Competing Powers*

Erik Madsen† Andrea Prat‡

March 18, 2026

Abstract

Motivated by the growing interest in geoeconomics, we develop a formal framework in which multiple superpowers—such as the United States and China—compete to shape the behavior of less powerful countries or firms. We model superpowers as competing principals who influence an agent's policy choice by threatening costly punishments. We characterize the set of equilibrium policy choices as a system of incentive-compatibility constraints on each actor. Equilibrium policy choices tend to favor more powerful principals, and larger policy concessions can be extracted when principals are more aligned. When all parties additionally have access to frictionless transfers, Coasian surplus-maximizing equilibria always exist regardless of the distribution of punishment power, with power impacting only the division of surplus. However, this result no longer holds if frictions are introduced.

Keywords: Geoeconomics, principal-agent models, competing principals, costly punishments, foreign policy

*We thank Pedro Dal Bó, Rohan Dutta, Francesco Fabbri, Debraj Ray, Kareen Rozen, Jesse Schreger, and seminar participants at Brown University, McGill University, and NYU for helpful comments and suggestions. We gratefully acknowledge Zihao Li and Cole Wittbrodt for their excellent research assistance.

†Department of Economics, New York University (emadsen@nyu.edu)

‡Columbia Business School and Department of Economics, Columbia University (andrea.prat@columbia.edu)

1 Introduction

Hegemonic powers, like the United States and China, can use their economic and military power to influence the policies of other countries. In recent years, economists have begun to model and measure this influence process in the rapidly growing literature on geoeconomics. Work in this literature (which we review below) models geopolitical influence as a principal-agent relationship, with the principal (a hegemonic country) making policy demands on one or more agents (other countries or firms) and credibly threatening them with negative consequences if they do not comply. A targeted agent chooses between complying with the policy demand or enduring the threatened consequences, yielding an obedience constraint on implementable demands. By analyzing the threats and demands which arise in equilibrium, this literature has shed light on the sources and consequences of geoeconomic power.

Existing work in geoeconomics has focused on settings involving a singl principle. However, today's world contains at least two countries competing to assert hegemonic power: China and the United States. As Clayton, Maggiori, and Schreger (2025b) observe, it would therefore be useful to develop a theoretical framework that allows for multiple hegemons. How do firms and countries respond when targeted by competing powers? How do hegemons wield their power when they are competing with other hegemons? How does the outcome of hegemonic competition depend on the set of influence tools available?

To answer these questions, we develop a model with several principals and a single agent. The agent chooses a policy that affects both his own payoff and the payoffs of the each principal. The policy space, which can be multi-dimensional, encompasses any political, economic or financial variable under the agent's control that affects the principals' payoff. The agent has an ideal policy within this space and suffers a loss from selecting another policy.

Each principal controls a set of *punishments* that she can impose on the agent in order to extract policy concessions. A punishment is any tool under the principal's control that imposes a cost on the agent (and, potentially, the principal). For instance, punishments could include military strikes, sanctions, exclusion from payment systems, or restrictions on trade or investment. Each principal possesses a limited set of credible punishments, and we measure the *power* of a principal by the maximal cost she can credibly threaten to impose on the agent. The principals simultaneously

announce *threats*, which commit each principal to a schedule of punishments as a function of the policy chosen by the agent. After observing these threats, the agent selects a policy and the principals carry out any punishments triggered by the choice.

The main result of the paper (Theorem 1) characterizes the policy choices that can be implemented via a subgame-perfect equilibrium with two principals. It reduces implementation to a small set of incentive-compatibility (IC) constraints on the three actors. The agent's IC constraint simply requires that his loss not exceed the largest punishment that can be inflicted by the two principals working in concert. Meanwhile, the principals' IC constraints require that their payoffs exceed particular reservation values, which are endogenously determined by the magnitude of the agent's loss. Reservation values are explicit functions of model primitives and can be used to efficiently compute equilibria in a general class of games with multiple hegemons.[1]

Conceptually, our characterization divides the set of implementable policies into two distinct regimes. The first contains *non-collaborative* outcomes, which could have been obtained by the more powerful principal acting alone. The second encompasses *collaborative* outcomes, which require more punishment power than either principal could individually bring to bear. There always exist non-collaborative outcomes which can be sustained in equilibrium. By contrast, collaborative outcomes involve a discontinuous jump in the weaker principal's reservation value and can be implemented only if the two principals are sufficiently aligned.

Our analysis formally verifies two intuitive claims regarding competition between principals. First, outcomes tend to be skewed in the direction of the more powerful principal. Each principal obtains her reservation value by using threats to push the agent toward alternative policies. The range of policies she can induce depends both on her own power as well as on the size of the agent's concession. More powerful principals can attain a wider range of policies and enjoy a greater reservation value.

Second, principals can extract greater concessions from the agent when their preferences are more aligned. When principals' preferences are sufficiently misaligned, all implementable policies are non-collaborative. When their preferences are sufficiently aligned, non-collaborative outcomes become implementable, and the maximum concession that can be extracted grows as alignment increases. If alignment is sufficiently strong, then there exist implementable outcomes which extract all of the agent's rent,

[1]The authors' homepage contains the beta version of a tool to compute and visualize equilibrium sets.

in the sense that the IC constraint of the agent binds and no set of threats could extract further concessions.

