proposition 3.5. *In the equilibrium of Game C, the set of delegated projects* $\bar{\mathcal{A}}_D$ *takes the form* $[\bar{\alpha}^L, \bar{\alpha}^U]$ *where* $\bar{\alpha}^L \leq \beta \leq \bar{\alpha}^U$*.*

Proof. See Appendix C.3.3. □

To understand the two thresholds $\bar{\alpha}^L$ and $\bar{\alpha}^U$, note that there are two effects of delegating a project. First, delegating a project changes the payoff from that project. Second, delegating a project affects the agent's belief, and hence effort, on the set of delegated projects. For projects with low impact ($\alpha < \beta$), both effects become more positive as α increases: the principal gains more from delegating a project of higher impact that she would not work on anyway, and such a project depresses less the agent's effort across delegated projects. Therefore, no gap can exist on the set of delegated projects on this interval, i.e., $[\bar{\alpha}^L, \beta]$, because if it is worth delegating projects with impacts below the gap, then it is worth delegating projects within this gap.

For projects of high impact ($\alpha > \beta$), the agent exerts less effort than the principal would under centralization. This payoff loss increases with the impact of the project. However, committing to delegating such projects also induces the agent to exert more effort on *all* delegated projects. In the proof we show that the first effect dominates the second, so the overall benefit of delegation decreases with project's impact. Therefore, no gap can exist on the set of delegated projects on this interval either, i.e., $[\beta, \bar{\alpha}^U]$.

Naturally, commitment to a delegation strategy improves the principal's payoff: any feasible strategy in the lemons game can be replicated in the commitment game. Compared to the threshold delegation strategy, part of the improvement comes from abandoning the very low-return projects to increase the agent's effort over delegated projects. Improvement may also come from delegating projects of impact higher than α^* to increase the agent's effort on other projects. In Appendix C.3.3 we show that the upper bound of the optimal delegation interval is strictly above α^* if the principal's payoff from interval delegation is concave in this upper bound, which holds, for example, if impacts are distributed uniformly and the agent's effort cost is quadratic.

3.4. Conclusion

This chapter emphasizes the importance of information environment in shaping a manager's delegation decisions. A manager is expected to delegate less when the subordinate does not know the returns of delegated projects and the manager cannot credibly share this information. Consequently the manager is worse off than when such information is transparent. Different remedies to restore credible communication or to improve the delegation strategy are also discussed.

Other forces that are assumed away in the chapter also play important roles in a manager's delegation decisions, and suggests different directions where this chapter can be extended. There could be interesting interactions, for instance, between a manager with multiple projects and multiple subordinates of heterogenous capacities. Deciding which agents should work on which projects then have more subtle signaling effects. Repeated interactions may also generate surprising delegation patterns: it may create commitment power as suggested in Section 3.3.1, but it might also affect the subordinates' effort incentive calculation.

The mechanism described in the chapter sheds light on the difficulty of understanding delegation decisions without understanding a manager's information environment. Testable implications are also suggested for when some features of the information environment are observed. More empirical and theoretical studies on the relationship between delegation and workload would further an understanding of this fundamental process in management.

Bibliography

Acemoglu, Egorov, and Sonin (2012) “Dynamics and Stability of Constitutions, Coalitions, and Clubs,” *American Economic Review*, 6, 1–25.

Aghion, Philippe and Jean Tirole (1997) “Formal and Real Authority in Organizations,” *Journal of Political Economy*, 105 (1), 1–29.

Akerlof, George A. (1970) “The Market for ”Lemons”: Quality Uncertainty and the Market Mechanism,” *The Quarterly Journal of Economics*, 84 (3), 488.

Alonso, Ricardo, Wouter Dessein, and Niko Matouschek (2008) “When Does Coordination Require Centralization?,” *American Economic Review*, 98 (1), 145–179.

Alonso, Ricardo and Niko Matouschek (2007) “Relational delegation,” *RAND Journal of Economics*, 38 (4), 1070–1089.

——— (2008) “Optimal Delegation,” *Review of Economic Studies*, 75 (1), 259–293.

American National Standards Institute (2021) “ANSI Essential Requirements: Due process requirements for American National Standards,” `https://www.ansi.org/essentialrequirements`.

Arnold, Gretchen (1995) “Dilemmas of Feminist Coalitions: Collective Identity and Strategic Effectiveness in the Battered Women’s Movement,” in Ferree, Myra Marx and Patricia Yancey Martin eds. *Feminist Organizations: Harvest of the New Women’s Movement*, Chap. 18, 276–290.

Austin, Julia (2017) “What It Takes to Become a Great Product Manager,” `https://hbr.org/2017/12/what-it-takes-to-become-a-great-product-manager`.

Bagnoli, Mark and Ted Bergstrom (2005) “Log-concave probability and its applications,” *Economic Theory*, 26 (2), 445–469.

Baker, George (1992) “Incentive Contracts and Performance Measurement,” *Journal of Political Economy*, 100 (3), 598–614.

Baker, George, Robert Gibbons, and Kevin J. Murphy (1994) “Subjective Performance Measures in Optimal Incentive Contracts,” *The Quarterly Journal of Economics*, 109 (4), 1125–1156.

——— (1999) “Informal authority in organizations,” *Journal of Law, Economics, and Organization*, 15 (1), 56–73.

Bandy, Joe and Jackie Smith eds. (2005) *Coalitions across Borders: Transnational Protest and the Neoliberal Order*, Oxford: Rowman & Littlefield Publishers, Inc. xvi, 262.

Barber, Lucy Grace (2002) *Marching on Washington: The Forging of an American Political Tradition*: University of California Press.

Barberà, Salvador, Carmen Beviá, and Clara Ponsatí (2015) “Meritocracy, egalitarianism and the stability of majoritarian organizations,” *Games and Economic Behavior*, 91, 237–257.

Barkan, Steven E. (1986) “Interorganizational Conflict in the Southern Civil Rights Movement,” *Sociological Inquiry*, 56 (2), 190–209.

Barnard, Chester I. (1938) *The Functions of the Executives*: Harvard University Press.

Baron, Justus, Jorge Contreras, Martin Husovec, and Pierre Larouche (2019) “Making the Rules. The Governance of Standard Development Organizations and their Policies on

Intellectual Property Rights,"Technical report, European Commission, Joint Research Centre.

Beamish, Thomas D. and Amy J. Luebbers (2009) "Alliance Building across Social Movements: Bridging Difference in a Peace and Justice Coalition," *Social Problems*, 56 (4), 647–676.

Best, James and Daniel Quigley (2017) "Persuasion for the Long Run."

Beyer, Anne, Ilan Guttman, and Iván Marinovic (2014) "Optimal contracts with performance manipulation," *Journal of Accounting Research*, 52 (4), 817–847.

Bonatti, Alessandro and Heikki Rantakari (2016) "The politics of compromise," *American Economic Review*, 106 (2), 229–259.

Branch, Taylor (1990) *Parting the Waters: America in the King Years, 1954-63*, 56: Simon & Schuster, 562.

Brooks, Ethel (2005) "Transnatinal Campaigns against Child Labor: The Garment Industry in Bangladesh," in *Coalitions across Borders: Transnational Protest and the Neoliberal Order*, Chap. 6.

Bystydzienski, Jill M. and Steven P. Schacht eds. (2001) *Forging Radical Alliances Across Difference: Coalition Politics for the New Millennium*: Rowman & Littlefield Publishers.

Cai, Yongshun (2017) *The Occupy Movement in Hong Kong: Sustaining Decentralized Protest*, Routledge contemporary China series ; 157, Abingdon, Oxon ; New York, NY: Routledge.

Cohen, David (2021) "Why the DOJ's latest IEEE move is a giveaway to Big Tech at the expense of US technology leadership," `https://bit.ly/3NWjRNN`.

Condorelli, Daniele and Balázs Szentes (2020) "Information design in the holdup problem," *Journal of Political Economy*, 128 (2), 681–709.

Crocker, Keith J. and Joel Slemrod (2008) "The economics of earnings manipulation and managerial compensation," *The RAND Journal of Economics*, 38 (3), 698–713.

Curtis, Russell L. and Louis A. Zurcher (1973) "Stable resources of protest movements: The multi-organizational field," *Social Forces*, 52 (1), 53–61.

Deimen, Inga and Dezso Szalay (2019) "Delegated expertise, authority, and communication," *American Economic Review*, 109 (4), 1349–1374.

Della Porta, Donatella (2009) "Consensus in Movements," in Della Porta, Donatella ed. *Democracy in Social Movements*, Chap. 3.

Della Porta, Donatella and Mario Diani (2006) *Social Movement: An Introduction*.

Dessein, Wouter (2002) "Authority and Communication in Organizations," *Review of Economic Studies*, 69, 811–838.

——— (2005) "Information and Control in Ventures," *The Journal of Finance*, LX (5), 2513–2549.

Dessein, Wouter, Andrea Galeotti, and Tano Santos (2016) "Rational inattention and organizational focus," *American Economic Review*, 106 (6), 1522–1536.

Doctorow, Cory (2017) "An open letter to the W3C Director, CEO, team and membership," `https://www.eff.org/deeplinks/2017/09/open-letter-w3c-director-ceo-team-and-membership`.

Dye, Ronald A. (1985) "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, 23 (1), 123.

——— (1988) "Earnings Management in an Overlapping Generations Model," *Journal of Accounting Research*, 26 (2), 195.

European Union (2016) "Consolidated versions of the Treaty on the Functioning of the European Union, OJ C 202, 7.6.2016."

Farrell, Joseph and Timothy Simcoe (2012) "Choosing the rules for consensus standardization," *RAND Journal of Economics*, 43 (2), 235–252.

Fehr, Ernst and Klaus M. Schmidt (2004) "Fairness and incentives in a multi-task principal-agent model," *Scandinavian Journal of Economics*, 106 (3), 453–474.

Gallo, Oihane and Elena Inarra (2018) "Rationing rules and stable coalition structures," *Theoretical Economics*, 13 (3), 933–950.

Gamson, William A. (1961) "A Theory of Coalition Formation," *American Sociological Review*, 26 (3), 373–382.

——— (1975) *The Strategy of Social Protest*: Dorsey Press.

Ganz, Marshall (2000) "Resources and resourcefulness: Strategic capacity in the unionization of California agriculture, 1959-1966," *American Journal of Sociology*, 10 (4), 1003–1062.

Garicano, Luis and Tano Santos (2004) "Referrals," *American Economic Review*, 94 (3), 499–525.

Garrett, Daniel Ferguson, George Georgiadis, Alex Smolin, and Balazs Szentes (2021) "Optimal Technology Design," *Mimeo* (May).

Gemmill, Gary R. and David L. Wilemon (1972) "The Product Manager as an Influence Agent," *Journal of Marketing*, 36 (1), 26.

General Secretariat of the Council (2009) *Rules of Procedure of the Council* (December).

Gibbons, Robert (1998) “Incentives in Organizations,” *Journal of Economic Perspectives*, 12 (4), 115–132.

Gottfried, Heidi and Penny Weiss (1994) “A Compound Feminist Organization: Purdue University’s Council on the Status of Women,” *Women and Politics*, 14 (2), 23–44.

Grossman, Sanford (1981) “The Informational Role of Warranties and Private Disclosure about Product Quality,” *The Journal of Law and Economics*, 24 (3), 461–483.

Grossman, Sanford and Oliver Hart (1986) “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration,” *Journal of Political Economy*, 94 (4), 691–719.

Gul, Faruk (2001) “Unobservable investment and the hold-up problem,” *Econometrica*, 69 (2), 343–376.

Harris, Milton and Artur Raviv (1979) “Optimal incentive contracts with imperfect information,” *Journal of Economic Theory*, 20 (2), 231–259.

——— (2005) “Allocation of Decision-making Authority,” *Review of Finance*, 9 (3), 353–383.

Hart, Oliver and John Moore (1990) “Property Rights and the Nature of the Firm,” *Journal of Political Economy*, 98 (6), 1119–1158.

Hermalin, Benjamin E and Michael L Katz (2009) “Information and the hold-up problem,” *RAND Journal of Economics*, 40 (3), 405–423.

Holmstrom, Bengt (1979) “Moral Hazard and Observability,” *The Bell Journal of Economics*, 10 (1), 74.

Holmstrom, Bengt and Paul Milgrom (1991) “Multitask Principal-Agent Analyses: Incentive Contracts , Asset Ownership , and Job Design,” *Journal of Law, Economics,* *and Organization*, 7 (January 1991), 24–52.

——— (1994) "The Firm as an Incentive System," *American Economic Review*, 84 (4), 972–991.

Jewitt, Ian, Ohad Kadan, and Jeroen M. Swinkels (2008) "Moral hazard with bounded payments," *Journal of Economic Theory*, 143 (1), 59–82.

Jung, Woon-Oh and Young K. Kwon (1988) "Disclosure When the Market Is Unsure of Information Endowment of Managers," *Journal of Accounting Research*, 26 (1), 146.

Kamenica, Emir and Matthew Gentzkow (2011) "Bayesian persuasion," *American Economic Review*, 101 (6), 2590–2615.

Kerr, Setven (1975) "On the Folly of Rewarding A, While Hoping for B," *Academy of Management Journal*, 18 (4), 769–783.

Kidd, Dustin and Keith McIntosh (2016) "Social Media and Social Movements," *Sociology Compass*, 10 (9), 785–794.

Koufopoulos, Kostas, Roman Kozhan, and Giulio Trigilia (2019) "Optimal security design under asymmetric information and profit manipulation," *Review of Corporate Finance Studies*, 8 (1), 146–173.

Kuhn, Anthony (2019) "In Hong Kong, Moderate And Radical Protesters Join Forces To Avoid Past Divisions," `https://n.pr/3LQVVJI`.

Lau, Stephanie (2008) "Information and bargaining in the hold-up problem," *RAND Journal of Economics*, 39 (1), 266–282.

Lauzier, Jean-Gabriel (2021) "Ex-post moral hazard and manipulation-proof contracts."

Leana, Carrie (1986) "Predictors and Consequences of Delegation," *Academy of Management Journal*, 29 (4), 754–774.

Levi, Margaret and Gillian H. Murphy (2006) “Coalitions of contention: The case of the WTO protests in Seattle,” *Political Studies*, 54 (4), 651–670.

Levitt, Arthur (1998) “The ”Numbers Game”,” `https://www.sec.gov/news/speech/speecharchive/1998/spch220.txt`.

Lewis, John and Michael D’Orso (1999) *Walking with the Wind: A Memoir of the Movement*: Harcourt Brace.

Li, Jin, Niko Matouschek, and Michael Powell (2017) “Power dynamics in organizations,” *American Economic Journal: Microeconomics*, 9 (1), 217–241.

Li, Siguang and Xi Weng (2017) “Random Authority,” *International Economic Review*, 58 (1), 211–235.

Lindbeck, Assar and Jörgen W. Weibull (1987) “Balanced-budget redistribution as the outcome of political competition,” *Public Choice*, 52 (3), 273–297.

March, James and Herbert Simon (1958) *Organizations*: John Wiley & Sons, Inc.

Marino, Anthony M., John G. Matsusaka, and Ján Zábojník (2010) “Disobedience and authority,” *Journal of Law, Economics, and Organization*, 26 (3), 427–459.

Milgrom, Paul (1988) “Employment Contracts , Influence Activities , and Efficient Organization Design,” *Journal of Political Economy*, 96 (1), 42–60.

Milgrom, Paul and Chris Shannon (1994) “Monotone Comparative Statics,” *Econometrica*, 62 (1), 157.

Morelli, Massimo and In Uck Park (2016) “Internal hierarchy and stable coalition structures,” *Games and Economic Behavior*, 96, 90–96.

Ng, Ellie (2014) “Hong Kong’s House, Divided,” `https://foreignpolicy.com/2014/11/25/hong-kongs-house-divided/`.

Obach, Brian (1999) "The Wisconsin Labor-Environmental Network," *Organization & Environment*, 12 (1), 45–74.

——— (2004) *Labor and the Environmental Movement*, Cambridge: The MIT Press.

Polletta, Francesca (2002) *Freedom Is an Endless Meeting*: University of Chicago Press.

——— (2013) "Participatory Democracy in Social Movements," *The Wiley-Blackwell Encyclopedia of Social and Political Movements.*

Pullum, Amanda (2018) "Foul Weather Friends: Enabling Movement Alliance through an Intentionally Limited Coalition," *Social Currents*, 5 (3), 228–243.

Rantakari, Heikki (2008) "Governing Adaptation," *Review of Economic Studies*, 75 (4), 1257–1285.

Ray, Debraj and Rajiv Vohra (2015) *Coalition Formation*, 4: Elsevier B.V. 239–326.

Reh, Christine (2012) "European Integration as Compromise: Recognition, Concessions and the Limits of Cooperation," *Government and Opposition*, 47 (3), 414–440.

Rustin, Bayard (2003) *Time on Two Crosses: The Collected Writings of Bayard Rustin*, San Francisco: Cleis Press.

Shavell, Steven (1979) "Risk Sharing and Incentives in the Principal and Agent Relationship," *The Bell Journal of Economics*, 10 (1), 55.

Simcoe, Timothy (2012) "Standard setting committees: Consensus governance for shared technology platforms," *American Economic Review*, 102 (1), 305–336.

——— (2014) "Governing the anticommons: Institutional design for standard- setting organizations," *Innovation Policy and the Economy*, 14 (1), 99–128.

Staggenborg, Suzanne (1986) "Coalition Work in the Pro-Choice Movement : Organizational and Environmental Opportunities and Obstacles," *Social Problems*, 33 (5), 374–390.

Stiglitz, Joseph (2006) *Making Globalization Work*, New York: W.W. Norton & Company, Inc.

Sun, Bo (2014) “Executive compensation and earnings management under moral hazard,” *Journal of Economic Dynamics and Control*, 41, 276–290.

Swank, Otto H. and Bauke Visser (2015) “Learning from others? Decision rights, strategic communication, and reputational concerns,” *American Economic Journal: Microeconomics*, 7 (4), 109–149.

Topkis, Donald M (1998) *Supermodularity and Complementarity*: Princeton University Press.

Tufekci, Zeynep (2017) *Twitter and The Tear Gas*: Yale University Press, 359.

Van Dyke, Nella and Bryan Amos (2017) “Social movement coalitions: Formation, longevity, and success,” *Sociology Compass*, 11 (7), 1–17.

Vogel, Ezra F. ed. (1975) *Modern Japanese Organization and Decision-Making*, Berkeley: University of California Press.

Weiss, Nancy J. (1986) “Creative Tensions in the Leadership of the Civil Rights Movement,” in Eagles, Charles W. ed. *The Civil Rights Movement in America*: University Press of Mississippi.

Wood, Lesley J. (2005) “Bridging the Chasms: The Case of Peoples’ Global Action,” in Bandy, Joe and Jackie Smith eds. *Coalitions across Borders: Transnational Protest and the Neoliberal Order*, Chap. 5, 95–117: Rowman & Littlefield Publishers, Inc.

X, Malcolm and Alex Haley (1965) *The Autobiography of Malcolm X*: Grove Press.

Yukl, Gary and Ping Ping Fu (1999) “Determinant of delegation and consultation by managers,” *Journal of Organizational Behaviour*, 20 (2), 219–232.

Zald, Mayer N. and Roberta Ash (1966) “Social Movement Organizations: Growth, Decay and Change,” *Social Forces*, 44 (3), 327.

Zald, Mayer N. and John D. McCarthy (2017) “Social movement industries: Competition and conflict among SMOs,” in *Social Movements in an Organizational Society: Collected Essays*, 161–182.

APPENDIX A

A Resource Model of Synergies and Incentives in Coalitions

In this section, we develop a model that endogenizes the synergies that members receive and the optimal incentive provision that the coordinator uses to induce members to take actions in Chapter 1. This model singles out resources as the source of synergy and the coordinator's right to use these resources as the basis for its incentive provision. It is inspired by the literature on social movements that discusses how the sharing of resources reinforces the cooperation among the constituent groups (Obach, 1999, 2004); for example, the successful episodes of the civil rights movement demonstrate that grassroots mobilization, political access and legal advocacy are all valuable but often supplied by different organizations (Barkan, 1986; Weiss, 1986).

In this model, a member joins the coalition by transferring the right to use his resource to the coalition coordinator. The coordinator coordinates by specifying how it would use members' resources conditional on the actions that members take. A member's synergy from the coalition is thus the net increase in his benefit as the result of this non-autarkic use of resources.

Formally, each member can transfer to the coalition the right to use one indivisible unit of his resource, and $x_i = 1$ indicates that member i makes such transfer. We assume that it takes time to reverse this decision so that members can take back this right only at the end of the game. In other words, members can leave the coalition only after actions have been taken on the issue of disagreement.

In addition to his payoff (or rather, losses) from the issue, each member has his private benefit (e.g., reputation, safety, maintanance), realized at the end of the game and determined by $f_i(y_i^1, y_i^2)$ where $y_i^j \in \{0, 1\}$ indicates whether the use of member j's resource is used to member i's benefit. We assume that f_i is strictly monotone in both arguments, and that $f_i(y_i^i = 1, y_i^j = 0) < f_i(y_i^i = 0, y_i^j = 1)$ where $j \neq i$ for all $i \in \{1, 2\}$, i.e., the coalition can improve upon autarky for each member.[1] We normalize the benefits by assuming the benefit from the autarkic use of resources is $f_i(y_i^i = 1, y_i^j = 0) = 0$ for all $i \in \{1, 2\}$ and $j \neq i$. Consequently, the benefit from "cross-use" is $\omega_i = f_i(y_i^i = 0, y_i^j = 1) > 0$, the benefit from both resources is $\omega_i^+ = f_i(y_i^i = 1, y_i^j = 1) > 0$ and the benefit from no resource is $\omega_i^- = f_i(y_i^i = 0, y_i^j = 0) < 0$.

We assume that the coordinator does not care about members' private benefits, and consequently it is without loss of generality to assume away free disposal as in equilibrium it cannot improve the coordinator's provision of incentive for the members. The members' benefits from four alternative uses of the resources are shown in the following table.

	$y_2^2 = 0$	$y_2^2 = 1$
$y_1^1 = 0$	(ω_1, ω_2)	(ω_1^-, ω_2^+)
$y_1^1 = 1$	(ω_1^+, ω_2^-)	$(0, 0)$

[1] A justification is that the marginal productivity of the other member's resource is very large when a member is endowed with none of it.

If the coalition is formed, the coordinator publicly commits to an arbitrary set $\{(a_1, a_2, y_1^1, y_2^2)\}$ where each element $(a_1, a_2, y_1^1, y_2^2) \in \mathbb{R}^2 \times \{0,1\}^2$ represents an actions pair and the resource use that the coordinator consequentially implement. Members simultaneously take action afterwards, and then chooses whether to take back the right to use his resource *before* the coordinator's planned use of resources becomes effective and private benefits are realized. Naturally, if the coalition is not formed or if a member takes back his resource, then only the autarkic benefits $(0, 0)$ is realized.

Instead of the giving a full solution, we highlight two key features of the players' equilibrium behavior in the enriched parts of the game, and point out how they correspond to the modeling elements in the baseline model.

We first highlight that it is optimal to associate one pair of actions—the pair that the coordinator wants to induce—with the "cross-use" of resources. Accordingly, in the baseline model we call (ω_1, ω_2) members' synergies, and let the coordinator choose a single pair of actions as its recommendation. It is optimal for the coordinator to induce a single pair of actions because of the convexity of its payoff function. Because "cross-use" is the only one that benefits both members compared to autarky, the coordinator must associate it with the pair of actions induced in equilibrium in order to incentivize participation by both members. It is then optimal to associate all other pairs of actions with the maximal penalty that the coordinator can impose on the members. Analogous to the "forcing contract" in risk-free principal-agent models (Harris and Raviv, 1979), this form of incentive generates the largest set of recommendations that member could accept, which is necessary for the coordinator's optimal recommendation.

We also highlight that the maximal penalty that the coordinator can impose is autarky. Accordingly, in the baseline model, we associate any deviation from the coordinator's recommendation with the forfeiture of synergy. Nominally, the maximal penalty is to use no resources on a deviating member (and it does not matter which member if they both deviate), but it induces the deviating member to take back the right to use his resource so that he receives the autarkic payoff 0 instead of the penalty payoff $\omega_i^- < 0$. Therefore autarky is the maximal "feasible" penalty, and, it is without loss of generality that we require a member to *leave* as he deviates from the coordinator's recommendation in the baseline model.

APPENDIX B

Partial disclosure

In this section, we provide a conjecture as suggested in Chapter 3. Consider a game where the principal is randomly capable of generating verifiable information about project impact à la Dye-Jung-Kwon (Dye, 1985; Jung and Kwon, 1988), and he chooses whether to reveal such information to the agent while choosing whether to delegate. We argue that the unique equilibrium strategy of the principal involves at most three thresholds $\alpha_1 < \alpha_2 \leq \alpha^{\dagger}$.[1]

Equilibrium responsibility assignment and information disclosure			
If P can generate information	$\alpha \in [\alpha^L, \alpha_1)$	$\alpha \in [\alpha_1, \alpha^{\dagger})$	$\alpha \in [\alpha^{\dagger}, \alpha^U]$
	$D, \emptyset$	$D, \{\alpha\}$	$C, \{\alpha\}$
If P cannot generate information	$\alpha \in [\alpha^L, \alpha_2)$	$\alpha \in [\alpha_2, \alpha^U]$	
	$D, \emptyset$	$C, \emptyset$	

To understand the threashold α_2, consider the case where the principal *can* generate information but faces the lowest-impact projects, i.e., $\alpha < \alpha_1$. Compared to the case where she can *always* generate information, she now prefers to hide their impacts and delegate them, pooling them with the cases where she cannot generate such information.

[1] Notice that α_1 may coincide with α^L and α_2 with $\alpha^{\dagger}$ if α^L is sufficiently high.

Therefore when the agent is delegated a project without being disclosed its impact, he exerts lower effort than in Game L. The analysis of Game L then implies that the principal optimally delegates projects without impacts below a threshold $\alpha_2 \leq \alpha^{\dagger}$.

APPENDIX C

Proofs

C.1. Incentives in Coalitions

In this section we prove results for the general N-member game. Specific results for the two-member game are given when appropriate.

C.1.1. Disintegration actions, exertion of power and partition of members

Ignore the non-generic case of $\bar{\theta} = \theta_i$ for some member i, which is ruled out by the sufficient disagreement assumption that we later derive. We first show that members are partitioned by $\bar{\theta}$ into two groups: $a_i^D < a_{-i}^D$ if $\theta_i < \bar{\theta}$, and vice versa. Notice that in the case of $N = 2$, $\kappa_1 \neq \kappa_2$ does not affect this result, as the partition of the two members is fixed: each member is always his own group.

By definition, a disintegration action must be the best-response to other disintegration actions, i.e.,

$$a_i^D = (1-\kappa)\theta_i + \kappa a_{-i}^D$$

Summing over i and dividing by N, and with some manipulation, we get

$$a_i^D - a_{-i}^D = \frac{N(1-\kappa)}{N-1+\kappa}(\theta_i - \bar{\theta})$$

which forms the basis for partitioning members into $\mathcal{I} \equiv \{i|\theta_i < \bar{\theta}\}$ and $\mathcal{J} \equiv \{j|\theta_j > \bar{\theta}\}$.

Throughout the analysis, this partition remains fixed in the sense that irrespective of the power exerted over the members, $a_i < a_{-i}$ if $\theta_i - \bar{\theta}$ and vice versa. Disintegration is but a special case of zero power exerted over the members. Fixed partition is guaranteed by the sufficient disagreement assumption, which stipulates that the distance between a member's ideology and the average ideology is sufficiently large compared to the coalition's power over its members, as the former determines the magnitude of power needed to reverse the order of a_i and a_{-i}.

To derive the sufficient disagreement assumption, we need to characterize the effects of power exerted over the members. Recall that $\tilde{a}_i(\mathbf{e}), \forall i \in \mathcal{N}$ is defined as the unique solution (uniqueness guaranteed by contraction mapping) to the following system:

$$\tilde{a}_i = a_i^{BR}(\tilde{\mathbf{a}}_{-i}) + e_i, \forall i \in \mathcal{N}$$

It is straightforward to derive the following derivatives (all are constants irrespective of $\mathbf{e}$) where $i \neq j$.

$$\frac{\partial \tilde{a}_i}{\partial e_i} = \frac{(N-1)(1-\kappa)+\kappa}{(N-1+\kappa)(1-\kappa)}$$

$$\frac{\partial \tilde{a}_{-i}}{\partial e_i} = \frac{\partial \tilde{a}_j}{\partial e_i} = \frac{\kappa}{(N-1+\kappa)(1-\kappa)}$$

$$\frac{\partial \tilde{a}_{-j}}{\partial e_i} = \frac{1}{(N-1+\kappa)(1-\kappa)}$$

As in the two-member model, the direct effect of power exertion on a member always dominates the (average) strategic effect on other members. Consequently, for all feasible exertion of power $\mathbf{e} \in \mathcal{E}$, if $\bar{\theta} > \theta_i$ then $\tilde{a}_{-i} - \tilde{a}_i$ is minimized at $(-\delta_1, -\delta_2, \cdots, \delta_i, \cdots, -\delta_N)$, and if $\bar{\theta} < \theta_i$ then $\tilde{a}_i - \tilde{a}_{-i}$ is minimized at $(\delta_1, \delta_2, \cdots, -\delta_i, \cdots, \delta_N)$. As a result, ensuring that $\tilde{a}_i \lesseqgtr \bar{a}_i$ if $\theta_i \lesseqgtr \bar{\theta}$ for all $\mathbf{e} \in \mathcal{E}$ is reduced to ensuring

$$\begin{cases} \tilde{a}_{-i}(\mathbf{e}) - \tilde{a}_i(\mathbf{e})|_{\mathbf{e}=(-\delta_1,-\delta_2,\cdots,\delta_i,\cdots,-\mathcal{E}_N)} > 0 & \bar{\theta} > \theta_i \\ \tilde{a}_i(\mathbf{e}) - \tilde{a}_{-i}(\mathbf{e})|_{\mathbf{e}=(\delta_1,\delta_2,\cdots,-\delta_i,\cdots,\mathcal{E}_N)} > 0 & \bar{\theta} < \theta_i \end{cases}$$

This condition is further simplified to the following expression (thanks to the constant first derivatives).

Assumption (Generalized sufficient disagreement)**.**

$$(1-\kappa)\left|\theta_i - \bar{\theta}\right| > \frac{N-2}{N}\delta_i + \bar{\delta}, \forall i \in \mathcal{N}$$

Notice that this is *only* a sufficient condition for the following results, as it is generally too strong to require fixed partition for any power exertion—some are strictly Pareto dominated and should not matter for any optimization problems analyzed later.

A special case is $N = 2$, where this condition is also necessary. This can be seen by deriving a necessary condition requiring that the partition remains the same at the coordinator's optimal recommendation—necessary for the optimality of the full exertion of power as described in Lemma 1.8. This condition is equivalent to

$$\begin{cases} (1-\kappa)(\bar{\theta} - \theta_\ell) > \delta_\ell - \frac{\sum_{i\in\mathcal{J}}\delta_i - \sum_{j\in\not{\mathcal{J}}}\delta_j}{N} & \theta_\ell < \bar{\theta} \\ (1-\kappa)(\theta_i - \bar{\theta}) > \delta_\ell + \frac{\sum_{i\in\mathcal{J}}\delta_i - \sum_{j\in\not{\mathcal{J}}}\delta_j}{N} & \theta_\ell > \bar{\theta} \end{cases}$$

which reduces to the sufficient disagreement assumption if and only if $N = 2$.

C.1.2. Optimal recommendation and condition for coalition formation

We now prove Lemma 1.8, which implies that the coordinator optimally exerts full power over the members, the direction of which depends on the member's group affiliation. We repeat the result here.

Lemma.

$$\begin{cases} e_i^C = \delta_i & \theta_i < \bar{\theta} \\ e_i^C = -\delta_i & \theta_i > \bar{\theta} \end{cases}$$

Proof. The effect of power exertion on overall coordination is given by the following expression

$$\begin{aligned} & -\frac{\partial \sum_{\ell \in \mathcal{N}} (a_\ell - a_{-\ell})^2}{\partial e_i} \\ = & -\frac{\partial (\tilde{a}_i - \tilde{a}_{-i})^2 + \sum_{j \neq i} (\tilde{a}_j - \tilde{a}_{-j})^2}{\partial e_i} \\ = & \frac{2(\tilde{a}_{-i} - \tilde{a}_i)(N-1)(1-\kappa) - 2(1-\kappa) \sum_{j \neq i} (\tilde{a}_{-j} - \tilde{a}_j)}{(N-1+\kappa)(1-\kappa)} \\ = & \frac{2N(\tilde{a}_{-i} - \tilde{a}_i)}{N-1+\kappa} \end{aligned}$$

As the sufficient disagreement assumption implies, the sign of $\tilde{a}_{-i} - \tilde{a}_i$ must be the same as $\bar{\theta} - \theta_i$. Therefore the optimal recommendation must be the corner solution described in the lemma. □

As the coordinator optimally exerts full power over the members, they are held to their respective outside value. Consequently, a member i is better off than staying out if and only if a_{-i}^C is closer than a_{-i}^D to θ_i. This leads to Proposition 1.4, which we repeat here.

Proposition. *The N-member coalition is formed if and only if*

$$-(1-\kappa)\cdot\min\{\delta_j|\theta_j>\bar{\theta}\}\leq\sum_{i:\theta_i<\bar{\theta}}\delta_i-\sum_{j:\theta_j>\bar{\theta}}\delta_j\leq(1-\kappa)\cdot\min\{\delta_i|\theta_i<\bar{\theta}\}$$

Proof. Suppose that $\theta_\ell<\bar{\theta}$ for some member ℓ. First we show that $\theta_\ell<a^D_{-\ell}$ and $\theta_\ell<a^C_{-\ell}$, and then we derive the condition for $a^C_{-\ell}\leq a^D_{-\ell}$.

Recall that $a^D_\ell<a^D_{-\ell}$ and $a^C_\ell<a^C_{-\ell}$ given the sufficient disagreement assumption. As $a^D_\ell=a^{BR}_\ell(\mathbf{a}^D_{-\ell})$ and $a^C_\ell>a^{BR}_\ell(\mathbf{a}^C_{-\ell})$, it follows that $\theta_\ell<a^D_{-\ell}$ and $\theta_\ell<a^C_{-\ell}$. This is proven by contradiction, for $\theta_\ell\geq a_{-\ell}$ must imply $a^{BR}_\ell(\mathbf{a}_{-\ell})\geq a_{-\ell}$.

The condition for $a^C_{-\ell}\leq a^D_{-\ell}$ can be succinctly expressed thanks to the constant first-order derivative:

$$\begin{aligned}&\sum_{m\in\mathcal{N}}\frac{\partial\tilde{a}_{-\ell}}{\partial e_m}e^C_m\leq 0\\ \Leftrightarrow\quad&\sum_{i\in\mathcal{I}}\delta_i-\sum_{j\in\mathcal{J}}\delta_j\leq(1-\kappa)\delta_\ell\end{aligned}$$

It is straightforward to derive a similar condition for the case of $\theta_\ell>\bar{\theta}$:

$$\sum_{i\in\mathcal{I}}\delta_i-\sum_{j\in\mathcal{J}}\delta_j\geq-(1-\kappa)\delta_\ell$$

The proposition then follows when such conditions are satisfied for all members. □

C.1.3. Strong Pareto improvement and constrained efficient recommendations

Assume that the coalition cannot form without any remedy. To be consistent with Section 1.3, we assume without loss of generality that $\mathcal{J}$ is the weak group, i.e., who refuses to join must be some member ℓ with $\theta_\ell < \bar{\theta}$ and the set of such members $\mathcal{J}^b \equiv \{\ell | \theta_\ell < \bar{\theta}$ and $\kappa \cdot \delta_\ell > \sum_{j \in \mathcal{J}} \delta_j - \sum_{i \in \mathcal{I} \setminus \{\ell\}} \delta_i\}$.

We first adapt the definition of constrained efficient recommendations for $N \geq 2$ members.

Definition (Constrained efficient recommendation)**.** $\mathbf{e} \in \mathcal{E}^*$ if

(Participation constraint) $\tilde{u}_\ell(\mathbf{e}) \geq U_i^D$, for all $\ell \in \mathcal{N}$,

(Acceptance constraint) $\mathbf{e} \in \mathcal{E}$, and

(Pareto efficiency) $\nexists \mathbf{e}' \in \mathcal{E}$ such that $\tilde{u}_\ell(\mathbf{e}') \geqslant \tilde{u}_\ell(\mathbf{e})$ for all $\ell \in \mathcal{N}$ with $\tilde{u}_\ell(\mathbf{e}') > \tilde{u}_\ell(\mathbf{e})$ for some i.

We now derive the strong Pareto improvement that allows for a (partial) characterization of $\mathcal{E}^*$. Notice that the following specification of u_ℓ includes all the quadratic payoffs discussed in this paper as special case.

Lemma C.1. *If $\mathbf{e} \in \mathcal{E}$, $\theta_i < \bar{\theta}$, $\theta_j > \bar{\theta}$, $e_i < \delta_i$ and $e_j > -\delta_j$, then there exists $\mathbf{e}' \in \mathcal{E}$ where $e_i' < e_i$ and $e_j' > e_j$ such that $\tilde{u}_\ell(\mathbf{e}') > \tilde{u}_\ell(\mathbf{e})$ for any player ℓ whose payoff is specified as*

$$\text{(C.1.1)} \quad u_\ell(\mathbf{a}, s) = z \cdot \omega_\ell - \sum_{m \in \mathcal{N}} \mu_{\ell m}(a_m - a_{-m})^2 - \sum_{m \in \mathcal{N}} \nu_{\ell m}[(a_m - \theta_\ell)^2 + (N-1)(a_{-m} - \theta_\ell)^2]$$

where $\omega_\ell > 0, \mu_{\ell m} \geq 0, \nu_{\ell m} \geq 0, \sum_{m \in \mathcal{N}}(\mu_{\ell m} + \nu_{\ell m}) > 0$ and $\theta_\ell \in \mathbb{R}$.

Proof. We construct such an improved recommendation $\mathbf{e}'$ for all players by marginally increasing e_i and decreasing e_j in the same proportion. In the case of two-member model with $\kappa_1 \neq \kappa_2$, the proportion is instead $(1 + \kappa_1) : (1 + \kappa_2)$. The net change in $\tilde{a}_i$ and $\tilde{a}_j$ would then be identical (and non-zero) in magnitude with $\tilde{a}_i$ marginally increased and

$\tilde{a}_j$ marginally decreased. There is no change in $\tilde{a}_m$ for $m \neq i \neq j$. Consequently, $\tilde{a}_{-i}$ marginally decreases and $\tilde{a}_{-j}$ marginally decreases by a (weakly) smaller magnitude—strictly so if $N \geq 3$—with $\tilde{a}_{-m}$ not changed for $m \neq i \neq j$.