Our baseline model allows principals to influence the agent using punishments but not transfers. The key distinction between these two instruments is that punishments move the payoffs of the agent and principal in the same direction; while transfers move payoffs in opposite directions. The most obvious example of a transfer is a monetary payment to the agent. But transfers could also take indirect forms which imperfectly distribute surplus, such as foreign aid, credit, or preferential access to products or resources.[2]

Competition between principals using monetary transfers (without punishments) has been studied under the guise of menu auctions (Bernheim and Whinston 1986; Grossman and Helpman 1994), which are often used by political economists to model the lobbying process. To create a bridge between that setting and ours, we extend our model to allow principals to promise both punishments and transfers. We prove two results.

First, if transfers are frictionless—utility can be perfectly transferred between the agent and each principal—then transfers and punishments do not interact as tools for influence. More precisely, any policy which can be implemented without punishments can also be implemented with punishments, regardless of the power of each principal. In particular, a well-known result regarding menu auctions implies that any policy maximizing the sum of payoffs of all players can be implemented. Furthermore, there exist "orthogonalized" equilibria implementing these policies, with transfers to the agent used to influence the policy choice and punishments used to extract payments from the agent. Within this class of equilibria, punishment power impacts the division of surplus but not the implemented policy.

The second result instead investigates what happens if transfers are not frictionless. In that case, surplus-maximizing policies may not be implementable, and the maximum achievable surplus may be affected by the distribution of power among principals. When the friction is sufficiently large, principals use threats exclusively and the equilibrium policy set is the same as in our baseline (threats-only) setting.

[2]Our use of the term "punishment" takes the high-payoff "no punishment" outcome as a reference point. In particular applications, discussions may instead take a low-payoff outcome as a reference, making punishments appear to be rewards: For instance, financial network inclusion or market access may be offered in return for a policy concession. Whenever such offers would benefit the principal to implement, they can be equivalently reframed as punishments.

Related Literature

Important contributions to the geoeconomics literature include Clayton, Maggiori, and Schreger (2025a,b), Thoenig (2023), Broner et al. (2025), Konrad (2024), and Kleinman, Liu, and Redding (2024). Most of these papers focus on influence by a single principal. One exception is Clayton, Maggiori, and Schreger (2025a, Appendix B.2), which studies a 2-principal setting under the restriction that there are no "z-externalities" (i.e., the principal's utility function does not include non-market outcomes) and only revenue-neutral wedges and transfers are used. As that paper discusses, however, a general model with multiple principals and costly punishment is missing. In our setting, by contrast, we place no restrictions on how the agent's policy enters the principals' payoff functions.[3]

Beyond the geoeconomics literature, there is a long tradition of modeling foreign influence on domestic policy. Aidt, Albornoz, and Hauk (2018) provides a comprehensive survey and a one-principal model with both transfers and sanctions. And in the realm of political economy, Dal Bó, Dal Bó, and Di Tella (2006) develop a political-economy model in which an interest group can use both bribes and threats to influence a political agent. The paper characterizes the equilibrium choice between the two instruments and examines the effect on corruption and politicians' quality of granting officials with immunity from prosecution.

As discussed above, our paper is also closely related to studies of menu auctions. Bernheim and Whinston (1986) and Grossman and Helpman (1994) analyzed a benchmark setting with a single agent and perfectly-transferrable utility. Later work extended the analysis to nonlinear utilities (Dixit, Grossman, and Helpman 1997) and multiple agents (Prat and Rustichini 2003). The game in our paper is very similar—in terms of player set, information, timing, and payoffs—to a menu auction. The key difference is that our principals offer punishments (tools that move the payoffs of the principal and the agent in the same direction) rather than transfers (tools that move the payoffs of the two parties in opposite directions). In Section 4 we allow for both punishments and transfers, allowing us to explicate the connection between our paper and a classic menu auction.

Finally, our paper joins a literature in contract theory interested in non-monetary

[3]Other aspects of foreign influence have also been modeled. For instance, Antràs and Padró i Miquel (2025) focus on foreign spending in elections, which impacts the policy preferences of the elected government.

incentive tools. Much of this literature has focused on the use of such tools to motivate hidden effort. For instance, Chwe (1990) studies a setting where a principal can impose costly punishments alongside incentive payments. Madsen, Williams, and Skrzypacz (2025) allow the principal to make intertemporal promises alongside immediate payments, opening a channel for punishment via threats to forego future collaboration. Miller and Rozen (2014) consider a setting with no rewards and unlimited costly punishments. And Dutta, Levine, and Modica (2018) analyze a 2-principal setting with a single incentive tool which, depending on model parameters, functions either a reward or a punishment. Compared to these papers, we abstract from noisy monitoring of the agent's action and forego systematic comparisons between punishment and (imperfect) monetary transfers as incentive tools. We instead focus on how the set of implementable actions is shaped by competition between multiple principals.