Given that $\tilde{a}_i < \tilde{a}_{-i}$ and $\tilde{a}_j > \tilde{a}_{-j}$ under either recommendation as guaranteed by the sufficient disagreement assumption, such a change strict decreases the term $\sum_{m\in\mathcal{N}} \mu_{\ell m}(a_m - \bar{a}_{-m})^2$. The change in the last term is slightly more complicated. Let ζ_m be an N-dimensional vector, each coordinate being 0 except for the mth coordinate being 1. Let $\tilde{a}_m$ and $\tilde{a}_{-m}$ be functions of $(\mathbf{e} + \varepsilon \cdot \zeta_i - \varepsilon \cdot \zeta_j)$. Then

$$\begin{aligned}
&\frac{\partial \sum_{m\in\mathcal{N}} \nu_{\ell m}[(\tilde{a}_m - \theta_\ell)^2 + (N-1)(\tilde{a}_{-m} - \theta_\ell)^2]}{\partial\varepsilon} \\
= \quad & \sum_{m\in\{i,j\}} \nu_{\ell m} \frac{\partial[(\tilde{a}_m - \theta_\ell)^2 + (N-1)(\tilde{a}_{-m} - \theta_\ell)^2]}{\partial\varepsilon} \\
= \quad & \sum_{m\in\{i,j\}} 2\nu_{\ell m}\left[(\tilde{a}_m - \theta_\ell)\left(\frac{\partial\tilde{a}_m}{\partial e_i} - \frac{\partial\tilde{a}_m}{\partial e_j}\right) + (N-1)(\tilde{a}_{-m} - \theta_\ell)\left(\frac{\partial\tilde{a}_{-m}}{\partial e_i} - \frac{\partial\tilde{a}_{-m}}{\partial e_j}\right)\right] \\
= \quad & \frac{2}{(N-1+\kappa)(1-\kappa)}\nu_{\ell i}\left[(\tilde{a}_i - \theta_\ell)(N-1)(1-\kappa) - (N-1)(\tilde{a}_{-i} - \theta_\ell)(1-\kappa)\right] \\
& - \frac{2}{(N-1+\kappa)(1-\kappa)}\nu_{\ell j}\left[(\tilde{a}_j - \theta_\ell)(N-1)(1-\kappa) - (N-1)(\tilde{a}_{-j} - \theta_\ell)(1-\kappa)\right] \\
= \quad & \frac{2(N-1)}{(N-1+\kappa)}\left[\nu_{\ell i}(\tilde{a}_i - \tilde{a}_{-i}) - \nu_{\ell j}(\tilde{a}_j - \tilde{a}_{-j})\right] \\
< \quad & 0
\end{aligned}$$

The inequality follows from $\tilde{a}_i < \tilde{a}_{-i}$ and $\tilde{a}_j > \tilde{a}_{-j}$. As a result, $\tilde{u}_\ell$ must strictly increase from $\mathbf{e}$ to $\mathbf{e}'$. □

This result implies that power must be fully exerted over at least one group in order for the recommendation not to be strictly Pareto dominated with respect to an arbitrary group of players whose preferences are as specified in Equation (C.1.1). With abuse of notation, let $\mathcal{E}^{\partial}$ be the set of such recommendations (which are all on the boundary of $\mathcal{E}$); obviously $\mathcal{E}^* \subseteq \mathcal{E}^{\partial}$.

We now show that power must be fully exerted over the strong group $\mathcal{J}$ if a recommendation in $\mathcal{E}^{\partial}$ can induce coalition formation. This is formalized in the following lemma, which extends to all recommendations in $\mathcal{E}^*$.

Lemma C.2. *For any $\mathbf{e}^{\partial} \in \mathcal{E}^{\partial}$ such that $\tilde{u}_i(\mathbf{e}^{\partial}) \geq U_i^D$ for all $i \in \mathcal{I}$, $e_j^{\partial} = -\delta_j$ for all $j \in \mathcal{J}$.*

Proof. If $e_j^{\partial} > -\delta_j$ for some $j \in \mathcal{J}$, the previous lemma implies that $e_i^{\partial} = \delta_i$ for all $i \in \mathcal{I}$. Consequently, $\tilde{a}_\ell(\mathbf{e}^{\partial}) > \tilde{a}_\ell(\mathbf{e}^C)$ for all $\ell \in \mathcal{N}$. Given that some member m with $\theta_m < \bar{\theta}$ receives $\tilde{u}_m(\mathbf{e}^C) < U_m^D$ as assumed, $\tilde{u}_m(\mathbf{e}^{\partial}) < \tilde{u}_m(\mathbf{e}^C) < U_m^D$ as member m is still held to his outside value, which becomes only worse as $\tilde{a}_{-m}$ increases from $a_{-m}^C > \theta_m$, as shown in the proof of Proposition 1.4. This is a contradiction and therefore $e_j^{\partial} = -\delta_j$ for all $j \in \mathcal{J}$. □

Corollary C.1. *For any $\mathbf{e}^* \in \mathcal{E}^*$, $e_j^* = -\delta_j$ for all $j \in \mathcal{J}$.*

As a result of Lemma C.2, any $\mathbf{e}^{\partial} \in \mathcal{E}^{\partial}$ must hold each member from the strong group $\mathcal{J}$ to his outside value. For them to be willing to participate, $\mathbf{e}^{\partial}$ must also exert sufficient power over the weak group in aggregate. This is formalized in the following lemma, which also extends to all recommendation in $\mathcal{E}^*$.

Lemma C.3. *For any $\mathbf{e}^{\partial} \in \mathcal{E}^{\partial}$ such that $\tilde{u}_\ell(\mathbf{e}^{\partial}) \geq U_\ell^D$ for all $\ell \in \mathcal{N}$, $\sum_{i \in \mathcal{I}} e_i^{\partial} \geq \sum_{j \in \mathcal{J}} \delta_j - (1-\kappa) \cdot \min_{j \in \mathcal{J}} \delta_j > 0$.*

Proof. Consider some member $\ell \in \mathcal{J}$. Because ℓ is held to his outside value, the condition for his participation is the same as in Proposition 1.4 but with $\mathbf{e}^C$ substituted with $\mathbf{e}^\partial$, i.e.,

$$\sum_{m \in \mathcal{N}} \frac{\partial \tilde{a}_{-\ell}}{\partial e_m} e_m^\partial \geq 0$$

$$\Leftrightarrow \quad \sum_{i \in \mathcal{I}} e_i^\partial - \sum_{j \in \mathcal{J}} \delta_j \geq -(1-\kappa)\delta_\ell$$

This condition is jointly satisfied for all members of the strong group $\mathcal{J}$ if and only if

$$\sum_{i \in \mathcal{I}} e_i^\partial \geq \sum_{j \in \mathcal{J}} \delta_j - (1-\kappa) \cdot \min_{j \in \mathcal{J}} \delta_j > 0$$

□

Corollary C.2. *For any* $\mathbf{e}^* \in \mathcal{E}^*$, $\sum_{i \in \mathcal{I}} e_i^* \geq \sum_{j \in \mathcal{J}} \delta_j - (1-\kappa) \cdot \min_{j \in \mathcal{J}} \delta_j > 0$.

In the two-member model where $\theta_1 < \theta_2$, the previous results are reduced to requiring $\mathbf{e}^*$ taking the form of $(e_1, -\delta_2)$ where $e_1 \geq \kappa_1 \delta_2$. Lemma 1.7 further characterizes $\mathcal{E}^*$ in this case, which we repeat here.

Lemma. $\mathcal{E}^* = \{(e_1, -\delta_2) | e_1 \in [\kappa_1 \delta_2, \bar{e}_1]\}$ *where* $\bar{e}_1 \in (\kappa_1 \delta_2, \delta_1)$ *is uniquely determined by* $\tilde{u}_1(\bar{e}_1, -\delta_2) = U_1^D$.

Proof. We show the result in three steps.

(1) Lemma C.1 establishes that all recommendations in $\mathcal{E}^*$ must take the form of $(e_1, -\delta_2)$.

(2) Among such recommendations, both members are guaranteed their respective disintegration payoff if and only if $e_1 \in [\kappa_1 \delta_2, \bar{e}_1]$.

First notice that $\tilde{u}_2(e_1, -\delta_2)$ strictly decreases in e_1 for all $e_1 \leq \delta_1$ (which we label as M1) and $\tilde{u}_1(\delta_1, e_2)$ decreases in e_2 for all $e_2 \geq -\delta_2$ (which we label as M2). In each case, the member i is held to his outside value, and the $\tilde{a}_j$ does not cross over to the other side of θ_i as can be proven by contradiction as in the proof of Proposition 1.4. Consequently, greater e_1 and lesser e_2 improves his outside value.

Lemma C.3 establishes that $\tilde{u}_2(e_1, -\delta_2) \geq U_2^D$ if $e_1 \geq \kappa_1\delta_2$. M1 then implies that $\tilde{u}_2(e_1, -\delta_2) \geq U_2^D$ only if $e_1 \geq \kappa_1\delta_2$.

In addition, there exists some $\bar{e}_1 \in (\kappa_1\delta_2, \delta_1)$ such that $\tilde{u}_1(e_1, -\delta_2) \geq U_1^D$ if and only if $e_1 \leq \bar{e}_1$.

Notice that $\tilde{u}_1(e_1, -\delta_2)$ strictly decreases in e_1 when $e_1 \geq \kappa_1\delta_2$ (which we label as M3). The sufficient disagreement assumption guarantees that $\tilde{a}_2(e_1, -\delta_2) > \theta_1$ whenever $e_1 \geq \kappa_1\delta_2$. Therefore $\tilde{u}_1(e_1, -\delta_2)$ must decrease in e_1 when $e_1 \geq \kappa_1\delta_2$ due to both the increasing distortion in member 1's action from his best response (i.e., increasing e_1^2) and the increasing $\tilde{a}_2(e_1, -\delta_2)$ which decreases the outside value.

What follow are the existence and the uniqueness of $\bar{e}_1 \in (\kappa_1\delta_2, \delta_1)$ such that $\tilde{u}_1(e_1, -\delta_2) = U_1^D$. Because $\tilde{u}_1(\delta_1, -\delta_2) < U_1^D$ and $\tilde{u}_1(\kappa_2\delta_1, -\delta_2) = u_1(a_1^D, a_2^{BR}(a_1^D) - \delta_2, 1) > u_1(a_1^D, a_2^D, 1) > U_1^D$, the intermediate value theorem and the strict monotonicity imply the result.

(3) There exists no acceptable recommendation that Pareto dominates $(e_1, -\delta_2)$ for all $e_1 \in [\kappa_1\delta_2, \bar{e}_1]$.

Suppose the contrary, i.e., some $(e_1', e_2') \in \mathcal{E}$ Pareto dominates some $(e_1, -\delta_2)$ where $e_1 \in [\kappa_1\delta_2, \bar{e}_1]$. Apply the Pareto improvement from Lemma C.1 to find (e_1'', e_2'') on the boundary of $\mathcal{E}$, which must also Pareto dominate $(e_1, -\delta_2)$. Then for some member i, $\tilde{u}_i(e_1'', e_2'') > \tilde{u}_i(e_1, -\delta_2)$.

If $i = 1$, then the previous mononicity results M2 and M3 imply that (e_1'', e_2'') must take the form $(e_1'', -\delta_2)$ where $e_1'' < e_1$. However, M1 then implies $\tilde{u}_2(e_1'', -\delta_2) < \tilde{u}_2(e_1, -\delta_2)$, which contradicts Pareto domination.

If $i = 2$, then the previous mononicity result M1 implies that (e_1'', e_2'') takes the form $(e_1'', -\delta_2)$ where $e_1'' > e_1$, or the form (δ_1, e_2''). In either case, M3 and M2 imply $\tilde{u}_1(e_1'', e_2'') < \tilde{u}_2(e_1, -\delta_2)$, which contradicts Pareto domination. □

We now show a key result about efficient remedies, namely the set of recommendations $\mathbf{e}^\partial \in \mathcal{E}^\partial$ that induces coalition formation is the set of constrained efficient recommendations. This result is later used to establish the constrained Pareto efficiency of the remedies.

Lemma C.4. $\mathcal{E}^* = \{\mathbf{e}^\partial \in \mathcal{E}^\partial | \tilde{u}_\ell(\mathbf{e}^\partial) \geq U_\ell^D, \forall \ell \in \mathcal{N}\}$.

Proof. The direction of $\mathcal{E}^* \subseteq \{\mathbf{e}^\partial \in \mathcal{E}^\partial | \tilde{u}_\ell(\mathbf{e}^\partial) \geq U_\ell^D, \forall \ell \in \mathcal{N}\}$ is a given, since $\mathbf{e}^\partial \in \mathcal{E}^\partial$ is weaker than the Pareto efficiency condition in the definition of $\mathcal{E}^*$.

Before we show the other direction, first observe the marginal effect of a directional change in power exertion. For any $\ell, m \in \mathcal{I}$, if e_ℓ is marginally increased and e_m decreased by the same magnitude, then $\tilde{u}_\ell$ increases and $\tilde{u}_m$ decreases while all other members' payoffs stay the same. Formally, let$\tilde{u}_\ell$, $\tilde{a}_\ell$ and $\tilde{a}_{-\ell}$ be functions of $(\mathbf{e} + \varepsilon \cdot \zeta_\ell - \varepsilon \cdot \zeta_m)$. Then

$$\begin{aligned}
&\frac{\partial \tilde{u}_\ell}{\partial \varepsilon}\\
=&\frac{\partial - \kappa(\tilde{a}_\ell - \tilde{a}_{-\ell})^2 - (1-\kappa)[(\tilde{a}_\ell - \theta_\ell)^2 + (N-1)(\tilde{a}_{-\ell} - \theta_\ell)^2]}{\partial \varepsilon}\\
=&-2\kappa(\tilde{a}_\ell - \tilde{a}_{-\ell})\left(\frac{\partial \tilde{a}_\ell}{\partial e_\ell} - \frac{\partial \tilde{a}_\ell}{\partial e_m} - \frac{\partial \tilde{a}_{-\ell}}{\partial e_\ell} + \frac{\partial \tilde{a}_{-\ell}}{\partial e_m}\right)\\
&-2(1-\kappa)\left[(\tilde{a}_\ell - \theta_\ell)\left(\frac{\partial \tilde{a}_\ell}{\partial e_\ell} - \frac{\partial \tilde{a}_\ell}{\partial e_m}\right) + (N-1)(\tilde{a}_{-\ell} - \theta_\ell)\left(\frac{\partial \tilde{a}_{-\ell}}{\partial e_\ell} - \frac{\partial \tilde{a}_{-\ell}}{\partial e_m}\right)\right]\\
=&-\frac{2\kappa N(\tilde{a}_\ell - \tilde{a}_{-\ell})}{N-1+\kappa} - \frac{2(N-1)(1-\kappa)}{N-1+\kappa}[(\tilde{a}_\ell - \theta_\ell) - (\tilde{a}_{-\ell} - \theta_\ell)]\\
=&-2(\tilde{a}_\ell - \tilde{a}_{-\ell})\\
>&0
\end{aligned}$$

and similarly, $\frac{\partial \tilde{u}_m}{\partial \varepsilon} < 0$ and $\frac{\partial \tilde{u}_n}{\partial \varepsilon} = 0, \forall n \neq \ell \neq m$ as neither $\tilde{a}_n$ nor $\tilde{a}_{-n}$ changes.

Now we show the direction of $\{\mathbf{e}^\partial \in \mathcal{E}^\partial | \tilde{u}_\ell(\mathbf{e}^\partial) \geq U_\ell^D, \forall \ell \in \mathcal{N}\} \subseteq \mathcal{E}^*$ by contradiction. Suppose $\mathbf{e}^\partial$ is Pareto dominated by some $\mathbf{e} \in \mathcal{E}$, then Lemma C.1 implies that we can find some $\mathbf{e}^* \in \mathcal{E}^*$ that Pareto dominates $\mathbf{e}^\partial$. Let $\mathcal{M}^*$ be the nonempty set of members for whom the Pareto dominance is strict.

There are five cases.

(1) If $\mathcal{J} \subseteq \mathcal{M}^*$, then $\sum_{i \in \mathcal{J}} e_i^* > \sum_{i \in \mathcal{J}} e_i^\partial$.

   (a) If $|\mathcal{J}| = 1$ then $e_i^* > e_i^\partial$ and consequently $\tilde{u}_i(\mathbf{e}^*) < \tilde{u}_i(\mathbf{e}^\partial)$ because $e_i^\partial > 0$ as shown previously. Contradiction is found.

   (b) Otherwise, $|\mathcal{J}| \geq 2$.

(i) If no member $i \in \mathcal{J}$ exists such that $e_i^* < e_i^\partial$, then because $\sum_{i \in \mathcal{J}} e_i^\partial > 0$, some $\ell \in \mathcal{J}$ must exists such that $e_\ell^\partial > 0$. Then his outside value decreases as $\mathbf{e}^\partial$ becomes $\mathbf{e}^*$, and distortion in his action weakly increases. Consequently, $\tilde{u}_\ell(\mathbf{e}^*) < \tilde{u}_\ell(\mathbf{e}^\partial)$. Contradiction is found.

(ii) Otherwise, there are two disjoint nonempty sets $\mathcal{J}^+ \equiv \{i \in \mathcal{J} | e_i^* > e_i^\partial\}$ and $\mathcal{J}^- \equiv \{i \in \mathcal{J} | e_i^* < e_i^\partial\}$. The change in power exertion over $\mathcal{J}^+$ is greater in magnitude than the change over $\mathcal{J}^-$. We can thus decompose the change in the payoff of any member $m \in \mathcal{J}^-$ from $\mathbf{e}^\partial$ to $\mathbf{e}^*$ in three parts. With abuse of notation, we describe how to apply the previous observation of directional marginal effect of power exertion. First, match the change from e_m^∂ to e_m^* with the change of the same magnitude from e_ℓ^∂ to e_ℓ^*, where ℓ is an indefinite member from $\mathcal{J}^+$; as e_ℓ^∂ reaches e_ℓ^*, we switch to a different member from $\mathcal{J}^+$ as ℓ. The observation regarding marginal change in power exertion implies that m is strictly worse off. Second, match the change in all other members from $\mathcal{J}^-$ with the change in members from $\mathcal{J}^+$. The same observation implies that m's payoff is constant. Third, there is residual change in members from $\mathcal{J}^+$, which makes m worse off as his outside value decreases while distortion in his action remains constant. Consequently, $\tilde{u}_m(\mathbf{e}^*) < \tilde{u}_m(\mathbf{e}^\partial)$. Contradiction is found.

(2) If $\mathcal{J} \cap \mathcal{M}^* = \emptyset$, then $\mathcal{J} \cap \mathcal{M}^* \neq \emptyset$ and $\sum_{i \in \mathcal{J}} e_i^* = \sum_{i \in \mathcal{J}} e_i^\partial$.

(a) If $|\mathcal{J}| = 1$ then this is impossible, for $e_i^* = e_i^\partial$, which implies that $\tilde{u}_\ell(\mathbf{e}^*) = \tilde{u}_\ell(\mathbf{e}^\partial)$. Therefore $\mathcal{J} \cap \mathcal{M}^* = \emptyset$ and contradiction is found.

(b) Otherwise, $|\mathcal{J}| \geq 2$. Since $\mathbf{e}^* \neq \mathbf{e}^\partial$, there are two disjoint nonempty sets $\mathcal{J}^+ \equiv \{i \in \mathcal{J} | e_i^* > e_i^\partial\}$ and $\mathcal{J}^- \equiv \{i \in \mathcal{J} | e_i^* < e_i^\partial\}$, and the change in power exertion over $\mathcal{J}^+$ is the same in magnitude as over $\mathcal{J}^-$. We can thus decompose the change in the payoff of any member $m \in \mathcal{J}^-$ from $\mathbf{e}^\partial$ to $\mathbf{e}^*$ in two parts, i.e., the first two parts of Case 1(b)ii. Similarly, $\tilde{u}_m(\mathbf{e}^*) < \tilde{u}_m(\mathbf{e}^\partial)$. Contradiction is found.

□

C.1.4. Delegated coordination

In this section, we characterize the conditions under which delegated coordination allows the coalition to form. Suppose that the delegate's preference $u_d(\mathbf{a}, s)$ takes the form as specified in Equation (C.1.1); it accommodates the general preference in the two-member model as specified in Equation (1.3.1) as special case. Furthermore, it accommodates any convex combination of members' and the coordinator's preferences.

We focus on the case where d represents a convex combination of members' and the coordinator's preferences, with weights $\gamma \equiv (\gamma_C, \gamma_1, \gamma_2, \cdots, \gamma_N)$. We show that if her recommendation $\mathbf{e}^d$ allows the coalition to form, then the remedy is efficient, and the interest of both groups must be represented.