2 Model

2.1 Setup

$N \geq 1$ principals compete to influence the policy choice made by an agent. The policy space is modeled by a compact subset of some Euclidean space $X \subset \mathbb{R}^n$, from which the agent chooses a single policy $x \in X$. This policy choice is payoff relevant for all parties. We assume that the agent has an ideal policy within this space, which we label 0 as a convention. He incurs a loss from deviating from this policy: selecting policy $x \in X$ incurs a loss $\ell(x) \geq 0$, with $\ell(0) = 0$. Meanwhile, each principal $i \in \{1, ..., N\}$ enjoys a payoff $g_i(x)$ from policy $x \in X$. We assume that ℓ and each g_i are continuous on X. We also impose a richness condition on the set of policies: For every $\epsilon > 0$, there exist distinct policies $x^1, ..., x^N$ satisfying $\ell(x^i) < \epsilon$ for each i. This condition is satisfied if 0 is an accumulation point of X; or, alternatively, if there exist N distinct policies satisfying $\ell(x) = 0$.

The principals influence the agent's decision by threatening to inflict costly punishments. Specifically, each principal is capable of inflicting a utility loss $p_i \in [0, \Pi_i]$ on the agent at a cost $c_i(p_i) \geq 0$. The upper bounds $\mathbf{\Pi} = (\Pi_1, ..., \Pi_N) > 0$ are exogenous limits on allowed punishments. We assume that principals have heterogeneous punishment power, and without we loss order them so that $\Pi_1 > \Pi_2 > ... > \Pi_N$.

We make no assumptions on c_i other than that $c_i(0) = 0$. (In particular, we allow for costless punishments as a special case.) Given punishments $\boldsymbol{p} = (p_1, ..., p_N)$ and policy x, the agent's net loss is:[4]

$$L(x, \boldsymbol{p}) = \ell(x) + \sum_{i=1}^{N} p_i$$

while principal i's net gain is

$$G_i(x, p_i) = g_i(x) - c_i(p_i).$$

We impose the following richness condition on principal payoffs: There exist distinct, nonzero policies $\underline{x}^2, ..., \underline{x}^N$ such that $\ell(\underline{x}^i) \leq \sum_{j \neq i} \Pi_j - \Pi_i$ for each $i \geq 2$ and $g_i(\underline{x}^j) \leq g_i(0)$ for all $i, j \geq 2$.

The timing of the interaction between principals and agent is as follows: First, all principals simultaneously commit to *threats* $\pi_i \in \mathcal{P}_i$, where $\mathcal{P}_i$ is the set of lower semicontinuous functions from X to $[0, \Pi_i]$.[5] Second, the agent observes all threats and selects a policy. Third, threats are carried out and payoffs are realized.

A (pure) *strategy profile* of this game is a triple $(\boldsymbol{\pi}, \chi)$ consisting of a profile of threats $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ for each principal along with a response function $\chi : \prod_{i=1}^{N} \mathcal{P}_i \to X$ for the agent. In a slight abuse of notation, we will let $L(x, \boldsymbol{\pi}) \equiv L(x, \boldsymbol{\pi}(x))$. A strategy profile $(\boldsymbol{\pi}^*, \chi)$ is a (pure-strategy) *equilibrium* if

$$\chi(\boldsymbol{\pi}) \in \arg\min_{x \in X} L(x, \boldsymbol{\pi})$$

for every $\boldsymbol{\pi} \in \prod_{i=1}^{N} \mathcal{P}_i$[6] and

$$\pi_i^* \in \arg\max_{\pi_i \in \mathcal{P}_i} G_i(\chi(\pi_i, \boldsymbol{\pi}_{-i}^*), \pi_i(\chi(\pi_i, \boldsymbol{\pi}_{-i}^*)))$$

for each principal i. This equilibrium notion enforces subgame perfection for the agent,

[4] To keep notation manageable, we assume that the agent's payoff is separable in the two principals' payoffs. The analysis can be extended to a non-separable payoff function as long as we maintain monotonicity.

[5] Lower semicontinuity ensures existence of agent best responses. Lemma A.1 in the Appendix characterizes a flexible class of l.s.c. threats which are sufficient for contructing equilibria.

[6] Since the policy space is compact and punishment schedules are l.s.c., the agent has at least one best response to any threat profile.

since his response must be optimal even off the equilibrium path.

An equilibrium yields an implemented policy as well as realized punishments. On-path punishments, however, are not useful for expanding the set of implementable policies. Intuitively, punishing the implemented policy only makes that policy more vulnerable to deviations from other principals. The following lemma formally verifies this claim.

Lemma 1. *Fix any $x^* \in X$, and suppose that $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium satisfying $\chi(\boldsymbol{\pi}^*) = x^*$. Then there exists an equilibrium $(\boldsymbol{\pi}', \chi')$ satisfying $\chi'(\boldsymbol{\pi}') = x^*$ and $\boldsymbol{\pi}'(x^*) = \mathbf{0}$.*

Suppose further that $c_i(p_i) > 0$ when $p_i > 0$ for every principal i. Then $\boldsymbol{\pi}^*(x^*) = \mathbf{0}$.

In light of this result, we focus on equilibria involving no on-path punishment. Given an equilibrium $(\boldsymbol{\pi}^*, \chi)$, we refer to $\chi(\boldsymbol{\pi}^*)$ as the equilibrium *outcome*. We say that an outcome $x \in X$ is *implementable* if there exists an equilibrium with outcome x, and we say that threats $\boldsymbol{\pi}^*$ *implement* x if there exists an equilibrium $(\boldsymbol{\pi}^*, \chi)$ with outcome x.

2.2 Interpretation

Our model builds on the interaction between a hegemon (the principal) and a foreign entity (the agent) analyzed in Clayton, Maggiori, and Schreger (2025b, Section 3.2). In that model, the foreign entity controls two (vectors) of variables, one related to private economic consumption, and another capturing geopolitical action. The policy choice in our model simultaneously captures both sets of outcomes. To focus our analysis on the interaction of multiple principals, our baseline setup has one agent only.[7]

The policies we contemplate can take many forms, such as the provision of public goods (military spending, public procurement, participation to consortia) or regulation of activities of citizens and firms (prohibition to trade with certain entities, disclosure requirements, etc.). Meanwhile, the punishments we contemplate can range from military attack or withholding of military protection to a vast array of geoeconomic tools such as export or import restrictions, exclusion from beneficial industrial policy, financial sanctions, and exclusion from multilateral organizations (Clayton, Maggiori, and Schreger 2025b).

[7] We plan to provide a multiple-agent multiple-principal extension.

3 Main Results

In this section, we characterize the set of implementable policies in the presence of $N = 2$ principals and explore important properties of this set.

3.1 Implementable set

Suppose that $N = 2$ principals are present. We now characterize the set of outcomes that are implementable. For any $\lambda \geq 0$, define

$$\overline{g}_i(\lambda) \equiv \sup\{g_i(x) \ : \ \ell(x) < \lambda\}, \quad \underline{g}_i(\lambda) \equiv \inf\{g_i(x) \ : \ \ell(x) \leq \lambda\}$$

to be the best and worst payoffs obtainable by each principal, subject to the agent's loss being no more than λ.[8] Also let $\Delta\Pi \equiv \Pi_1 - \Pi_2$.

The following theorem characterizes lower bounds on equilibrium payoffs for each principal, along with an upper bound on the agent's loss, which are necessary and sufficient for implementability. We will refer to these bounds as the IC constraints on implementability.

Theorem 1. *Fix $x^* \in X$. Then:*

- *If $\ell(x^*) \leq \Pi_1$, then x^* is implementable if and only if*

$$g_1(x^*) \geq \overline{g}_1(\Delta\Pi), \quad g_2(x^*) \geq \underline{g}_2(\Delta\Pi).$$

- *If $\ell(x^*) \in (\Pi_1, \Pi_1 + \Pi_2]$, then x^* is implementable if and only if*

$$g_1(x^*) \geq \overline{g}_1(\ell(x^*) - \Pi_2), \quad g_2(x^*) \geq \overline{g}_2(\ell(x^*) - \Pi_1).$$

- *If $\ell(x^*) > \Pi_1 + \Pi_2$, then x^* is not implementable.*

The sole IC constraint on the agent is that his equilibrium loss be no more than $\Pi_1 + \Pi_2$, the largest punishment that can be inflicted by the two principals working in concert. Were this bound breached, the agent could profitably deviate to his ideal policy regardless of the threats he faced. On the other hand, whenever the bound

[8] For technical reasons, the supremum must exclude outcomes involving a loss of exactly λ.

is satisfied, there exist threats under which the agent cannot profitably deviate. In particular, *grim trigger* threats, under which both principals punish any deviation from x^* as harshly as possible, enforce any x^* satisfying $\ell(x^*) \leq \Pi_1 + \Pi_2$. However, as we will discuss shortly, grim trigger threats are not capable of sustaining most equilibrium outcomes due to IC constraints on the principals.

The remaining IC constraints take the form of *reservation values* which must be delivered to each principal. Delivering each principal her reservation value ensures that she cannot implement any alternative policy that she prefers to x^*. The reservation values are derived by identifying a deviation set of policies which each principal can implement, and then minimizing this set through a judicious choice of threats by the competing principal.

In general, any outcome x' that a principal can achieve can be achieved by imposing no punishment for x' and the maximum punishment for all other policies. Were both principals to use grim trigger threats in equilibrium, deviating could implement any policy x' which the agent prefers to both x^* and 0 (the latter being the best policy which does not avoid punishment). In other words, under grim trigger threats the achievable deviations for each principal i are those satisfying

$$\ell(x') + \Pi_{-i} < \min\{\ell(x^*) + \Pi_i, \Pi_1 + \Pi_2\},$$

or $\ell(x') < \min\{\ell(x^*) + \Delta\Pi, \Pi_1\}$ for principal 1 and $\ell(x') < \min\{\ell(x^*) - \Delta\Pi, \Pi_2\}$ for principal 2. Note that these constraints imply higher reservation values than those achieved by Theorem 1.

Reservation values can be reduced through more sophisticated threats. The key is for each principal i to select a *deterrent policy* $\underline{x}_i$, which is unattractive to principal $-i$ and which principal i punishes minimally. Were principal $-i$ to attempt a deviation to policy x' by threatening to punish x^*, she must ensure that this threat does not push the agent to instead deviate to $\underline{x}_i$ (which principal i likes less than x^*) rather than x'. The presence of such a deterrent policy shrinks principal $-i$'s deviation set.

The proof of Theorem 1 establishes that a single deterrent policy is sufficient to minimize each principal's reservation value. Essentially, each principal needs to provide only one attractive alternative to the agent to undermine a given deviation, and the attractiveness of this alternative is independent of the attempted deviation. Hence, the same alternative works to deter all deviations simultaneously. However,

each principal must use a distinct deterrent policy, and the optimal deterrent will depend in general on the outcome being implemented. (The richness conditions we have imposed on the policy space and payoffs ensure that the optimal deterrents for each principal do not coincide.)