Lemma C.5. *If $\tilde{u}_\ell(\mathbf{e}^d) \geq U_\ell^D$ for all $\ell \in \mathcal{N}$ then $\mathbf{e}^d \in \mathcal{E}^*$ and $0 < \sum_{i \in \mathcal{J}} \gamma_i < 1$.*

Proof. Lemma C.1 implies that $\mathbf{e}^d \in \mathcal{E}^\partial$ (or it can be strictly improved for the delegate). The constrained efficiency is then established in Lemma C.4.

We now show that $\sum_{i\in\mathcal{I}}\gamma_i > 0$. If $\sum_{i\in\mathcal{I}}\gamma_i = 0$, i.e., only the preferences of $\mathcal{J}^d \subseteq \mathcal{J}\cup\{C\}$ are represented in $u_d(\mathbf{a}, s)$, then it is in the interest of $\mathcal{J}^d$ to recommend $e_i^d = \delta_i, \forall i \in \mathcal{I}$ to improve their outside values (and coordination for C). Furthermore, $e_j^d \geq 0, \forall j \in \mathcal{J}$. For $j \in \mathcal{J}\backslash\mathcal{J}^d$, $e_j^d = \delta_j$ for the same reason; for $j \in \mathcal{J}^d$, if $e_j^d < 0$, then it is a strict improvement of $\mathcal{J}^d$ to recommend $e_j^d = 0$, for it improves their outside values (and coordination for C) and reduces distortion in member j's action. Such recommendation with $e_i^d = \delta_i, \forall i \in \mathcal{I}$ and $e_j^d \geq 0, \forall j \in \mathcal{J}$ contradicts the requirement for coalition to form as previously argued, and therefore $\sum_{i\in\mathcal{I}}\gamma_i > 0$.

We now show that $\sum_{i\in\mathcal{J}}\gamma_i < 1$. If $\sum_{i\in\mathcal{J}}\gamma_i = 1$, i.e., only the preferences of $\mathcal{I}^d \subseteq \mathcal{I}\cup\{C\}$ are represented in $u_d(\mathbf{a}, s)$, then $e_i^d \leq 0, \forall i \in \mathcal{I}$ and $e_j^d = -\delta_j, \forall j \in \mathcal{J}$. This is shown in the symmetric way as the previous argument. Lemma C.3 then shows that, for all $j \in \mathcal{J}$ to participate, $\sum_{i\in\mathcal{I}} e_i^d > 0$ must be satisfied. Yet $\sum_{i\in\mathcal{I}} e_i^d \leq 0$, and therefore $\sum_{i\in\mathcal{J}}\gamma_i < 1$. □

For two-member model, the existence of such weights is shown with the intermediate value theorem; the existence result complement the previous lemma to produce Proposition 1.2. Let the delegate's optimal recommendation takes the form of $(e_1^\gamma, -\delta_2)$ where $\gamma \equiv \gamma_1$ indexes the recommendation, i.e.,

$$e_1^\gamma = \underset{e_1\in[-\delta_1,\delta_1]}{\operatorname{argmax}}\ \tilde{u}_\gamma(e_1, -\delta_2)$$

It is straightforward to show that $\tilde{u}_\gamma(e_1, -\delta_2)$ is strictly concave in e_1, and therefore the maximizer must be unique. In addition, because $[-\delta_1, \delta_1]$ is closed and $\tilde{u}_\gamma(e_1, -\delta_2)$ is continuous in e_1 and γ, e_1^γ must be continuous in γ.

Notice the two extreme cases: if $\gamma = 0$, then $e_1^0 = \delta_1$ as shown in the baseline model; if $\gamma = 1$, then $e_1^1 < 0$ as shown in the previous proof. The intermediate value theorem then implies that some $\hat{\gamma} \in (0, 1)$ must exist so that $e_1^{\hat{\gamma}} = \kappa_1\delta_2 \in (0, \delta_1)$, and consequently $(e_1^{\hat{\gamma}}, -\delta_2)$ is constrained efficient.

C.1.5. Member approval

In this section, we first show that under any non-degenerate approval rule, the coordinator optimally makes an acceptable recommendation that gains necessary approval. We then show that the coordinator's optimal approval rule must be non-degenerate. We then show that if any such rule must constitute an efficient remedy if it allows the coalition to form. This result implies the superiority of *consensus*, i.e., giving each member a veto that triggers an inefficient remedy is Pareto efficient and constitutes a Pareto improvement on the remedy. We further show that some members' veto must be inconsequential in this case.

We first generalize the monotonicity condition which continues to be imposed on approval rules that the coordinator can choose from.

Condition (Monotonicity)**.** For any $i \in \mathcal{N}$ and $\mathbf{m}_{-i} \in \{0, 1\}^{N-1}$, $r(m_i = 1, \mathbf{m}_{-i}) = 1$ if $r(m_i = 0, \mathbf{m}_{-i}) = 1$.

Notice that there continues to be two degenerate approval rules, i.e., $r(\cdot) = 1$ and $r(\cdot) = 0$. Among non-degenerate approval rules, we refer to a rule as giving member i a *veto* if $r(m_i = 0, \mathbf{m}_{-i}) = 0$ for any $\mathbf{m}_{-i} \in \{0, 1\}^{N-1}$. Not every rule gives anyone a veto, and a member has a veto in multiple rules just as in the two-member model.

As argued for the two-member model, we assume that each member sends his weakly dominating message:

$$\tilde{m}_i(\mathbf{e}^P, U_i^{\emptyset}) = \begin{cases} 1 & \tilde{u}_i(\mathbf{e}^P) \geq U_i^{\emptyset} \\ 0 & \tilde{u}_i(\mathbf{e}^P) < U_i^{\emptyset} \end{cases}$$

For any $\alpha^{\emptyset} \in \Delta\mathbb{R}^N$ (mixed) fallback recommendation, we construct an acceptable recommendation $\bar{\mathbf{a}} \in \mathbb{R}^N$ and show that it Pareto dominates $\alpha^{\emptyset}$. Consequently, we show that for any (fixed) non-degenerate approval rule it is optimal for the coordinator to choose $\mathbf{e}^P$ from acceptable recommendations that can gain members' approval, and consequently $\mathbf{e}^R = \mathbf{e}^P$ is induced in equilibrium.

Lemma C.6. *Under any non-degerate approval rule r, the coordinator optimally recommends some $\mathbf{e}^P \in \mathcal{E}$ such that $r \circ \tilde{\mathbf{m}}(\mathbf{e}^P, \mathbf{U}^{\emptyset}) = 1$.*

Proof. We construct $\bar{\mathbf{a}}$ by replacing any unacceptable recommendation in the support of $\alpha^{\emptyset}$ with $\mathbf{a}^D$, taking the expection of this new distribution, and applying the same *strong* Pareto improvement from Lemma C.1.

Notice that $\bar{\mathbf{a}} \in \mathcal{A} \equiv \{\tilde{\mathbf{a}}(\mathbf{e}) | \mathbf{e} \in \mathcal{E}\}$, i.e., it is acceptable, because $\mathcal{A}$ is convex so the expectation and the *strong* Pareto improvement both preserve this property. In addition, the concavity of members' and the coordinator's payoff functions and the *strong* Pareto improvement ensure that $u_i(\bar{\mathbf{a}}, 1) \geq U_i^{\emptyset}$ for any $i \in \mathcal{N} \cup \{C\}$, strictly so for all players if $\alpha^{\emptyset}$ is strictly mixed because of the concave payoff functions, or if $\alpha^{\emptyset}$ is pure but not constrained efficient because of the *strong* Pareto improvement.

In addition, the coordinator cannot optimally make an unacceptable recommendation either: even if it gets approved and induces $(\mathbf{a}^D, 0)$, the coordinator can still do better by recommending $\mathbf{a}^D$, which must be approved as well and induce $(\mathbf{a}^D, 1)$. As a result, the coordinator's optimal recommendation must be acceptable and gets approvd under r. □

This result implies that neither of the degenerate rules can be uniquely optimal as argued in Section 1.3.3. The automatic approval prevents the coalition from forming because the coordinator optimally recommends $\mathbf{a}^C$. The automatic rejection is worse than giving each member a veto, as the coordinator can at least recommend $\bar{\mathbf{a}}$ under the latter rule, which must be approved as each member is guaranteed at least his fallback payoffs.

Notice that the coordinator must receive weakly more than its fallback payoff, strictly more if $\alpha^\emptyset$ is strictly mixed, or pure but not constrained efficient. This result therefore establishes that augmenting any inefficient remedy with consensus decision-making constitutes a Pareto improvement as argued in Section 1.3.3.1.

The Pareto efficiency of consensus is implied by the following result: if coalition is formed under some nondegenerate approval rule, then the equilibrium recommendation $\mathbf{e}^R = \mathbf{e}^P$ must be constrained efficient.

Lemma C.7. *Fix some nondegenerate approval rule r and some fallback recommendation $\alpha^\emptyset$, and let $\mathbf{e}^P$ be the equilibrium recommendation. If $\tilde{u}_\ell(\mathbf{e}^P) \geq U_\ell^D$ for all $\ell \in \mathcal{N}$ then $\mathbf{e}^P \in \mathcal{E}^*$.*

Proof. Lemma C.1 implies that $\mathbf{e}^P \in \mathcal{E}^\partial$ (or it can be strictly improved for the coordinator). The constrained efficiency is then established in Lemma C.4. □

We now show that under consensus, some members' veto must be inconsequential as the constraint regarding his payoff is not binding. As a special case, the strong member is always the one in the two-member model.

Lemma C.8. *If the coalition forms when each member has a veto, there must be some member whose constraint is not binding.*

Proof. Given that the optimal recommendation $\mathbf{e}^P \in \mathcal{E}^*$ satisfies $e_i^P \leq 0, \forall i \in \mathcal{I}$ and $e_j^P = -\delta_j, \forall j \in \mathcal{J}$, there are a total of $|\mathcal{I}|$ "free" variables $(e_i)_{i \in \mathcal{I}}$ with $N > |\mathcal{I}|$ constraints. Consequently, some constraints must not be bindings. Generically, we can identify (some of) those members, and they are from the strong group.

(1) Suppose $|\mathcal{I}| = 1$. Then the weak member i's constraint is binding but no member from the strong group can have his constraint binding. Recall that $\mathbf{e}^P \in \mathcal{E}^*$ implies that $e_j^P = -\delta_j, \forall j \in \mathcal{J}$. Yet if i's constraint is not binding, $\mathbf{e}^P$ must be unchanged if his constraint disappears. This implies that the coordinator's recommendation optimally satisfies $e_i^P = \delta_i$. Consequently $\mathbf{e}^P \notin \mathcal{E}^*$.

Suppose some member $\ell \in \mathcal{J}$ has his constraint binding. This implies that his equilibrium payoff is strictly lowered if his constraint disappears. Let the new recommendation be $\mathbf{e}' \in \mathcal{E}^{\partial}$, and $\mathbf{e}'$ continues to satisfy $e_j' = -\delta_j, \forall j \in \mathcal{J}$ given Lemma C.2, while $\tilde{u}_\ell(\mathbf{e}') < \tilde{u}_\ell(\mathbf{e}^P)$. Consequently, $e_i' < e_i^P$ and $\tilde{u}_i(\mathbf{e}') > \tilde{u}_i(\mathbf{e}^P)$. This contradicts the optimality of $\mathbf{e}'$ as an increase in e_i' must not affect i's constraint while improving all other players' payoffs (including the coordinator's). Therefore no member from the strong group can have his constraint binding.

(2) Suppose $|\mathcal{J}| \geq 2$. Recall that $\mathbf{e}^P \in \mathcal{E}^*$ implies that $e_j^P = -\delta_j, \forall j \in \mathcal{J}$. Therefore any member $m \in \mathcal{J}$ has his constraint binding if and only if $U_m^O(\tilde{a}_{-m}(\mathbf{e}^P)) = U_m^\emptyset$, or equivalently

$$\sum_{i \in \mathcal{J}} e_i^P - \sum_{j \in \mathcal{J}} \delta_j = (N - 1 + \kappa)(1 - \kappa)[U_m^{-1}(U_m^\emptyset) - a_{-m}^D] - (\kappa - 1)\delta_m$$

where U_m^{-1} is the inverse function of U_m^O. Notice that this equation generically cannot be satisfied for all $m \in \mathcal{J}$, and in this case only one member's constraint is binding.

□

We now prove Proposition 1.3. Part of the results are shown as special case of previous lemmas. We repeat the proposition here.

Proposition. *It is optimal for the coordinator to give the weak member a veto; doing so induces the coalition to form if and only if* $U_1^\emptyset \in [U_1^D, \tilde{u}_1(\kappa_1\delta_2, -\delta_2)]$.

Proof. Lemma C.8 has shown that if some $(e_1^P, e_2^P) \in \mathcal{E}^*$ is induced in equilibrium when each member has a veto, then the same recommendation is induced when only the weak member has the veto.

In the case that no member has a veto, $(e_1^C, e_2^C) = (\delta_1, -\delta_2)$ must yield at least one member at least his fallback payoff and is therefore the equilibrium recommendation; otherwise we can construct $(\bar{a}_1, \bar{a}_2) \in \mathcal{E}^\partial$ from $(\delta_1, -\delta_2)$, which must strictly Pareto dominate $(\delta_1, -\delta_2)$. However, the monotonicity results M1 and M2 used in the proof of Lemma 1.7 imply a contradiction.

If only the strong member has a veto is binding, then $\tilde{u}_2(e_1^P, e_2^P) \geq \tilde{u}_2(e_1^C, e_2^C)$. Consequently (e_1^P, e_2^P) takes the form of (δ_1, e_2), implying that $\tilde{u}_1(e_1^P, e_2^P) \leq \tilde{u}_1(e_1^C, e_2^C) < U_1^D$, so the coalition cannot form.

Recall that the degenerate approval rules are always suboptimal. Consequently the previous results has established the optimality of giving the weak member a veto.

We now show that when the weak member has a veto, the coalition can form if and only if $U_1^{\emptyset} \in [U_1^D, \tilde{u}_1(\kappa_1\delta_2, -\delta_2)]$. We first show necessity: if the coalition can form, the constraint from the weak member's veto must be binding and $(e_1^P, e_2^P) \in \mathcal{E}^*$. Therefore (e_1^P, e_2^P) takes the form of $(e_1, -\delta_1)$ where $e_1 \in [\kappa_1\delta_2, \bar{e}_1]$ and $\tilde{u}_1(e_1, -\delta_2) = U_1^{\emptyset}$. Consequently, $U_1^{\emptyset} \in [U_1^D, \tilde{u}_1(\kappa_1\delta_2, -\delta_2)]$. As for sufficiency, notice that the coordinator's program is equivalent to

$$\max_{(e_1^P, e_2^P) \in \mathcal{E}} \tilde{u}_L(e_1^P, e_2^P)$$

$$\text{subject to } \tilde{u}_1(e_1^P, e_2^P) \geq U_1^{\emptyset}$$

When $U_1^{\emptyset} \in [U_1^D, \tilde{u}_1(\kappa_1\delta_2, -\delta_2)]$, as $U_1^D > \tilde{u}_1(\delta_1, -\delta_2)$, the constraint must be binding and the monotonicity result M1 implies that the optimal recommendation (e_1^P, e_2^P) must take the form of $(e_1, -\delta_1)$. Because $\tilde{u}_L(e_1, -\delta_2)$ increases in e_1 for all $e_1 \in [-\delta_1, \delta_1]$, the coordinator's program is equivalent to

$$\max e_1 \in [-\delta_1, \delta_1]$$

$$\text{subject to } \tilde{u}_1(e_1, -\delta_2) = U_1^{\emptyset}$$

It is straightforward to show that the solution must fall within $[\kappa_1\delta_2, \bar{e}_1]$: any $e_1 > \bar{e}_1$ must yield $\tilde{u}_1 < U_1^D \leq U_1^{\emptyset}$, whereas if any $e_1 < \kappa_1\delta_2$ yields $\tilde{u}_1 \in [U_1^D, \tilde{u}_1(\kappa_1\delta_2, -\delta_2)]$ then there must exist a higher $e_1' \in [\kappa_1\delta_2, \bar{e}_1]$ that yields the same payoff. □

C.2. Performance Manipulation and Incentive Design

C.2.1. Baseline model

Let

$$\Pr[\kappa = H] = p(e, \phi) + q(m, \phi)$$

where $p_e \geq 0, p_\phi \geq 0, p_{e\phi} > 0, q_m \geq 0, q_\phi \leq 0$ and $q_{m\phi} < 0$ and they are all bounded away from infinity. Let p be multiplicatively separable in (e, ϕ), i.e., $p(e, \phi) = f(e) \cdot g(\phi)$. Recall that

$$U^P = u(e) - h(m) - t$$

$$U^A = t - c(e)$$

where u, h and c are weakly positive and strictly increasing; $c_e(0) = 0$ and $c_{ee} > 0$.

C.2.1.1. Principal's optimal effort inducement. Before we prove Lemma 1, we first show some preliminary results.

The principal sets $t(\kappa = L) = 0$ as previously argued. With abuse of notation, let $t = t(\kappa = H)$. The agent's *ex post* program is

$$\max_{e \in [0,1]} \; t(p + q) - c$$

And the first-order condition characterizes the interior optimal effort $e^A(t,\phi)$:

$$tp_e(e^A) - c_e(e^A) = 0 \tag{C.2.1}$$

Then $t^P(e,\phi) = \frac{c_e(e)}{p_e(e,\phi)}$, the bonus level needed to induce effort level e. Consequently, $t^P_\phi = -\frac{p_{e\phi}c_e}{p_e^2} \leq 0$, strictly for any interior effort level, and $t^P_e = \frac{c_{ee}p_e - p_{ee}c_e}{p_e^2} > 0$. In addition, $t^P_{e\phi} < 0$ from p's multiplicative separability in (e,ϕ).

Notice that

(1) $t^P_e > 0$. More effort requires higher bonus.

(2) $t^P_\phi < 0$. Better alignment reduces the bonus level needed to induce any level of effort.

(3) $t^P_{e\phi} < 0$. Better alignment also reduces the marginal bonus needed to induce effort.[1]

The principal's interim program is

$$\max_{t \geq 0} \; u(e^A(t,\phi)) - h(m) - \mathbb{E}[t|e^A(t,\phi), m, \phi]$$

Given the strict monotonicity of t^P in e, we can rewrite the program as

$$\max_{e\in[0,1]} \; u(e) - h(m) - t^P \cdot [p(e,\phi) + q(m,\phi)]$$

[1] More generally, this result follows the logsupermodularity (as implied by multiplicative separability) of $\Pr[\kappa = H]$ in (e,ϕ), a sufficient but not necessary condition.

which we will assume to be concave in e (which can be guaranteed by c_e being weakly convex; see Jewitt et al. (2008) for proof) and whose solution is interior. Denote the solution to this partial program as $e^P(m, \phi)$, determined by the first-order condition

$$u_e - t^P p_e - \underbrace{t_e^P \cdot (p+q)}_{\text{motivational rent}} = 0 \tag{C.2.2}$$

We first show the following lemma that performance measure alignment has no effect on bonus payment towards effort (or on the marginal bonus payment, or the marginal of the marginal, etc.) at any level of effort inducement. This would greatly reduce the number of effects to consider throughout the paper.

Lemma C.9. *Let* $t_k^P \equiv \frac{\partial^k t^P(e,\phi)}{\partial e^k}$ *and* $p_\ell \equiv \frac{\partial^\ell p(e,\phi)}{\partial e^\ell}$ *for any* $(k, \ell) \in \mathbb{N}^2$. *Specifically,* $t_0^P \equiv t^P$ *and* $p_0 \equiv p$. *Then* $\frac{\partial t_k^P \cdot p_\ell}{\partial \phi} = 0$.

Proof. Given the multiplicative separability of p in (e, ϕ), we can write $p(e, \phi) = f(e) \cdot g(\phi)$. The decomposition is not necessarily unique; we use an arbitrary one.