The proof further identifies effective deterrents. For principal 2, the agent's ideal policy is always an effective deterrent. Intuitively, this deterrent is as attractive as possible to the agent. The only wrinkle is that it might itself be an attractive policy for principal 1, and so such a deterrent works only for equilibrium policies satisfying $g_1(x^*) \geq g_1(0)$. However, because principal 1 is the more powerful one, she can implement a range of policies no matter 2's punishment schedule. Indeed, the set of policies x' such that $\ell(x') < \Delta\Pi$ can be implemented by principal 1 no matter principal 2's threats. Hence $g_1(x^*) \geq g_1(0)$ must hold in any equilibrium, and there is no loss in choosing 0 as a deterrent. The reservation value for principal 1 obtained in Theorem 1 is derived by characterizing the outcomes she can achieve when principal 2 uses the agent's ideal policy as a deterrent.

A similar logic holds for principal 1's deterrent policy, provided that the auxiliary constraint $g_2(x^*) \geq g_2(0)$ is redundant. When $\ell(x^*) > \Pi_1$, the fact that x^* is the agent's best on-path policy turns out to imply that principal 2 can achieve any deviation outcome x' satisfying $\ell(x') + \Pi_1 < \ell(x^*)$. In that case, $g_2(x^*) \geq g_2(0)$ must hold in any equilibrium. Since principal 2 is already using 0 as a deterrent and some principal must punish 0 enough that it is not attractive on-path, principal 1 must use some other deterrent. Our richness condition ensures that there exist policies x' imposing arbitrarily small losses on the agent, which serve as suitable deterrents. (In particular, if there exists a policy $x' \neq 0$ such that $\ell(x') = 0$, then x' can be used as a deterrent.)

However, when $\ell(x^*) \leq \Pi_1$, the outcome 0 *cannot* necessarily be achieved by principal 2 in equilibrium, and the constraint $g_2(x^*) \geq g_2(0)$ is stronger than necessary. It turns out that, in this regime, principal 1 can implement any outcome satisfying $\ell(x') \leq \Delta\Pi$ no matter principal 2's threats, allowing her to enforce a very harsh deterrent. In particular, by choosing the *worst* outcome in this set as a deterrent, principal 1 can hold principal 2 to a low reservation value.

A key feature of the implementable set is that principal 2's reservation value rises discontinuously as the agent's loss crosses the key threshold Π_1. Intuitively, outcomes satisfying $\ell(x^*) \leq \Pi_1$ are *non-collaborative*, because principal 1 can implement them

without principal 2's support. On the other hand, outcomes satisfying $\ell(x^*) > \Pi_1$ are *collaborative*, because principal 1 needs principal 2's support to implement them. This requirement opens up a new range of deviations for principal 2.

3.2 Impact of power

Theorem 1 has some surprising implications for the impact of power on the implementable set. One implication is that, within the collaborative set, increases in *either* principal's power expand what is implementable. The following result formalizes this claim:

Proposition 1. *Fix power levels* $\mathbf{\Pi}$ *and* $\mathbf{\Pi}' \geq \mathbf{\Pi}$. *For every outcome* $x^* \in X$ *such that* $\ell(x^*) > \Pi_1'$ *and* x^* *is implementable under power levels* $\mathbf{\Pi}$, *the outcome is implementable under power levels* $\mathbf{\Pi}'$.

In other words, so long as collaboration continues to be required, increases in a principal's power do not allow it to block an outcome that was previously achievable. Rather, increased power allows each principal to more effectively deter deviations by the other principal, relaxing incentive constraints on implementability.

Within the non-collaborative set, increases in principal power need not expand the implementable set. In particular, increases in a principal's power tightens her IC constraint within the non-collaborative set, which may eliminate outcomes that were previously implementable.

Nonetheless, the following proposition establishes that implementable outcomes for each principal improve when her power increases, in the sense that there exist implementable outcomes at least as attractive to her as each outcome that was previously implementable.

Proposition 2. *Fix power levels* $\mathbf{\Pi}$ *and* $\mathbf{\Pi}'$ *such that* $\Pi_i' > \Pi_i$ *and* $\Pi_{-i}' = \Pi_{-i}$ *for some principal* i. *For every outcome* $x^* \in X$ *such that* $\ell(x^*) \leq \Pi_1'$ *and* x^* *is implementable under power levels* $\mathbf{\Pi}$, *at least one of the following holds:*

- x^* *is implementable under power levels* $\mathbf{\Pi}'$.
- *There exists an* x^{**} *which is implementable under power levels* $\mathbf{\Pi}'$ *and such that* $g_i(x^{**}) > g_i(x^*)$.

Intuitively, an implementable outcome x^* is ruled out by a principal's increasing power only if her reservation value grows to exceed $g_i(x^*)$. But the reservation value is determined by a particular deviation outcome which the principal can credibly implement. Hence, if the reservation value rules out x^*, then it can itself be implemented to improve on x^*.

An immediate corollary of the previous propositions is the following global comparative static regarding the maximal payoff each principal can achieve in any equilibrium:

Corollary 1. *The best implementable outcome for each principal yields a payoff which is nondecreasing in her power.*

3.3 Worked example

We now illustrate how the characterization in Theorem 1 can be applied to construct the equilibrium set.

Consider a setting with the 2-dimensional policy space $X = \mathbb{R}^2$, the agent loss function $\ell(x) = ||x||_2$, and principal gain functions $g_1(x) = x_1$ and $g_2(x) = \beta x_1 + \sqrt{1-\beta^2}\, x_2$, with $\beta \in [-1, 1]$.[9] In other words, the agent suffers a loss equal to the Euclidean distance from the origin, while each principal benefits linearly from movement along a particular orientation in policy space. As a normalization, we take principal 1's policy orientation to be along the first dimension. The parameter β measures the extent of alignment between the principals' policy orientations, with $\beta = 1$ representing perfect alignment, $\beta = 0$ signifying orthogonal interests, and $\beta = -1$ capturing perfect misalignment.

With linear gain functions and a Euclidean distance loss function, the best- and worst-case payoffs $\overline{g}_i$ and $\underline{g}_i$ are simply $\overline{g}_i(\lambda) = \lambda$ and $\underline{g}_i(\lambda) = -\lambda$. Hence, a policy x^* satisfying $||x^*||_2 \leq \Pi_1$ is implementable iff

$$x_1^* \geq \Delta\Pi, \quad \beta x_1^* + \sqrt{1-\beta^2}\, x_2^* \geq -\Delta\Pi.$$

Meanwhile, a policy satisfying $||x^*||_2 \in (\Pi_1, \Pi_1 + \Pi_2]$ is implementable iff

$$x_1^* \geq ||x^*||_2 - \Pi_2, \quad \beta x_1^* + \sqrt{1-\beta^2}\, x_2^* \geq ||x^*||_2 - \Pi_1.$$

[9] Although the policy space is not compact, it can be compactified without impacting the implementable set by restricting X to a ball with radius Π_1.

Figure 1 plots the regions satisfying these inequalities for several values of β when $\mathbf{\Pi} = (3, 2)$. (These powers are chosen simply for concreteness. All of the qualitative properties described below hold for any choice of $\mathbf{\Pi}$.)

These figures reveal several interesting possibilities regarding the implementable set. First, there may be implementable outcomes which are worse for both principals than some other implementable outcome. In this example, all outcomes in the interior of the implementable set can be improved for both principals without leaving the set. Intuitively, the simultaneous issuance of threats can lead to coordination failures.

Second, the principals' IC constraints may limit the size of the concession that can be extracted from the agent, particularly if their preferences are not well-aligned. While concessions up to the maximal size $\Pi_1 + \Pi_2$ are implementable in the first two cases ($\beta = 1$ and $\beta = 0$), maximal concessions are bounded below this level in the second two ($\beta = -0.6$ and $\beta = -0.97$). Figure 2 further illustrates this point by plotting the largest implementable concession as a function of β. For sufficiently small β, the maximum concession lies below $\Pi_1 + \Pi_2$.

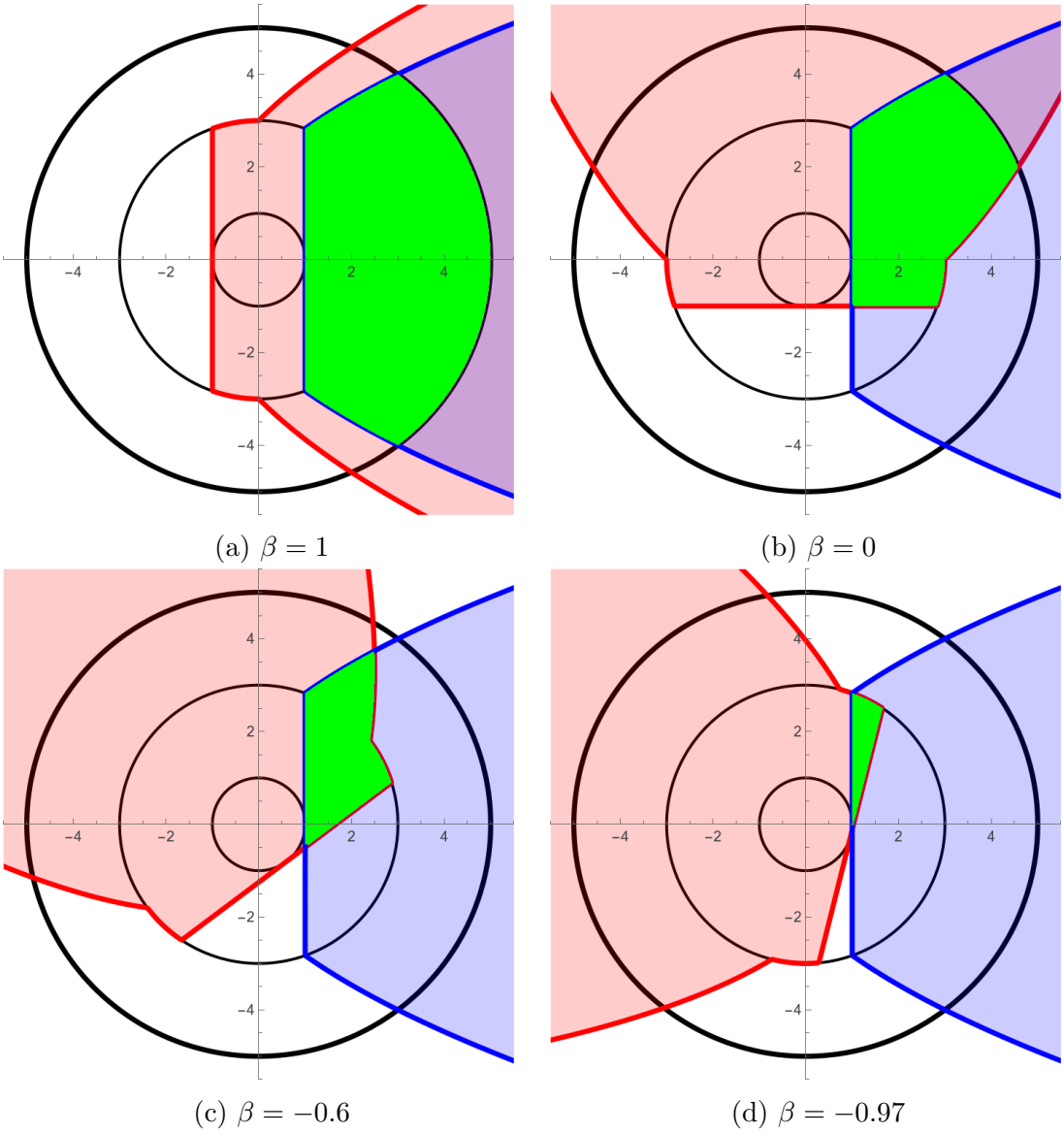
Third, collaborative outcomes may not be possible if the principals are insufficiently aligned. In particular, in the fourth case ($\beta = -0.97$) all implementable outcomes involve a loss no larger than Π_1 and are therefore non-collaborative. Figure 2 further illustrates this point. For sufficiently small β, the maximum concession is Π_1, i.e., non-collaborative.