Notice that

$$\begin{aligned} t_k^P &\equiv \frac{\partial^k t^P(e,\phi)}{\partial e^k} \\ &= \frac{\partial^k}{\partial e^k} \frac{c_e(e)}{p_e(e,\phi)} \\ &= \frac{\partial^k}{\partial e^k} \frac{c_e(e)}{f_e(e) \cdot g(\phi)} \\ &= \left(\frac{\partial^k}{\partial e^k} \frac{c_e(e)}{f_e(e)} \right) \cdot g(\phi)^{-1} \end{aligned}$$

and

$$\begin{aligned} p_\ell &\equiv \frac{\partial^\ell p(e,\phi)}{\partial e^\ell} \\ &= \frac{\partial^\ell}{\partial e^\ell}\left(f(e)\cdot g(\phi)\right) \\ &= \left(\frac{\partial^\ell}{\partial e^\ell} f(e)\right)\cdot g(\phi) \end{aligned}$$

Therefore $t_k^P \cdot p_\ell = \frac{\partial^k}{\partial e^k}\frac{c_e(e)}{f_e(e)} \cdot \frac{\partial^\ell}{\partial e^\ell} f(e)$, independent of ϕ. Hence the result. □

The next corollary is a special case of this lemma, and is useful for showing in later results.

Corollary C.3. *Fix any level of effort inducement. The bonus paid towards effort is independent from performance measure alignment, i.e.*

$$\frac{\partial t^P(e,\cdot)p(e,\cdot)}{\partial \phi} = 0$$

Notice that manipulation increases motivational rent while performance measure alignment decreases it. We would therefore expect both the principal's effort inducement and her value to move in the opposite direction. Hence the following lemma.

We repeat Lemma 2.1 here.

Lemma. *Principal's optimal effort inducement and interim value both decrease in manipulation m and increase in performance measure alignment ϕ.*

Proof. Let $\mathcal{S} \equiv \mathcal{S}(e, m, \phi) < 0$ (by assumption) be the second derivative of U_P w.r.t. e.

The principal's equilibrium effort inducement decreases in manipulation,

$$e^P_m = \frac{t^P_e q_m}{\mathcal{S}(e^P, \cdot, \cdot)} < 0$$

and increases in performance measure alignment (invoking Lemma C.9 twice),

$$e^P_\phi = \frac{t^P_{e\phi} q + t^P_e q_\phi}{\mathcal{S}(e^P, \cdot, \cdot)} > 0$$

The envelope theorem implies that the principal's value decreases with manipulation,

$$\begin{aligned} V^P_m &= -h_m - t^P q_m \\ &< 0 \end{aligned} \tag{C.2.3}$$

and together with Corollary C.3, that the principal's value increases with performance measure alignment.

$$\begin{aligned} V^P_\phi &= -\frac{\partial t^P p}{\partial \phi} - \frac{\partial t^P q}{\partial \phi} \\ &= -t^P q_\phi - t^P_\phi q \\ &> 0 \end{aligned}$$

□

We repeat Proposition 2.1 here.[2]

Proposition. *The optimal bonus is lower than the "second-best" benchmark.*

Proof. $t^P(e^P(0,\phi),\phi) > t^P(e^P(m,\phi),\phi)$ where $m > 0$ follows from $\frac{\partial t^P(e^P(m,\phi),\phi)}{\partial m} = t^P_e \cdot e^P_m < 0$. □

[2] This proposition does not rely on any restriction on the functional form of $p(e, \phi)$.

C.2.1.2. Agent's optimal manipulation. The agent's *ex ante* program is

$$\max_{m\in[0,1]} V^A(m,\phi) \equiv t^P \cdot (p+q) - c(e^P)$$

where $e^P \equiv e^P(m,\phi)$, $t^P \equiv t^P(e^P,\phi)$, $p \equiv p(e^P,\phi)$ and $q \equiv q(m,\phi)$. We will assume that this program is concave in m and its solution is interior. The *ex ante* first-order condition is

$$t^P \cdot (p_e e^P_m + q_m) + t^P_e e^P_m \cdot (p+q) - c_e e^P_m = 0$$

Given (the agent's) *ex post* first-order condition (C.2.1), we can rewrite the *ex ante* first-order condition as

$$\underbrace{t^P q_m}_{\text{direct}} + \underbrace{t^P_e e^P_m \cdot (p+q)}_{\text{strategic}} = 0 \tag{C.2.4}$$

comprising two effects of marginal increase in manipulation:

(1) the direct gain due to the increased probability of bonus payment;

(2) the strategic loss due to the lowered bonus level since principal induces less effort.

To better see how the agent's optimal manipulation changes with performance measure alignment, we define $\rho \equiv \frac{p_e}{q_m}$ as the (marginal) ratio of performance measure sensitivities, which can be interpreted as a "multiplier" between effort and manipulation inducements with any incentive contract, and is therefore embedded in the effects of manipulation on effort inducement and rent extraction, which underpin both the direct and strategic effects of the agent's manipulation choice.

Lemma C.10. *The effects of the agent's manipulation decision on the motivational rent are fully characterized by the resulting effort inducement decision by the principal.*

Proof. It is straight-forward to observe this property with the direct effect by rewriting it as

$$t^P q_m = \frac{c_e}{p_e} q_m = \frac{c_e}{\rho}$$

The strategic effect is more complex. We first rewrite the interim first-order condition (C.2.2) as

$$p + q = \frac{u_e - c_e}{t_e^P}$$

Plug it in the second derivative of the principal's interim value

$$\begin{aligned}
\mathcal{S}(e^P, m, \phi) =& u_{ee} - c_{ee} - \frac{t_{ee}^P}{t_e^P}(u_e - c_e) - \frac{c_{ee} p_e - p_{ee} c_e}{p_e} \\
=& u_{ee} - 2c_{ee} - \left(\frac{c_{eee} p_e - p_{eee} c_e}{c_{ee} p_e - p_{ee} c_e} - 2\frac{p_{ee}}{p_e} \right)(u_e - c_e) + \frac{p_{ee} c_e}{p_e} \\
=& u_{ee} - 2c_{ee} - \left(\frac{c_{eee} f_e - f_{eee} c_e}{c_{ee} f_e - f_{ee} c_e} - 2\frac{f_{ee}}{f_e} \right)(u_e - c_e) + \frac{f_{ee} c_e}{f_e}
\end{aligned}$$

where f (from the decomposition of $p(e, \phi) = f(e) \cdot g(\phi)$) and its derivatives are functions of e only. Hence $\mathcal{S}(e^P, \cdot, \cdot)$ can be expressed equivalently as $\mathcal{S}(e^P)$.

The strategic effect can therefore be rewritten as

$$\begin{aligned}
&t^P_e e^P_m \cdot (p+q) \\
=&(u_e - c_e)e^P_m \\
=&(u_e - c_e)\frac{t^P_e q_m}{\mathcal{S}} \\
=&\frac{(u_e - c_e)}{\mathcal{S}}\frac{c_{ee}p_e - p_{ee}c_e}{p_e^2}q_m \\
=&\frac{(u_e - c_e)}{\mathcal{S}}\left(\frac{c_{ee}}{\rho} - \frac{p_{ee}c_e}{p_e\rho}\right) \\
=&\frac{(u_e - c_e)}{\mathcal{S}\rho}\left(c_{ee} - \frac{f_{ee}c_e}{f_e}\right)
\end{aligned}$$

Therefore both the direct and strategic effects, when multiplied by ρ, are functions of e only. Hence the result. □

We now prove Proposition 2.2, which we repeat here.

Proposition. *Irrespective of ϕ, the agent's optimal manipulation makes the principal induce the same level of effort.*

Proof. Recall the *ex ante* first-order condition as

$$\underbrace{t^P q_m}_{\text{direct}} + \underbrace{t^P_e e^P_m \cdot (p+q)}_{\text{strategic}} = 0$$

The direct effect is proportional to ρ^{-1} is straight-forward: the bonus level is proportional to ϕ^{-1} and the marginal increase in payment probability is $1-\phi$. For the strategic effect, ρ^{-1} enters through e^P_m, the principal's effort inducement reaction to manipulation. It is proportional to ρ^{-1} for the same reason. We can therefore multiply the *ex ante* first-order condition by ρ and write

$$c_e + \frac{(u_e - c_e)}{\mathcal{S}} \left(c_{ee} - \frac{f_{ee} c_e}{f_e} \right) = 0$$

This condition has a unique solution e^* (otherwise $m^A(\phi)$ cannot be unique). Hence $m^A(\phi)$ must satisfy $e^P(m^A(\phi), \phi) = e^*$. □

As a corollary, we can characterize how the optimal manipulation changes with performance measure alignment. Take the total derivative of $e^P(m^A(\phi), \phi) = e^*$ w.r.t. ϕ, we get

$$m^A_\phi = -\frac{e^P_\phi}{e^P_m} > 0$$

Corollary C.4. *The agent's optimal manipulation increases in performance measure alignment.*

C.2.1.3. *Ex ante* values. The principal's *ex ante* value given the agent's optimal manipulation choice $m^A(\phi)$ is given by

$$\bar{V}^P(\phi) \equiv V^P(m^A(\phi), \phi) = u(e^P) - h(m^A) - t^P \cdot (p + q)$$

where $m^A \equiv m^A(\phi)$, $e^P \equiv e^P(m^A, \phi)$, $t^P \equiv t^P(e^P, \phi)$, $p \equiv p(e^P, \phi)$ and $q \equiv q(m^A, \phi)$.

Let $\bar{e}^P_\phi \equiv e^P_m m^A_\phi + e^P_\phi$ be the total derivative of $e^P(m^A(\phi), \phi)$ w.r.t. ϕ (i.e. taking into account the agent's manipulation choice). The mariginal effect of performance measure alignment in the principal's *ex ante* value is therefore

$$\frac{\partial \bar{V}^P}{\partial \phi} = u_e \cdot \bar{e}^P_\phi - h_m \cdot m^A_\phi - (t^P_e \bar{e}^P_\phi + t^P_\phi) \cdot (p+q) - t^P \cdot (p_e \bar{e}^P_\phi + p_\phi + q_m m^A_\phi + q_\phi)$$

Before discussing the sign of this effect, we first show the following lemma.

Lemma C.11. *From the ex ante perspective, the performance measure alignment does not affect the agent's motivational rent from manipulation, i.e.*

$$\frac{\partial t^P \cdot q(m^A(\phi), \phi)}{\partial \phi} = 0$$

Proof. Recall that the agent's optimal manipulation $m^A(\phi)$ yields a fixed level of effort inducement e^* by the principal. e^* therefore must satisfy the principal's interim first-order condition (C.2.2), which can be rewritten as

$$t^P(e^*, \phi) \cdot p_e(e^*, \phi) + t^P_e(e^*, \phi) \cdot p(e^*, \phi) + t^P_e(e^*, \phi) \cdot q(m^A(\phi), \phi) = u_e$$

Take derivative w.r.t. ϕ we get

$$\frac{\partial t^P \cdot p_e}{\partial \phi} + \frac{\partial t^P_e \cdot p}{\partial \phi} + \frac{\partial t^P_e \cdot q^A(\phi)}{\partial \phi} = 0$$

where $q^A(\phi) \equiv q(m^A(\phi), \phi)$.

Lemma C.9 implies that the first two terms are 0 (for e^* is fixed), which means that

$$\frac{\partial t^P_e \cdot q^A}{\partial \phi} = 0$$

Recall that $t^P_{e\phi} = -\frac{p_{e\phi}}{p_e} t^P_e = -\frac{g_\phi}{g} t^P_e$ where $f(e) \cdot g(\phi) = p(e, \phi)$ is a decomposition of the latter; rewrite the previous condition as

$$\begin{aligned} &\frac{\partial t^P_e \cdot q^A}{\partial \phi} \\ =& t^P_e \cdot q^A_\phi + t^P_{e\phi} \cdot q^A \\ =& t^P_e \left(q^A_\phi - \frac{g_\phi}{g} \cdot q^A \right) \\ =& 0 \end{aligned}$$

Given that $t^P_e \neq 0$, q^A must satisfy

$$q^A_\phi - \frac{g_\phi}{g} \cdot q^A = 0$$

Now we turn to the bonus paid towards manipulation, and recall that $t^P_\phi = -\frac{p_{e\phi}}{p_e} t^P = -\frac{g_\phi}{g} t^P$.

$$\begin{aligned} &\frac{\partial t^P \cdot q(m^A(\phi), \phi)}{\partial \phi} \\ =& t^P \cdot q^A_\phi + t^P_\phi \cdot q^A \\ =& t^P \cdot \left(q^A_\phi - \frac{g_\phi}{g} \cdot q^A \right) \\ =& 0 \end{aligned}$$

□

In addition, we can characterize the functional form of $q^A(\phi)$, i.e. $q(m^A(\phi), \phi)$, by rewriting the previous condition as

$$\frac{\partial \ln q^A}{\partial \phi} = \frac{\partial \ln g}{\partial \phi}$$

which means $q^A(\phi) = Q \cdot g(\phi)$ where Q is some positive constant or function independent of ϕ.

We can now prove Proposition 2.3, which we repeat here.

Proposition C.1. *The principal's ex ante payoff decreases in performance measure alignment.*

PROOF. First we rearrange the terms

$$\begin{aligned}\frac{\partial \bar{V}^P}{\partial \phi} = \; & [u_e - t_e^P \cdot (p+q) - t^P \cdot p_e] \cdot \bar{e}_\phi^P - (t_\phi^P \cdot p + t^P \cdot p_\phi) \\ & - (t_\phi^P \cdot q + t^P \cdot q_m m_\phi^A + t^P \cdot q_\phi) - h_m \cdot m_\phi^A \\ = \; & [u_e - t_e^P \cdot (p+q) - t^P \cdot p_e] \cdot \bar{e}_\phi^P - \frac{\partial t^P p}{\partial \phi} - \frac{\partial t^P q}{\partial \phi} - h_m \cdot m_\phi^A\end{aligned}$$

The first term is the effect through the change in the principal's effort inducement choice. Proposition 2.2 implies that $\bar{e}_\phi^P = 0$, i.e. the agent's manipulation guarantees that the principal's effort inducement will not change with ϕ. Therefore this term is 0.

The rest of the terms are effects with effort fixed. The second term is the effect through the change in bonus paid towards effort. Corollary (C.3) implies that this term is always 0. The third term is the effect through the change in bonus paid towards manipulation, taking into account the agent's manipulation choice. Lemma C.11 implies that this term is 0 in equilibrium.

Hence only the last term remains, representing the direct harm to the principal:

$$\frac{\partial \bar{V}^P}{\partial \phi} = -h_m \cdot m_\phi^A < 0$$

□

C.2.2. Policy intervention

C.2.2.1. Additional signals. We prove Lemma 2.2, which we repeat here.

Lemma. *When σ is contractible, the principal's optimal contract rewards the agent only when $\kappa = H$ and $\sigma = E$.*

PROOF. Let $t_{\kappa,\sigma}$ be the contractual transfer to agent i under the realization (κ, σ). Agent's ex-post program is

$$\max_{e\in[0,1]} \mathbb{E}[t|e] - c(e)$$

We can write

$$\begin{aligned}\mathbb{E}[t|e] &= \sum_{\kappa}\sum_{\omega}\sum_{\sigma} \Pr[\omega]\Pr[\kappa|\omega] r_{\sigma|\omega} t_{\kappa,\sigma} \\ &= T + \tau e\end{aligned}$$

where

$$T \equiv \phi(t_{L,E}r_{E|\mathcal{E}}+t_{L,M}r_{M|\mathcal{E}})+(1-\phi)r_{E|\mathcal{M}}[t_{H,E}m+t_{L,E}(1-m)]+(1-\phi)r_{M|\mathcal{M}}[t_{H,M}m+t_{L,M}(1-m)]$$

$$\tau \equiv \phi[r_{E|\mathcal{E}}(t_{H,E} - t_{L,E}) + r_{M|\mathcal{E}}(t_{H,M} - t_{L,M})]$$

When $e \in [0, 1]$ does not bind, then first-order condition that determines the agent's effort choice is

$$c_e(e) = \tau \tag{C.2.5}$$

Let $e^A(\tau)$ be the agent's optimal choice which increases in τ. The principal's interim program is

$$\max_{t,e} u(e) - \mathbb{E}[t|e]$$

subject to

$$\begin{cases} e = e^A(\tau) & \text{(IC)} \\ t \geq 0 & \text{(LL)} \end{cases}$$

Write out $\mathbb{E}[t|e]$ and plug in (IC),

$$\max_{t \geq 0} u(e^A(\tau)) - T - \tau e^A(\tau)$$

For finding out the optimal contract for the principal, we first eliminate the suboptimal ones.

Notice that either $t_{H,E} - t_{L,E} < 0$ or $t_{H,M} - t_{L,M} < 0$ cannot be optimal given the (LL) constraint; the principal could set $t_{H,E} = t_{L,E} = 0$ or $t_{H,M} = t_{L,M} = 0$ that would induce higher effort (as τ increases) at lower cost. By the same token, $t_{L,E} > 0$ or $t_{L,M} > 0$ cannot be optimal since the principal can find a better contract t' that induces the same effort at lower cost ($t'_{H,E} = t_{H,E} - t_{L,E}$, $t'_{H,M} = t_{H,M} - t_{L,M}$ and $t'_{L,E} = t'_{L,M} = 0$).

We can now rewrite

$$T = (1-\phi)(r_{E|\mathcal{M}} t_{H,E} + r_{M|\mathcal{M}} t_{H,M})m$$

$$\tau \equiv \phi(r_{E|\mathcal{E}} t_{H,E} + r_{M|\mathcal{E}} t_{H,M})$$

Given that $\frac{r_{E|\mathcal{M}} r_{M|\mathcal{E}}}{r_{E|\mathcal{E}}} < r_{M|\mathcal{M}}$, by setting $t'_{H,M} = 0$ and $t'_{H,E} = t_{H,E} + \frac{r_{M|\mathcal{E}}}{r_{E|\mathcal{E}}} t_{H,M}$ we find a better contract t' that induces the same effort at lower cost.

Therefore the optimal contract must have $t_{L,E} = t_{H,M} = t_{L,M} = 0$ and $t_{H,E} > 0$. □

We now prove Lemma 2.3, which we repeat here.

Lemma. *Any outcome of a given performance measure that yields high performance with probability $p(e) + q(m)$ can be replicated by another performance measure that yields high performance with probability $\alpha p(e) + \alpha q(m)$ where $\alpha > 0$.*

Proof. For κ that yields high performance with probability $p(e) + q(m)$, let $\hat{\kappa}$'s probability be $\alpha p(e) + \alpha q(m)$. We verify the following three results.

(1) For any t that induces e from the agent with κ, $\hat{t} = \frac{t}{\alpha}$ would also induce e with $\hat{\kappa}$. This is straight-forward from the first-order condition (C.2.1).

(2) Any m that makes the principal to induce e with κ would also make her do so with $\hat{\kappa}$. This is straight-forward from the first-order condition (C.2.2) noting that $\hat{t} = \frac{t}{\alpha}$.

(3) The expected bonus payment given any (e, m) is identical with either (κ, t) or $(\hat{\kappa}, \hat{t})$. So are the principal's and the agent's payoffs.

Therefore, any of the principal's or the agent's programs always yields the same outcome with either κ or $\hat{\kappa}$. □

C.2.2.2. Costly manipulation. With manipulation cost,

$$U^A = t - c(e) - g_m(m)$$

the agent's *ex ante* first-order condition (multiplied by ρ) becomes

$$\text{(C.2.6)} \qquad \underbrace{t^P q_m}_{\text{direct}} + \underbrace{t^P_e e^P_m \cdot (p+q)}_{\text{strategic}} - \rho g_m = 0$$

The ambiguous effect of ϕ can also be inferred in Equation (2.2.1). On one hand, when $\rho \to 0$ the last term is dominated by the trade-off in the first two terms. When manipulation is very cost-effective (i.e. when the performance measure is badly aligned and manipulation is hardly costly), the trade-offs that we previously discussed is very relevant; better alignment impairs the deterrence effect and yields more manipulation. Manipulation is lower than in the costless case, more so the better the alignment (given the convex manipulation cost); effort inducement therefore increases with better alignment. The principal, however, would still be harmed overall, yet less so than the benchmark case. At some point the benefit from more effort inducement would offset the harm from more manipulation, and the principal's *ex ante* payoff levels off. On the other hand, when $\rho \to \infty$, manipulation becomes too ineffective to be of real concern: the first-order condition holds only if m is sufficiently small. To put it in another way, ρg_m is the marginal cost of manipulation when its effectiveness is taken into account; better alignment makes the set-up too expensive; the agent thus makes the set-up less manipulative and thus benefits the principal.