Fourth, concession-maximizing outcomes need not maximize either principal's payoff within the implementable set. In particular, in the third case ($\beta = -0.6$) each principal's preferred outcome within the implementable set involves a concession of size $\Pi_1 = 3$, while concessions much larger than this are implementable.

Finally, the frontier of implementable outcomes may exhibit a non-convex kink where it crosses from non-collaborative to collaborative outcomes. In particular, in both the second and third cases ($\beta = 0$ and $\beta = -0.6$), such a kink arises. This kink is generated by the discontinuously higher reservation value needed to maintain incentive compatibility for principal 2 inside the collaborative set.

3.4 Many principals

We now consider the case of $N \geq 3$ principals. This setting is more complex than the 2-principal case considered above, because the deterrence outcomes used to discipline



(a) $\beta = 1$ (b) $\beta = 0$

(c) $\beta = -0.6$ (d) $\beta = -0.97$

Figure 1: Implementable outcomes for the example in Section 3.3. Red outcomes are IC for principal 1, blue outcomes are IC for principal 2, and green outcomes are implementable.

one principal may be attractive to another principal. As a result, analogs of the IC constraints in Theorem 1 are no longer sufficient for implementability.

Nonetheless, versions of these constraints are still necessary, as we now establish. Let $\bar{\Pi} \equiv \sum_{i=1}^{N} \Pi_i$ be the combined power of all principals, and for each principal i, let $\bar{\Pi}_{-i} \equiv \bar{\Pi} - \Pi_i$ be the combined power of all principals except i. We first establish necessary conditions when principal 1 is powerful, in the sense that her power exceeds

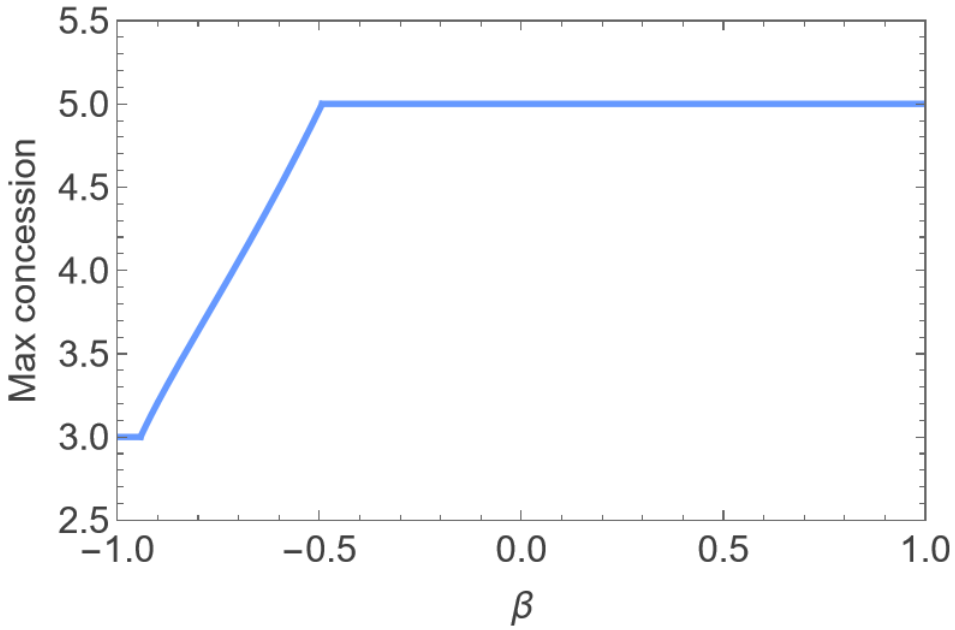


Figure 2: Maximum implementable concession for the example in Section 3.3.

the combined power of all other principals.