The optimal manipulation set-up in the quadratic-linear specification is a single-peaked function of ρ with the peak at $\hat{\rho} = \sqrt{\frac{3\gamma}{4\mu}}$. The comparative statics are intuitive: the ratio of costs ($\frac{\gamma}{\mu}$) indicates the relative importance of the motivational rent and manipulation cost;

high ratio therefore results in the dominance of the deterrence effect of the performance measure misalignment.

C.2.3. Two-period model

We first prove the case where manipulation negatively affects the realization of high performance in the first period. In Section C.2.3.1, we show that the same result can come from the interaction between effort and manipulation on the agent's cost.

We assume that the realization of the performances are $\Pr[\kappa_1 = H] = p(e_1, \phi) - r(m, \phi)$ and $\Pr[\kappa_2 = H] = p(e_2, \phi) + q(m, \phi)$, where r is the performance deterioration due to the set-up. We assume that $r_m \geq 0, r_\phi \leq 0$ and $r_{m\phi} \leq 0$.[3] The principal's payoff is $U^P = u(e_1) + u(e_2) - h(m) - t_1 - t_2$ and the agent's payoff is $U^A = t_1 + t_2 - c(e_1) - c(e_2)$. Again we assume that (a) the solution (e, m) is always interior, and (b) both the agent's and the principal's optimization problems are strictly concave.

Let $(\hat{e}^A, \hat{m}^A)$ as functions of t be the unique optimal choice in the agent's first-period program

$$\max_{(e,m)\in[0,1]^2} U^A(e, m, t) \equiv \mathbb{E}[t|e, m] + V^A(m) - c(e)$$

Let $\tilde{e}^A(m, t)$ be the unique optimal effort choice in the agent's *partial* program where the manipulation level m is fixed.

$$\max_{e\in[0,1]} U^A(e, m, t) \equiv \mathbb{E}[t|e, m] + V^A(m) - c(e)$$

[3] The signs of r_ϕ and $r_{m\phi}$ strengthen the definition of "improvement" in the sense that an improved performance measure is aslo better insulated from the set-up activity itself. This assumption does not affect the results in this section and will be more discussed in Section 2.3.1.

Lemma C.12. *$\frac{\partial \hat{e}^A(t)}{\partial t} = \frac{\partial \tilde{e}^A(\hat{m}^A(t),t)}{\partial t} > 0$ and $\frac{\partial \hat{m}^A(t)}{\partial t} < 0$.*

Proof. We first characterize $(\hat{e}^A, \hat{m}^A)$ and $\tilde{e}^A$.

$(\hat{e}^A, \hat{m}^A)$ is determined by the first-order conditions

$$\begin{cases} U_e^A(\cdot,\cdot,t) \equiv tp_e(\cdot) - c_e(\cdot) = 0 \\ U_m^A(\cdot,\cdot,t) \equiv -tr_m(\cdot) + V_m^A(\cdot) = 0 \end{cases} \tag{C.2.7}$$

while the second-order conditions is satisfied for all $(e,m) \in [0,1]^2$

$$\begin{cases} U_{ee}^A \equiv tp_{ee} - c_{ee} < 0 \\ (U_{em}^A)^2 - U_{ee}^A U_{mm}^A \equiv (tp_{ee} - c_{ee})(tr_{mm} - V_{mm}^A) < 0 \end{cases}$$

Take the derivative of (C.2.7) w.r.t. t:

$$\hat{e}_t^A(t) = -\frac{p_e}{U_{ee}^A}\bigg|_{(\hat{e}^A,\hat{m}^A)} > 0$$

and

$$\hat{m}_t^A(t) = -\frac{r_m(\cdot)}{tr_{mm} - V_{mm}^A} < 0 \tag{C.2.8}$$

where the inequalities come from the second-order conditions.

Similarly, $\tilde{e}^A$ is determined by the first-order condition

$$U_e^A(\cdot,m,t) \equiv tp_e - c_e = 0 \tag{C.2.9}$$

while the second-order conditions is satisfied for all $(e, m) \in [0, 1]^2$

$$U^A_{ee} \equiv tp_{ee} - c_{ee} < 0$$

Take the derivative of (C.2.9) w.r.t. t:

$$\tilde{e}^A_t(m, t) = -\frac{p_e}{U^A_{ee}}\bigg|_{(\tilde{e}^A, m)} > 0$$

where the inequality comes from the second-order condition.

Note that

$$\tilde{e}^A(\hat{m}^A(t), t) = \hat{e}^A(t) \tag{C.2.10}$$

by comparing (C.2.7) and (C.2.9) at $m = \hat{m}^A(t)$. Therefore,

$$\hat{e}^A_t(t) = \tilde{e}^A_t(\hat{m}^A(t), t) > 0$$

Hence the result. □

For the rest of the results, we are agnostic on how effort and manipulation interact in the first period; we assume only that $\hat{e}_t \geq \tilde{e}_t$ and $\hat{m}_t < 0$ to accommodate the alternative interpretation in Section C.2.3.1.

We prove Proposition 2.6 in this more general setting, which we repeart here. The proof also includes that of Lemma 2.4.

Proposition. *The principal sets the first-period bonus higher than it would be if the agent's manipulation were fixed at its equilibrium level.*

Proof. Let $\hat{t}$ be the unique optimal bonus choice in the principal's first-period program

$$\max_{t\geq 0} \hat{U}^P(t) \equiv u(\hat{e}(t)) - \mathbb{E}[t|\hat{e}(t)] + V^P(\hat{m}(t))$$

and with abuse of notation let the associated equilibrium manipulation level be $\hat{m} \equiv \hat{m}(\hat{t})$, and similarly $\hat{m}_t \equiv \hat{m}_t(\hat{t})$ be its local derivative w.r.t t.

Let $\tilde{t}(m)$ be the unique optimal bonus choice in the principal's *auxiliary* first-period program with fixed manipulation m

$$\max_{t\geq 0} \tilde{U}^P(m,t) \equiv u(\tilde{e}(m,t)) - \mathbb{E}[t|\tilde{e}(m,t)] + V^P(m)$$

Note that in this program, m matters to the principal only through its effect on the agent's effort cost. This proposition is equivalent to $\hat{t} > \tilde{t}(\hat{m})$, which can be shown by comparing the derivatives of $\hat{U}^P(t)$ and $\tilde{U}^P(\hat{m},t)$ w.r.t. t:

$$\hat{U}^P_t(t) = [u_e(\hat{e}(t)) - t\cdot p_e(\hat{e}(t))]\cdot \hat{e}_t(t) - p(\hat{e}(t)) + V^P_m(\hat{m}(t))\cdot \hat{m}_t(t)$$

$$\tilde{U}^P_t(\hat{m},t) = [u_e(\tilde{e}^A(\hat{m},t)) - t\cdot p_e(\tilde{e}(\hat{m},t))]\cdot \tilde{e}_t(\hat{m},t) - p(\tilde{e}(\hat{m},t))$$

By definition we have $\hat{U}^P_t(\hat{t}) = 0$. Recall that we assume both programs to be strictly concave. Hence $\hat{t} > \tilde{t}(\hat{m})$ iff $0 > \tilde{U}^P_t(\hat{m},\hat{t})$.

With abuse of notation let $\hat{e} \equiv \hat{e}(\hat{t})$, $\hat{e}_t \equiv \hat{e}_t(\hat{t})$, $\tilde{e} \equiv \tilde{e}(\hat{m},\hat{t})$ and $\tilde{e}_t \equiv \tilde{e}_t(\hat{m},\hat{t})$. Equation (C.2.14) implies that $\hat{e} = \tilde{e}$. Therefore

$$\tilde{U}_t^P(\hat{m},\hat{t}) = \frac{\tilde{e}_t}{\hat{e}_t}[\hat{U}_t^P(t) - V_m^P(\hat{m}) \cdot \hat{m}_t] - \frac{\hat{e}_t - \tilde{e}_t}{\hat{e}_t} p(\hat{e}_t)$$

Given that $V_m^P(\hat{m}) < 0$ from Equations (C.2.3) and $\hat{m}_t < 0$, it must be that

$$\tilde{U}_t^P(\hat{m},\hat{t}) < -\frac{\hat{e}_t - \tilde{e}_t}{\hat{e}_t} p(\hat{e}_t)$$

Given that $\hat{e}_t \geq \tilde{e}_t$, $\tilde{U}_t^P(\hat{m},\hat{t}) < 0$ and thus $\hat{t} > \tilde{t}(\hat{m})$. □

In addition, we assume that $\tilde{t}(m)$ is weakly increasing (e.g. if $\tilde{U}^A$ satisfies supermodularity in (m,t)) in the indirect interaction case; in other words, the principal's bonus decision is weakly increasing in the agent's cost of effort.

For direct interaction, no additional assumption is needed; it can be derived from the existing assumptions (particularly from the separability between m and e in the first-period). The intuition is that higher m actually lowers the probability of bonus payment while not affecting the effort inducement. Effort inducement is therefore cheaper, prompting the principal to induce more effort by setting higher bonus.

Lemma C.13. *With direct interaction, the principal's optimal bonus increases with manipulation in the auxiliary program.*

Proof. Recall the program: for a fixed level of manipulation m

$$\max_t \tilde{U}^P(m,t) = u(\tilde{e}) - \mathbb{E}[t|\tilde{e}] + V^P(m)$$

Take derivative w.r.t. t

$$\tilde{U}_t^P = u_e\tilde{e}_t - tp_e\tilde{e}_t - (p - r)$$

Take derivative again w.r.t. m

$$\tilde{U}^P_{tm} = r_m > 0$$

$\tilde{U}^P$ is supermodular in (m, t) and hence $\tilde{t}_m > 0$. □

Then we can prove Proposition 2.7, which we repeat here.

Proposition C.2. *The optimal first-period bonus is higher than the "second-best" benchmark.*

Proof. Recall that $\tilde{t}(m)$ is the unique optimal bonus choice in the principal's *partial* first-period program with fixed manipulation m

$$\max_{t \geq 0} \tilde{U}^P(m, t) \equiv u(\tilde{e}(m, t)) - \mathbb{E}[t|\tilde{e}(m, t)] + V^P(m)$$

$\tilde{t}(0)$ would coincide with the "second best" benchmark, i.e. the optimal contract when manipulation is infeasible. The previous proposition shows that $\hat{t} > \tilde{t}(\hat{m})$. Given that $\tilde{t}$ is weakly increasing, we then have $\hat{t} > \tilde{t}(0)$. □

C.2.3.1. Indirect interaction. As a detour, in this section we show that the interaction between effort and manipulation need not come from their opposing effect on high performance, but from the cost side. We refer to this case as indirection interaction, and the additional complication is dealt here.

Let $\Pr[\kappa_1 = H] = p(e_1, \phi)$ and $\Pr[\kappa_2 = H] = p(e_2, \phi) + q(m, \phi)$. The principal's payoff is $U^A = u(e_1) + u(e_2) - h(m) - t_1 - t_2$ and the agent's payoff is $U^A = t_1 + t_2 - k(e_1, m) - k(e_2, 0)$ where k is the effort cost strictly increasing and convex in (e, m). $k_e(0, \cdot) = 0$ and

$k_{em} > 0$. Without loss of generality, we let $k(e, 0) = c(e)$ from the previous model. This cost structure suggests that effort and set-up are substitutes in the short term: the more the agent manipulates for the future gain, the more costly his effort exertion becomes in the current period. Notice that if instead $k_{em} = 0$, the analysis is identical to that in Section 2.2.2 with an additional first-period effort inducement problem independent of the rest of the analysis.

Let $(\hat{e}^A, \hat{m}^A)$ as functions of t be the unique optimal choice in the agent's first-period program

$$\max_{(e,m)\in[0,1]^2} U^A(e, m, t) \equiv \mathbb{E}[t|e] + V^A(m) - k(e, m)$$

Let $\tilde{e}^A(m, t)$ be the unique optimal effort choice in the agent's *partial* program where the manipulation level m is fixed.

$$\max_{e\in[0,1]} U^A(e, m, t) \equiv \mathbb{E}[t|e] + V^A(m) - k(e, m)$$

Lemma C.14. *$\frac{\partial \hat{e}^A(t)}{\partial t} > \frac{\partial \tilde{e}^A(\hat{m}^A(t),t)}{\partial t} > 0$ and $\frac{\partial \hat{m}^A(t)}{\partial t} < 0$.*

Proof. We first characterize $(\hat{e}^A, \hat{m}^A)$ and $\tilde{e}^A$.

$(\hat{e}^A, \hat{m}^A)$ is determined by the first-order conditions

$$\text{(C.2.11)} \qquad \begin{cases} U^A_e(\cdot, \cdot, t) \equiv tp_e(\cdot) - k_e(\cdot, \cdot) = 0 \\ U^A_m(\cdot, \cdot, t) \equiv V^A_m(\cdot) - k_m(\cdot, \cdot) = 0 \end{cases}$$

while the second-order conditions is satisfied for all $(e, m) \in [0,1]^2$

$$\begin{cases} U^A_{ee} \equiv tp_{ee} - k_{ee} < 0 \\ (U^A_{em})^2 - U^A_{ee}U^A_{mm} \equiv k^2_{em} - (tp_{ee} - k_{ee})(V^A_{mm} - k_{mm}) < 0 \end{cases}$$

which is implied by extending second-order conditions in the previous section, replacing effort cost $c(\cdot)$ with $k(\cdot, m)$ and imposing those conditions for all $m \in [0,1]$.

Take the derivative of (C.2.11) w.r.t. t and substitute out $\hat{m}^A_t$; we get

$$\hat{e}^A_t(t) = -\frac{p_e}{U^A_{ee} - \frac{(U^A_{em})^2}{U^A_{mm}}}\Bigg|_{(\hat{e}^A, \hat{m}^A)} > -\frac{p_e}{U^A_{ee}}\Bigg|_{(\hat{e}^A, \hat{m}^A)} > 0$$

where the inequalities come from the second-order conditions. Auxilliarily,

$$\hat{m}^A_t(t) = \frac{U^A_{em}}{U^A_{mm}}\hat{e}^A_t = \frac{k_{em}}{U^A_{mm}}\hat{e}^A_t < 0 \tag{C.2.12}$$

Similarly, $\tilde{e}^A$ is determined by the first-order condition

$$U^A_e(\cdot, m, t) \equiv tp_e - k_e(\cdot, m) = 0 \tag{C.2.13}$$

while the second-order conditions is satisfied for all $(e, m) \in [0,1]^2$

$$U^A_{ee} \equiv tp_{ee} - k_{ee} < 0$$

which is implied from the previous conditions.

Take the derivative of (C.2.13) w.r.t. t; we get

$$\tilde{e}_t^A(m,t) = -\frac{p_e}{U_{ee}^A}\bigg|_{(\tilde{e}^A,m)} > 0$$

where the inequality comes from the second-order condition.

Note that

$$\tilde{e}^A(\hat{m}^A(t),t) = \hat{e}^A(t) \tag{C.2.14}$$

by comparing (C.2.11) and (C.2.13) at $m = \hat{m}^A(t)$. Therefore,

$$\hat{e}_t^A(t) > -\frac{p_e}{U_{ee}^A}\bigg|_{(\hat{e}^A,\hat{m}^A)} = \tilde{e}_t^A(\hat{m}^A(t),t) > 0$$

Hence the result. □

C.3. Incentives in Delegation

Throughout the proofs, with abuse of notation we use $e_A(\alpha)$ to express the agent's *optimal* effort on a project of impact α. We also assume that the agent's effort cost is more general than the quadratic function, i.e., c' is log-concave, $\lim_{e\to 0}(\ln c'(e))' > 1$ and γ is sufficiently small.

C.3.1. Symmetric-information benchmark

We prove Proposition 3.1 that the principal delegates projects whose impact is below a threshold, which we repeat here.

Proposition. *There exists a threshold $\alpha^\dagger \in (\beta, \alpha^U]$ such that*

(1) If $\alpha \leq \alpha^\dagger$, the principal delegates and exerts 1 unit of effort on the β-impact project. The agent exerts $h(\gamma\alpha)$ unit of effort on the α-impact project.

(2) Otherwise, the principal centralizes and exerts 1 unit of effort on the α-impact project.

Proof. We first analyze the centralization payoff. Under centralization, the principal's effort choice is the corner solution: she exerts 1 unit of effort on whichever project with higher impact. Her payoff is thus

$$u_C(\alpha) = \max\{\alpha, \beta\}$$

Under delegation, the agent's effort e_A satisfies $c'(e_A) = \gamma\alpha$ and therefore $e_A(\alpha) = h(\gamma\alpha)$. The principal's payoff is thus

$$u_D(\alpha) = \alpha e_A(\alpha) + \beta$$

Notice that for $\alpha \leq \beta$, $u_C(\alpha) < u_D(\alpha)$. For $\alpha \in (\beta, 1]$, we show that u_D crosses u_C at most once and from above. To do so, we show that $u_D(\cdot)$ is strictly convex and that $u'_D(\cdot) \leq 1$ for all $\alpha \in (\beta, 1]$.

We first show that $u_D(\cdot)$ is strictly convex. Note that $u''_D(\alpha) = 2e'_A(\alpha) + \alpha e''_A(\alpha)$, $e'_A(\alpha) = \frac{\gamma}{c''}$ and $e''_A(\alpha) = -\frac{\gamma^2 c'''}{(c'')^3}$. Since $c''(\cdot) > 0$ and $c'''(\cdot) \leq 0$, $u_D(\cdot)$ is strictly convex.

We now show that $\lim_{e \downarrow 0} \frac{c'(e)}{c''(e)} < 1$ is sufficient for the existence of $\hat{\gamma}$ such that $\forall \gamma \leq \hat{\gamma}$, $u'_D(\cdot) \leq 1$.

Because u_D is strictly convex, $u'_D(\cdot) \leq 1$ if $u'_D(1) \leq 1$.

Note that $u'_D(1) = e_A(1) + \frac{c'(e_A(1))}{c''(e_A(1))}$. For $\gamma = 0$, $e_A(1) = 0$ and $\lim_{e \downarrow 0} \frac{c'(e)}{c''(e)} < 1$ implies that $u'_D(1) < 1$. Continuity of u'_D then guarantees the existence of $\hat{\gamma}$ such that $\forall \gamma \leq \hat{\gamma}$, $u'_D(1) < 1$ which guarantees $u'_D(\cdot) < 1$.

Recall that u_C is a straight-line with slope of 1 in the range of $[\beta, 1]$ and that $u_C(\beta) < u_D(\beta)$. Therefore u_D crosses u_C at most once and from above. $\square$

Note that the power function $c(e) = \lambda e_A^\theta$ with $\theta \in (1, 3]$, $\lambda > 0$ satisfies these assumptions. A sufficient condition for $u'_D(\cdot) \leq 1$ is $\gamma \leq \lambda\theta/e$. In addition, the exponential cost function $c(e_A) = \exp(\theta e_A) - 1$ with $\theta > 1$ also satisfies these assumptions.

C.3.2. Lemons game

Before we analyze the lemons game, we first establish that A's posterior belief μ can be reduced to its expectation $\alpha_\mu \equiv \mathbb{E}_\mu[\alpha]$ for the purpose of characterizing A's strategy and P's payoff. Recall that A exerts uniform effort across all the delegated projects because he cannot differentiate one from another. This effort level e_A solves $\max_e \mathbb{E}_\mu \left[\gamma \alpha e - c\left(e\right)\right]$, which implies $e_A = h\left(\gamma \mathbb{E}_\mu\left[\alpha\right]\right) = h\left(\gamma \alpha_\mu\right) = e_A\left(\alpha_\mu\right)$.

Let $\hat{u}_D(\alpha, \alpha_\mu) \equiv \alpha \cdot e_A\left(\alpha_\mu\right) + \beta$ be P's ex-post payoff from delegating project α when A's posterior expectation is α_μ. $\hat{u}_D(\alpha, \alpha_\mu)$ is increasing in both arguments, and linear in the first. The slope of $\hat{u}_D(\cdot, \alpha_\mu)$, a line segment, is $e_A\left(\alpha_\mu\right) < 1$. The shape of $\hat{u}_D(\cdot, \alpha_\mu)$ will be exploited in the characterizations of PBEs. Note that in equilibrium, μ (and consequently α_μ) must coincide with the distribution of projects that P delegates. Hence P's expected payoff from delegation, whenever she induces μ (and consequently α_μ) as A's posterior, must in equilibrium be $\mathbb{E}_\mu[\alpha \cdot e_A\left(\alpha_\mu\right) + \beta] = \alpha_\mu \cdot e_A\left(\alpha_\mu\right) + \beta$. In other words, in equilibrium P's expected payoff from delegation coincides with $u_D(\alpha_\mu)$, his payoff from delegating project α_μ in the symmetric-information benchmark. Henceforth when there is no confusion we will use $u_D(\alpha_\mu)$ to express P's expected payoff from delegation in equilibrium when her strategy is delegating a distribution of μ over α.

We now prove Lemma 3.1 that the principal delegates projects whose impact is below a threshold, which we repeat here.

Lemma. *In any equilibrium of Game L where the set of delegated projects $\mathcal{A}_D$ has a non-zero measure, $\mathcal{A}_D$ takes the form of $[\alpha^L, \tilde{\alpha}]$ where $\tilde{\alpha} \geq \alpha^L$.*

Proof. Let the $\alpha_\mu > 0$ be the expection of α under delegation and pick any $\alpha_1 \in \mathcal{A}_D \equiv \{\alpha : g(\alpha) = D\}$ and α_2 such that $\alpha_1 > \alpha_2 \geq \alpha^L$. We will show that $\alpha_2 \in \mathcal{A}_D$.

$\alpha_1 \in \mathcal{A}_D$ implies that $\hat{u}_D\left(\alpha_1, \alpha_\mu\right) \geq u_C\left(\alpha_1\right)$. Then there are 2 cases:

$\alpha_2 \in (0, \beta]$ Given that $\alpha_\mu > 0$, $\hat{u}_D\left(\alpha_2, \alpha_\mu\right) > \beta = u_C\left(\alpha_2\right)$.

$\alpha_2 \in (\beta, 1]$ Then $\alpha_1 \in (\beta, 1]$. Notice that $\frac{\partial \widehat{u}_D(\alpha, \alpha_\mu)}{\partial \alpha} = e_A\left(\alpha_\mu\right) < 1 = u_C'\left(\alpha\right)$ for all $\alpha \in (\beta, 1]$. Hence

$$\begin{aligned} \hat{u}_D\left(\alpha_2, \alpha_\mu\right) &= \hat{u}_D\left(\alpha_1, \alpha_\mu\right) + \int_{\alpha_1}^{\alpha_2} e_A\left(\alpha_\mu\right) \mathrm{d}\alpha \\ &< u_C\left(\alpha_1\right) + \int_{\alpha_1}^{\alpha_2} 1 \mathrm{d}\alpha \\ &= u_C\left(\alpha_2\right) \end{aligned}$$

Therefore $\alpha_2 \in \mathcal{A}_D$. Hence $\mathcal{A}_D$ must extend all the way down to α^L. □

We now prove Lemma 3.2 that a "threshold delegation" equilibrium is identified when $\overline{u}_D(\cdot)$ intersects $u_C(\cdot)$ at some $\alpha^* \in (\alpha^L, \alpha^U)$. We repeat the lemma here.

Lemma. *An equilibrium exists for Game L in which the principal delegates according to a threshold* $\alpha^* \in [\alpha^L, \alpha^U]$ *if and only if* $\overline{u}_D(\alpha^*) = u_C(\alpha^*)$. *The agent's effort exertion is* $h(\gamma \mathbb{E}_{\mu_0}[\alpha | \alpha \leq \alpha^*])$.

Proof. Define $\overline{a}\left(\alpha\right) \equiv \mathbb{E}_{\mu_0}\left[\tilde{\alpha} | \tilde{\alpha} \leq \alpha\right]$ as the expectation over $\left[\alpha^L, \alpha\right]$, then $\overline{u}_D$ can be rewritten as $\overline{u}_D\left(\alpha\right) \equiv \hat{u}_D\left(\alpha, \overline{a}\left(\alpha\right)\right)$. Fix the threshold α^*. Define $u_D^*(\alpha) \equiv \hat{u}_D\left(\alpha, \overline{a}\left(\alpha^*\right)\right)$ as P's ex-post delegation payoff when A expects that P delegates according to a threshold at α^*.

We first show that $\overline{u}_D\left(\alpha^*\right) = u_C\left(\alpha^*\right)$ is necessary for the specified PBE. Given that the delegation threshold in the PBE is at α^*, $\overline{u}_D\left(\alpha^*\right) = u_D^*\left(\alpha^*\right)$, i.e. P's ex-post delegation payoff at α^* in the PBE. If either $u_D^*\left(\alpha^*\right) > u_C\left(\alpha^*\right)$ or $u_D^*\left(\alpha^*\right) < u_C\left(\alpha^*\right)$, P would then deviate and either delegate $\alpha^* + \varepsilon$ or centralizes $\alpha^* - \varepsilon$ for some small $\varepsilon > 0$, both feasible since $\alpha^* \in (\alpha^L, \alpha^U)$.

We now show that $\overline{u}_D\left(\alpha^*\right) = u_C\left(\alpha^*\right)$ is also sufficient for the specified PBE.

Firstly we show that $\alpha^* > \beta$. $\alpha^* \geq \alpha^L$ implies $\overline{u}_D\left(\alpha^*\right) \geq \overline{u}_D\left(\alpha^L\right) = u_D\left(\alpha^L\right) > u_D\left(0\right) = \beta$. Therefore $u_C\left(\alpha^*\right) = \overline{u}_D\left(\alpha^*\right) > \beta$, it must be that $\alpha^* > \beta$.

We now show that the strategies constitute a PBE. It is easy to verify that of A's belief is consistent with the P's delegation strategy, then his effort is optimal. Now we verify P's incentive to switch g at each $\alpha \in \left[\alpha^L, \alpha^U\right]$ except for at α^*, where both responsibility assignments yield the same ex-post payoff. Notice that P's ex-post delegation payoff $u_D^*(\cdot)$

is a line segment with slope smaller than 1. Her incentive to deviate is verified on two intervals.

$\alpha \in [\alpha^L, \alpha^*)$ On this interval, we need to show $u_C(\cdot) < u_D^*(\cdot)$. For $\alpha \leq \beta$ (which may not exist), $u_C(\alpha) = \beta < u_D^*(\alpha)$. For $\alpha > \beta$, $u_C(\cdot)$ is a line segment with slope 1, with the *right* endpoint coinciding with $u_D^*(\cdot)$ at α^*. Hence $u_C(\alpha) < u_D^*(\alpha)$ as well.

$\alpha \in (\alpha^*, \alpha^U]$ On this interval, we need to show $u_C(\cdot) > u_D^*(\cdot)$. $u_C(\cdot)$ is again a line segment with slope 1, with the *left* endpoint coinciding with $u_D^*(\cdot)$ at α^*. Hence $u_C(\alpha) > u_D^*(\alpha)$ on this interval.

As neither P nor A would deviate, a PBE is found. □

We now prove Proposition 3.2 that a unique "threshold delegation" equilibrium exists for the lemons game, and the threshold is lower than in the symmetric-information benchmark. We repeat the proposition here.

Proposition. *When $\alpha^L < \alpha^\dagger < \alpha^U$, there exists a unique equilibrium with delegation threshold $\alpha^* \in (\alpha^L, \alpha^U]$. Furthermore, $\alpha^* < \alpha^\dagger$.*

Proof. We first show that $\overline{u}_D'(\alpha) < 1$ for all $\alpha \in [\alpha^L, \alpha^U]$. Then given that u_C has slope 1 on $(\beta, 1]$, the mean value theorem would rule out more than one intersection point on $(\beta, 1]$. Then given that $\overline{u}_D(\alpha^L) = u_D(\alpha^L) > u_C(\alpha^L)$ and $\overline{u}_D(\alpha^U) < u_D(\alpha^U) < u_C(\alpha^U)$, the uniqueness result follows.

Decompose the two derivatives:

$$\overline{u}_D'(\alpha) = e_A\left(\overline{a}\left(\alpha\right)\right) + \alpha \cdot e_A'\left(\overline{a}\left(\alpha\right)\right) \cdot \overline{a}'\left(\alpha\right)$$

$$u'_D(\alpha) = e_A(\alpha) + \alpha \cdot e'_A(\alpha)$$

Since $e'_A(\cdot)$ is weakly increasing because $c''' \leq 0$, a sufficient condition for $\overline{u}'_D(\alpha) \leq u'_D(\alpha)$ is $\overline{a}'(\alpha) \leq 1$. This is equivalent to the log-concavity of $\mathcal{F}_0(\alpha)$, which is equivalent to $f_0(\alpha)\mathcal{F}_0(\alpha) \leq F_0(\alpha)^2$.

$$\begin{aligned}
\overline{a}'(\alpha) &\equiv \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\frac{\int_{\alpha^L}^{\alpha} x\mathrm{d}F_0(x)}{F_0(\alpha)}\right) \\
&= \frac{\alpha f_0(\alpha)}{F_0(\alpha)} - \frac{f_0(\alpha)\int_{\alpha^L}^{\alpha} x\mathrm{d}F_0(x)}{F_0(\alpha)^2} \\
&= \frac{f_0(\alpha)}{F_0(\alpha)^2}\left(\alpha F_0(\alpha) - \int_{\alpha^L}^{\alpha} x\mathrm{d}F_0(x)\right) \\
&= \frac{f_0(\alpha)}{F_0(\alpha)^2}\left(\int_{\alpha^L}^{\alpha} F_0(x)\mathrm{d}x\right) \\
&= \frac{f_0(\alpha)\mathcal{F}_0(\alpha)}{F_0(\alpha)^2} \\
&\leq 1
\end{aligned}$$

Now we show that $\alpha^* < \alpha^\dagger$. Notice that $\forall \alpha \in [\alpha^\dagger, \alpha^U], \overline{u}_D(\alpha) \equiv \hat{u}_D(\alpha, \overline{a}(\alpha)) < \hat{u}_D(\alpha, \alpha) = u_D(\alpha) \leq u_C(\alpha)$. The first inequality is guaranteed by $\forall \alpha_\mu > 0, \frac{\partial \widehat{u}_D(\alpha, \alpha_\mu)}{\partial \alpha_\mu} = \alpha \cdot e_A(\alpha_\mu) > 0$ and $\forall \alpha > \alpha^L, \overline{a}(\alpha) < \alpha$. The second inequality is guaranteed by the "single-crossing" specification of u_D, which implies that if u_D intersects u_C at $\alpha^\dagger$, then $\forall \alpha \in [\alpha^\dagger, \alpha^U], u_D(\alpha) \leq u_C(\alpha)$. Notice also that $\overline{u}_D(\alpha^L) = u_D(\alpha^L) > u_C(\alpha^L)$. Therefore $\overline{u}_D$ and u_C must intersect on $(\alpha^L, \alpha^\dagger)$ given the continuity of both functions.

□

We now prove prove Corollary 3.1 that the principal's payoff in the lemons game is lower than in the benchmark game. We repeat the corollary here.

Corollary. *The principal's expected payoff in Game L is lower than in Game B.*

PROOF. Let $g^\dagger$ be the delegation strategy in the SPNE of Game B and g^* in any PBE of Game L. Let $\mathcal{A}_D \equiv \{\alpha \in [\alpha^L, \alpha^U] : g^*(\alpha) = D\}$. P's expected payoff in that PBE is given by

$$\begin{aligned}
&\int_{\mathcal{A}_D} u_D\left(\mathbb{E}\left[\alpha|\mathcal{A}_D\right]\right) \mathrm{d}\mu_0 + \int_{[\alpha^L,\alpha^U]\setminus\mathcal{A}_D} u_C\left(\alpha\right) \mathrm{d}\mu_0 \\
\leq &\int_{\mathcal{A}_D} u_D\left(\alpha\right) \mathrm{d}\mu_0 + \int_{[\alpha^L,\alpha^U]\setminus\mathcal{A}_D} u_C\left(\alpha\right) \mathrm{d}\mu_0 \\
\leq &\int_{\alpha^L}^{\alpha^U} \max_g \left\{g_g\left(\alpha\right)\right\} \mathrm{d}\mu_0 \\
= &\int_{\alpha^L}^{\alpha^U} u_{g^\dagger(\alpha)}\left(\alpha\right) \mathrm{d}\mu_0
\end{aligned}$$

The last term is P's expected payoff in the SPNE of Game B.

Notice that the first inequality turns into equality if and only if $\mathcal{A}_D$ has zero-measure, which means that the PBE is a "full centralization" PBE, which is equivalent to $\alpha^\dagger \geq \alpha^U$ as discussed later. The second inequality turns to equality if and only if $g^* = g^\dagger$. Hence the two expected utilities coincide if and only if P always chooses $g = C$ in both equilibria. Given that the "full centralization" PBE implies that SPNE must also admit $g = C$. Hence $\alpha^\dagger \geq \alpha^U$ is necessary and sufficient for the two expected payoffs to coincide.

Let the difference between the two expected utilities be Δu. By rearranging terms we get

$$\Delta u = \underbrace{\int_{\alpha^L}^{\alpha^U} u_{g^\dagger(\alpha)}\left(\alpha\right) - u_{g^*(\alpha)}\left(\alpha\right) \mathrm{d}\mu_0}_{\text{loss from assignment}} + \underbrace{\int_{\mathcal{A}_D} u_D(\alpha) - u_D\left(\mathbb{E}\left[\alpha|\mathcal{A}_D\right]\right) \mathrm{d}\mu_0}_{\text{loss from effort}}$$

The first term is weakly positive, representing the loss of payoff from having suboptimal responsibility assignment in the equilibrium. The second term is also weakly positive, representing the loss of payoff from A exerting uniform effort across all delegated projects. Note that if $g = C$ in both equilibria, both losses shrink to 0. □

Equilibria when $\alpha^L < \alpha^\dagger < \alpha^U$ is not satisfied. For completeness, we also discuss the two other types of equilibria of Game L ruled out by the assumption that $\alpha^L < \alpha^\dagger < \alpha^U$.

Firstly, there could be a "full centralization" equilibrium in which the principal never delegates. This equilibrium exists if and only if she prefers to centralize the project of the least impact from the pool of delegatable projects, i.e., $\alpha^L \geq \alpha^\dagger$.

Lemma C.15. *An equilibrium exists for Game L in which the principal never delegates if and only if $u_C(\alpha^L) \geq \overline{u}_D(\alpha^L)$. The equilibrium admits the following strategies and (off-path) belief:*

$$g = C$$

$$\alpha_\mu = \alpha^L$$

$$e_A = h\left(\gamma\alpha^L\right)$$

Proof. Note that $\overline{u}_D\left(\alpha^L\right) = \hat{u}_D\left(\alpha^L, \alpha^L\right) = u_D\left(\alpha^L\right)$, since $\overline{a}\left(\alpha^L\right) = \alpha^L$.

$u_C\left(\alpha^L\right) \geq \overline{u}_D\left(\alpha^L\right)$ is necessary. Any PBE with full centralization strategy must have A's off-path belief as some $\alpha_\mu \in \left[\alpha^L, \alpha^U\right]$. Then no-deviation by P requires that $u_C\left(\alpha^L\right) \geq \hat{u}_D\left(\alpha^L, \alpha_\mu\right)$. Given that $\hat{u}_D$ is increasing in α_μ, $u_C\left(\alpha^L\right) > \hat{u}_D\left(\alpha^L, \alpha^L\right) = \overline{u}_D\left(\alpha^L\right)$.

$u_C\left(\alpha^L\right) \geq \overline{u}_D\left(\alpha^L\right)$ is also sufficient. It is easy to verify that A's belief is consistent and he would not deviate given his belief. P's effort choice is also optimal. To verify P would not switch to $g = D$, there are two cases regarding α^L, and $u_C(\cdot) \geq \hat{u}_D\left(\cdot, \alpha^L\right)$ in both cases.

$\alpha^L = 0$: $u_C(\cdot) \geq \beta = \hat{u}_D\left(\cdot, \alpha^L\right)$.

$\alpha^L > 0$: $u_C\left(\alpha^L\right) \geq u_D\left(\alpha^L\right)$ implies that α^L. Then $\hat{u}_D\left(\cdot, \alpha^L\right)$ is a line segment with slope smaller than 1, whereas $u_C(\cdot)$ is a line segment with slope 1. Given that $u_C\left(\alpha^L\right) \geq \hat{u}_D\left(\alpha^L, \alpha^L\right)$, $u_C(\cdot) \geq \hat{u}_D\left(\cdot, \alpha^L\right)$.

□

In this PBE, P never delegates because when she delegates, A would assume that the project is the least important type and exert little effort; the loss from A's shirking would therefore be too large to justify delegating any project.

In contrast, there could be a "full delegation" PBE in which P always delegates. This PBE exists if and only if P prefers to delegate the most important project, taking into account A's optimal effort.

Lemma C.16. *A PBE exists for Game L in which P always delegates if and only if* $\overline{u}_D(\alpha^U) \geq u_C(\alpha^U)$. *The PBE admits the following strategies and belief:*

$$g = D$$

$$\alpha_\mu = \mathbb{E}_{\mu_0}[\alpha]$$

$$e_A = h\left(\gamma \mathbb{E}_{\mu_0}[\alpha]\right)$$

Proof. $\overline{u}_D\left(\alpha^U\right) \geq u_C\left(\alpha^U\right)$ is necessary. Given that A's belief is α_0 in the PBE, P's incentive not to switch to $g = C$ requires that $\hat{u}_D\left(\alpha^U, \alpha_0\right) = \overline{u}_D\left(\alpha^U\right) \geq u_C\left(\alpha^U\right)$.

$\overline{u}_D\left(\alpha^U\right) \geq u_C\left(\alpha^U\right)$ is also sufficient. It is easy to verify that A's belief is consistent and he would not deviate given his belief. P's effort choice is also optimal. To verify P would not switch to $g = C$, we show that $u_C(\alpha) \leq \hat{u}_D\left(\alpha, \alpha_0\right), \forall \alpha \in \left[\alpha^L, \alpha^U\right]$. There are two cases.

$\alpha \leq \beta$: (which may not exist) $u_C(\alpha) = \beta < \hat{u}_D\left(\alpha, \alpha_0\right)$.

$\alpha > \beta$: $\hat{u}_D\left(\cdot, \alpha_0\right)$ is a line segment with slope smaller than 1, whereas $u_C(\cdot)$ is a line segment with slope 1. Since $u_C\left(\alpha^U\right) \leq \hat{u}_D\left(\alpha^U, \alpha_0\right)$, it must be that $u_C(\alpha) \leq \hat{u}_D\left(\alpha, \alpha_0\right)$.

□

In this PBE, A expects to be delegated all projects and exerts effort accordingly. P prefers to delegate all projects given A's insufficient effort, and A's belief and effort choice are therefore justified.

C.3.3. Commitment game

We prove Proposition 3.5 that the principal delegates projects over an interval in the commitment game.[4] The rough intuition for the interval result is that, on the interval

[4] Even though the cases of either $\alpha^L \geq \beta$ and $\alpha^U \leq \beta$ are ruled out by the assumption that $\alpha^L < \alpha^\dagger < \alpha^U$, they are however easily characterized when the assumption is not maintained: let $\bar{\alpha}^L$ be α^L in the former case, and $\bar{\alpha}^U$ be α^U in the latter case. All the bounds are well defined in this way.

$[\alpha^L, \beta]$, the change in P's payoff from delegating a set of projects *of an arbitrary probability mass* increases in α, and similarly that on the interval $[\beta, \alpha^U]$, the change in P's payoff decreases in α. Therefore it is never optimal to have a gap of a positive probability mass in the set of delegated projects.

We repeat the proposition here.

Proposition. *In the equilibrium of Game C, the set of delegated projects $\bar{\mathcal{A}}_D$ takes the form $[\bar{\alpha}^L, \bar{\alpha}^U]$ where $\bar{\alpha}^L \leq \beta \leq \bar{\alpha}^U$.*

Proof. Notice that the cases of either $\alpha^L \geq \beta$ and $\alpha^U \leq \beta$ are ruled out by the assumption that $\alpha^L < \alpha^\dagger < \alpha^U$.

Fix some probability measure μ with CDF F and PDF f over $[\alpha^L, \alpha^U]$. With abuse of notation, we denote $\mathbb{E}[\mathcal{A}] \equiv \mathbb{E}[\alpha|\mathcal{A}]$ as the expected value of α in the set $\mathcal{A}$ given the probability measure μ.

Existence of $\bar{\alpha}^L \in [\alpha^L, \beta]$. We first prove the existence of the lower bound of the delegation interval $\bar{\alpha}^L \in [\alpha^L, \beta]$. We do so by showing that P strictly benefits from increasing (by a little) an arbitrary open set of delegated projects on this interval. Intuitively, this change in delegated projects increases both the effort from the agent and the average impact of the delegated projects without affecting the payoff from centralized projects. So the effect is unambiguously positive. This implies that it cannot be optimal to have a gap of a positive probability mass in the set of delegated projects on this interval, as we can always find a strict improvement by shifting some delegated projects to the right. Consequently, there must be a lower bound $\bar{\alpha}^L \in [\alpha^L, \beta]$ such that delegated projects of positive probability mass must be above $\bar{\alpha}^L$.

Consider two open subsets $\mathcal{A}_1 \equiv (\alpha_1, \alpha_2) \subset [\alpha^L, \beta]$ and $\mathcal{A}_2 \equiv (\alpha_1 + \varepsilon, \alpha_2 + \delta) \subset [\alpha^L, \beta]$ where $\mu(\mathcal{A}_1) = \mu(\mathcal{A}_2) = m$ and $0 < \varepsilon \leq \alpha_2 - \alpha_1$. Let $\mathcal{A}_0$ be an arbitrary subset of $[\alpha^L, \alpha^U] \backslash (\alpha_1, \alpha_2 + \delta)$, and M_0 be its probability mass. Consider P's payoff change from delegating $\mathcal{A}_0 \cup \mathcal{A}_1$ to delegating $\mathcal{A}_0 \cup \mathcal{A}_2$ (while centralizing the rest).

Denote this payoff change as Δu^L. Recall that P works on the β-impact project when $\alpha \leq \beta$ irrespective of whether she delegates or centralizes the α-impact project. Therefore,

$$\begin{aligned}
\Delta u^L = & \int_{\mathcal{A}_0 \cup \mathcal{A}_2} \left[\alpha e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_2\right]\right) - \beta\right] \mathrm{d}F(\alpha) - \int_{\mathcal{A}_0 \cup \mathcal{A}_1} \left[\alpha e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_1\right]\right) - \beta\right] \mathrm{d}F(\alpha) \\
= & \left[u_D \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_2\right]\right) - u_D \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_1\right]\right)\right] \cdot (M_0 + m) \\
> & 0
\end{aligned}$$

where $u_D(\alpha) \equiv \alpha e_A(\alpha) + \beta$ is P's payoff from delegating the α-impact project as defined in Appendix C.3.1. Given that u_D is increasing in α and that $\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_2\right] > \mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_1\right]$, Δu^L is always strictly positive.

Existence of $\bar{\alpha}^U \in [\beta, \alpha^U]$. We now prove the existence of the upper bound $\bar{\alpha}^U \in [\beta, \alpha^U]$. We do so by showing that P is strictly harmed by increasing *infinitesimally* an arbitrary set of delegated projects on this interval. This implies that it cannot be optimal to have a gap of a positive probability mass in the set of delegated projects on this interval, as we can always find an improvement by shifting some delegated projects to the left. Consequently, there must be an upper bound $\bar{\alpha}^U \in [\beta, \alpha^U]$ such that such that delegated projects of positive probability mass must be below $\bar{\alpha}^U$.

This proof depends on a result weaker than the previous one and is thus more involved because of the ambiguous effect of lowering a set of delegated projects on $[\beta, \alpha^U]$. It decreases the effort from the agent and but increases the payoff from high-impact projects because delegation is costly compared to centralization.

Consider two open subsets $\mathcal{A}_3 \equiv (\alpha_3, \alpha_4) \subset [\beta, \alpha^U]$ and $\mathcal{A}_4 \equiv (\alpha_3 + \varepsilon, \alpha_4 + \delta) \subset [\beta, \alpha^U]$ where $\mu(\mathcal{A}_3) = \mu(\mathcal{A}_4)$ and $0 < \varepsilon \leq \alpha_4 - \alpha_3$. Let $\mathcal{A}_0$ be an arbitrary subset of $[\alpha^L, \alpha^U] \backslash (\alpha_3, \alpha_4 + \delta)$, and Φ be its probability mass. Consider P's payoff change from delegating $\mathcal{A}_0 \cup \mathcal{A}_3$ to delegating $\mathcal{A}_0 \cup \mathcal{A}_4$ (while centralizing the rest). We are interested in this change as $\varepsilon \to 0$.

Denote this payoff change as $\Delta u^U(\varepsilon)$. Recall that P works on the α-impact project when $\alpha \geq \beta$ if she centralizes it. Therefore,

$$\begin{aligned}
\Delta u^U(\varepsilon) = & \int_{\mathcal{A}_0 \cup \mathcal{A}_4} \left[\alpha e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_4\right]\right) - \alpha\right] \mathrm{d}F(\alpha) - \int_{\mathcal{A}_0 \cup \mathcal{A}_3} \left[\alpha e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_3\right]\right) - \alpha\right] \mathrm{d}F(\alpha) \\
= & \int_{\mathcal{A}_0 \cup [\alpha_3 + \varepsilon, \alpha_4]} \alpha \left[e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_4\right]\right) - e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_3\right]\right)\right] \mathrm{d}F(\alpha) \\
& + \int_{\alpha_3}^{\alpha_3 + \varepsilon} \left[\alpha - \alpha e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_3\right]\right)\right] \mathrm{d}F(\alpha) + \int_{\alpha_4}^{\alpha_4 + \delta} \left[\alpha e_A \left(\mathbb{E}\left[\mathcal{A}_0 \cup \mathcal{A}_4\right] - \alpha\right)\right] \mathrm{d}F(\alpha)
\end{aligned}$$

We will show that

$$\lim_{\varepsilon \to 0} \frac{\Delta u^U(\varepsilon)}{\varepsilon} < 0$$

Let $\phi = F(\alpha_3 + \varepsilon) - F(\alpha_3) = F(\alpha_4 + \delta) - F(\alpha_4)$, and define the offset $z(\alpha, \phi)$ by $\phi = F(\alpha + z) - F(\alpha)$. Therefore $\varepsilon = z(\alpha_3, \phi)$ and $\delta = z(\alpha_4, \phi)$. Define a function mapping ϕ to the set $\mathcal{A}(\phi) = \mathcal{A}_0 \cup [\alpha_3 + z(\alpha_3, \phi), \alpha_4]$. Define a function mapping (α, ϕ) to the interval $\mathcal{A}_\phi(\alpha, \phi) = (\alpha, \alpha + z(\alpha, \phi))$. Notice that $\mathcal{A}_\phi(\alpha_3, \phi) = (\alpha_3, \alpha_3 + \varepsilon)$ and $\mathcal{A}_\phi(\alpha_4, \phi) = (\alpha_4, \alpha_4 + \delta)$.

Let $\Delta u^P(\alpha,\phi)$ be the change in P's payoff when she delegates $\mathcal{A}_\phi(\alpha,\phi)$ in addition to $\mathcal{A}(\phi)$. Then

$$\Delta u^P(\alpha,\phi) = u_D\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha,\phi)\right]\right)\cdot\Phi - u_D\left(\mathbb{E}\left[\mathcal{A}(\phi)\right]\right)\cdot(\Phi-\phi) - \mathbb{E}\left[\mathcal{A}_\phi(\alpha,\phi)\right]\cdot\phi$$

Notice that

$$\begin{aligned}\Delta u^U(\varepsilon) = & \int_{\mathcal{A}(\phi)} \alpha e_A\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha_4,\phi)\right]\right)\mathrm{d}F(\alpha)\\ & + \int_{\mathcal{A}_\phi(\alpha_4,\phi)} \left[\alpha e_A\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha_4,\phi)\right]-\alpha\right)\right]\mathrm{d}F(\alpha)\\ & - \int_{\mathcal{A}(\phi)} \alpha e_A\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha_3,\phi)\right]\right)\mathrm{d}F(\alpha)\\ & - \int_{\mathcal{A}_\phi(\alpha_3,\phi)} \left[\alpha e_A\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha_3,\phi)\right]\right)-\alpha\right]\mathrm{d}F(\alpha)\\ = & \quad \Delta u^P(\alpha_4,\phi) - \Delta u^P(\alpha_3,\phi)\end{aligned}$$

And therefore, using l'Hôpital's rule with

$$\frac{\partial z(\alpha,\phi)}{\partial\phi} = \frac{1}{f(\alpha+z(\alpha,\phi))}$$

we derive

$$\begin{aligned}&\lim_{\varepsilon\to0}\frac{\Delta u^U(\varepsilon)}{\varepsilon}\\ =&\lim_{\phi\to0}\left[\frac{\phi}{z(\alpha_3,\phi)}\cdot\frac{\Delta u^P(\alpha_4,\phi)-\Delta u^P(\alpha_3,\phi)}{\phi}\right]\\ =&f(\alpha_3)\cdot\lim_{\phi\to0}\frac{\Delta u^P(\alpha_4,\phi)-\Delta u^P(\alpha_3,\phi)}{\phi}\end{aligned}$$

We now derive $\lim_{\phi\to0}\frac{\Delta u^P(\alpha_4,\phi)-\Delta u^P(\alpha_3,\phi)}{\phi}$.

Let $U_D(\alpha, \phi, x) \equiv u_D\left(\mathbb{E}\left[\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)\right]\right) \cdot (\Phi - \phi + x)$, i.e., P's payoff from delegating projects on the set $\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)$. Then

$$\Delta u^P(\alpha, \phi) = U_D(\alpha, \phi, \phi) - U_D(\alpha, \phi, 0) - \mathbb{E}\left[\mathcal{A}_\phi(\alpha, \phi)\right] \cdot \phi$$

Divide both sides by the probability ϕ and take the limit

$$\begin{aligned}\lim_{\phi \to 0} \frac{\Delta u^P(\alpha, \phi)}{\phi} &= \lim_{\phi \to 0} \frac{U_D(\alpha, \phi, \phi) - U_D(\alpha, \phi, 0)}{\phi} - \lim_{\phi \to 0} \mathbb{E}\left[\mathcal{A}_\phi(\alpha, \phi)\right] \\ &= \left.\frac{\partial U_D(\alpha, 0, x)}{\partial x}\right|_{x=0} - \alpha\end{aligned}$$

By the chain rule,

$$\frac{\partial U_D(\alpha, \phi, x)}{\partial x} = \frac{\partial u_D\left(\mathbb{E}\left[\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)\right]\right)}{\partial x} \cdot (\Phi - \phi + x) + u_D\left(\mathbb{E}\left[\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)\right]\right)$$

Given that

$$\begin{aligned}&\frac{\partial u_D\left(\mathbb{E}\left[\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)\right]\right)}{\partial x} \\ &= u_D'\left(\mathbb{E}\left[\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)\right]\right) \\ &\quad \cdot \left(\frac{[\alpha + z(\alpha, \phi)] \cdot f(\alpha + z(\alpha, \phi))}{\Phi - \phi + x} \cdot \frac{\partial z(\alpha, x)}{\partial x} - \frac{\mathbb{E}\left[\mathcal{A}(\phi) \cup \mathcal{A}_\phi(\alpha, x)\right]}{\Phi - \phi + x}\right)\end{aligned}$$

and

$$\frac{\partial z(\alpha, x)}{\partial x} = \frac{1}{f(\alpha + z(\alpha, x))}$$

we derive

$$\begin{aligned}
&\frac{\partial U_D(\alpha,\phi,x)}{\partial x} \\
=\;& u_D'\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha,x)\right]\right)\cdot\left(\alpha+z(\alpha,\phi)-\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha,x)\right]\right) \\
&+u_D\left(\mathbb{E}\left[\mathcal{A}(\phi)\cup\mathcal{A}_\phi(\alpha,x)\right]\right)
\end{aligned}$$

Hence

$$\begin{aligned}
&\lim_{\phi\to 0}\frac{\Delta u^P(\alpha,\phi)}{\phi} \\
=&u_D'\left(\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)\cdot\left(\alpha-\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)+u_D\left(\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)-\alpha \\
=&\left[u_D'\left(\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)-1\right]\cdot\alpha+C
\end{aligned}$$

where

$$C=-u_D'\left(\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)\cdot\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]+u_D\left(\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)$$

is independent of α.

Therefore,

$$\lim_{\varepsilon\to 0}\frac{\Delta u^U(\varepsilon)}{\varepsilon}=f(\alpha_3)\cdot\left[u_D'\left(\mathbb{E}\left[\mathcal{A}_0\cup[\alpha_3,\alpha_4]\right]\right)-1\right]\cdot(\alpha_4-\alpha_3)$$

So $\lim_{\varepsilon\to 0}\frac{\Delta u^U(\varepsilon)}{\varepsilon}<0$ as $u_D'(\cdot)<1$ as shown in Appendix C.3.1. □

We now assume that P's payoff from such interval delegation is concave in $\bar{\alpha}^U$ so that the first-order condition over $\bar{\alpha}^U$ partially characterizes the optimal delegation interval(s).[5] In this case, we can show that the upper bound of the delegation interval must be above the threshold in Game L.

Proposition C.3. *Let $\bar{u}(\bar{\alpha}^L, \bar{\alpha}^U)$ be the principal's payoff from delegating projects on the interval $\left[\bar{\alpha}^L, \bar{\alpha}^U\right]$, and assume that $\bar{u}$ is concave in $\bar{\alpha}^U$. Then the upper bound of the delegation interval must be strictly above α^*.*

Proof. Recall that $\mathcal{A}_D = \left[\alpha^L, \alpha^*\right]$ is the equilibrium set of delegated projects in Game L, and let $\bar{\mathcal{A}}_D = \left[\bar{\alpha}^L, \bar{\alpha}^U\right]$ be an optimal set of delegated projects in Game C.

Fix $\bar{\alpha}^L \geq \alpha^L$ and let $\bar{\mathcal{A}}_D(\bar{\alpha}^U)$ map $\bar{\alpha}^U$ to the set of delegated projects. With abuse of notation, let $\bar{u}(\bar{\alpha}^U)$ map $\bar{\alpha}^U$ to P's payoff in Game C:

$$\bar{u}(\bar{\alpha}^U) = \beta \cdot F(\bar{\alpha}^L) + \int_{\bar{\alpha}^L}^{\bar{\alpha}^U} [e_A(\mathbb{E}[\bar{\mathcal{A}}_D(\bar{\alpha}^U)]) \cdot \alpha + \beta]\mathrm{d}F(\alpha) + \int_{\bar{\alpha}^U}^{\alpha^U} \alpha \mathrm{d}F(\alpha)$$

Notice that

$$\frac{\partial \mathbb{E}[\bar{\mathcal{A}}_D(\bar{\alpha}^U)]}{\partial \bar{\alpha}^U} = \frac{\partial}{\partial \bar{\alpha}^U} \frac{\int_{\bar{\mathcal{A}}_D} \alpha \mathrm{d}F(\alpha)}{\mu(\bar{\mathcal{A}}_D)} = \frac{(\bar{\alpha}^U - \mathbb{E}[\bar{\mathcal{A}}_D]) \cdot f(\bar{\alpha}^U)}{\mu(\bar{\mathcal{A}}_D)}$$

[5] This is the case, for instance, when projects follow a uniform distribution and the effort cost is quadratic, e.g., $c(e) = \frac{e^2}{2}$. Specifically, we can show that $\bar{u}''$ as defined in the following proof is

$$\begin{aligned} &\bar{u}''(\bar{\alpha}^U) \\ =&\frac{4\gamma\mathbb{E}[\bar{\mathcal{A}}_D(\bar{\alpha}^U)] - 4}{4(\alpha^U - \alpha^L)} \\ <&0 \end{aligned}$$

where $\gamma \leq 1/2$ by the assumption that $u_D'(1) \leq 1$.

Therefore, at $\bar{\mathcal{A}}_D = \bar{\mathcal{A}}_D(\bar{\alpha}^U)$,

$$
\begin{aligned}
&\bar{u}'(\bar{\alpha}^U) \\
=&-\bar{\alpha}^U \cdot f(\bar{\alpha}^U) + [e_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \bar{\alpha}^U + \beta] \cdot f(\bar{\alpha}^U) + e'_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \frac{\partial \mathbb{E}[\bar{\mathcal{A}}_D]}{\partial \bar{\alpha}^U} \cdot \mathbb{E}[\bar{\mathcal{A}}_D] \cdot \mu(\bar{\mathcal{A}}_D) \\
=&[e_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \bar{\alpha}^U - \bar{\alpha}^U + \beta] \cdot f(\bar{\alpha}^U) + e'_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \frac{\partial \mathbb{E}[\bar{\mathcal{A}}_D]}{\partial \bar{\alpha}^U} \cdot \mathbb{E}[\bar{\mathcal{A}}_D] \cdot \mu(\bar{\mathcal{A}}_D) \\
=&f(\bar{\alpha}^U) \cdot \left\{[e_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \bar{\alpha}^U - \bar{\alpha}^U + \beta] + e'_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot (\bar{\alpha}^U - \mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \mathbb{E}[\bar{\mathcal{A}}_D]\right\}
\end{aligned}
$$

Given that P is indifferent between centralizing and delegating α^*-impact project in Game L, $\alpha^* = \beta + e_A\left(\mathbb{E}\left[\mathcal{A}_D\right]\right) \cdot \alpha^*$. Now we derive $\bar{u}'(\bar{\alpha}^U)$ at $\bar{\alpha}^U = \alpha^*$:

$$
\begin{aligned}
&u'(\alpha^*) \\
=&f(\alpha^*) \cdot \left\{\left[e_A(\mathbb{E}[\bar{\mathcal{A}}_D]) - e_A\left(\mathbb{E}\left[\mathcal{A}_D\right]\right)\right] \cdot \alpha^* + e'_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot (\alpha^* - \mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \mathbb{E}[\bar{\mathcal{A}}_D]\right\} \\
\geq&f(\alpha^*) \cdot e'_A(\mathbb{E}[\bar{\mathcal{A}}_D]) \cdot (\alpha^* - \mathbb{E}[\bar{\mathcal{A}}_D]) \cdot \mathbb{E}[\bar{\mathcal{A}}_D] \\
>&0
\end{aligned}
$$

where the first inequality follows the result that $\mathbb{E}[\bar{\mathcal{A}}_D(\alpha^*)] \geq \mathbb{E}[\mathcal{A}_D]$, and the second inequality follows from the result that $\alpha^* > \bar{\mathcal{A}}_D(\alpha^*)$ since it is the upper bound of $\bar{\mathcal{A}}_D(\alpha^*)$ with positive probability mass.

Given that the upper bound of an optimal delegation interval must satisfy $\bar{u}' = 0$, and since the concavity assumption implies that $\bar{u}'$ is weakly decreasing, we conclude that the upper bound must be strictly above α^*. □