Proposition 3. *Suppose that* $\Pi_1 > \bar{\Pi}_{-1}$. *Fix* $x^* \in X$. *If* x^* *is implementable, then:*

- $\ell(x^*) \leq \bar{\Pi}$,
- *If* $\ell(x^*) \leq \Pi_1$, *then:*

$$g_1(x^*) \geq \overline{g}_1(\Pi_1 - \bar{\Pi}_{-1}), \quad g_i(x^*) \geq \underline{g}_i(\bar{\Pi}_{-i} - \Pi_i), \quad i = 2, ..., N$$

- *If* $\ell(x^*) > \Pi_1$, *then for each principal* $i = 1, ..., N$:

$$g_i(x^*) \geq \begin{cases} \overline{g}_i(\ell(x^*) - \bar{\Pi}_{-i}), & \ell(x^*) > \bar{\Pi}_{-i} \\ \underline{g}_i(\bar{\Pi}_{-i} - \Pi_i), & \ell(x^*) \leq \bar{\Pi}_{-i} \end{cases}, \quad i = 1, ..., N$$

These bounds generalize the IC constraints in the 2-principal setting, and they reduce to those bounds when $N = 2$. As in the 2-principal case, there is an asymmetry between the IC constraint on principal 1 and all other principals: Principal 1 can always force the agent to implement any outcome x satisfying $\ell(x) \leq \Pi_1 - \bar{\Pi}_{-1}$, whereas each other principal cannot guarantee any particular outcome when $\ell(x^*)$ is

small. Also as in the 2-principal case, reservation values are systematically larger for principals with higher power.

One new feature in the many-principal environment is a staggered rise in the reservation value across principals as $\ell(x^*)$ grows. With $N \geq 3$ principals, principal 1's reservation value begins rising once $\ell(x^*)$ rises above Π_1. However, principal 2's reservation value does not rise until $\ell(x^*)$ rises above $\bar{\Pi}_{-2} > \Pi_1$. And principal 3's reservation value does not begin rising until $\ell(x^*)$ rises above $\bar{\Pi}_{-3} > \bar{\Pi}_{-2}$, and so on.

We next establish analogous bounds when principal 1 is not powerful, in the sense that her power is no greater than the combined power of all other principals.

Proposition 4. *Suppose that* $\Pi_1 \leq \bar{\Pi}_{-1}$. *Fix* $x^* \in X$. *If* x^* *is implementable, then:*

- $\ell(x^*) \leq \bar{\Pi}$,
- *For each principal* $i = 1, ..., N$:

$$g_i(x^*) \geq \begin{cases} \overline{g}_i(\ell(x^*) - \bar{\Pi}_{-i}), & \ell(x^*) > \bar{\Pi}_{-i} \\ \underline{g}_i(\bar{\Pi}_{-i} - \Pi_i), & \ell(x^*) \leq \bar{\Pi}_{-i} \end{cases}$$

When principal 1 is not powerful, the IC constraints on the principals become symmetric. In particular, principal 1 cannot guarantee herself any particular outcome for small values of $\ell(x^*)$. However, reservation values are still systematically higher for more powerful principals and rise in staggered fashion, with principal i's reservation value beginning to rise once $\ell(x^*)$ exceeds $\bar{\Pi}_{-i} = \bar{\Pi} - \Pi_i$, a threshold which rises with i.

We complement these necessary conditions with tighter sufficient conditions ensuring implementability.

Proposition 5. *Fix* $x^* \in X$. *Then* x^* *is implementable if* $\ell(x^*) \leq \bar{\Pi}$ *and*

$$\begin{cases} g_i(x^*) \geq \overline{g}_i(\max\{\ell(x^*), \Pi_i\} - \bar{\Pi}_{-i}), & \max\{\ell(x^*), \Pi_i\} > \bar{\Pi}_{-i} \\ g_i(x^*) > g_i(0), & \textit{otherwise} \end{cases}$$

for each principal $i = 1, ..., N$.

Note that these sufficient conditions coincide with the necessary conditions characterized above whenever $\ell(x^*) > \bar{\Pi}_{-N}$, i.e., in the "fully collaborative" region in which

all principals' punishment power is jointly required to implement the outcome. Otherwise, there is a gap between the necessary and sufficient conditions for the weakest principals whose punishment power is not necessary to implement the outcome.

4 Transfers Too

So far, we have assumed that punishments are the only tool principals can use to influence the agent. However, in some contexts they may also be able to make transfers. These two influence tools differ in one key respect: While a punishment moves the principal's and agent's payoffs in the same direction, a transfer moves their payoffs in opposite directions.

This section makes two points. First, if transfers are frictionless—utility can be perfectly transferred between the agent and each principal—then transfers and punishments do not interact as tools for influence. More precisely, any policy which can be implemented without punishments can also be implemented with punishments, regardless of the power of each principal. In particular, any policy maximizing the sum of the payoffs of the three players can be implemented. Furthermore, there exist "orthogonalized" equilibria implementing these policies, with transfers used to influence the policy choice and punishments used to extract transfers from the agent. Within this class of equilibria, punishment power impacts the division of surplus but not the implemented policy.

Second, orthogonalization depends on the availability of frictionless transfers. In the presence of transaction costs, surplus-maximizing policies may not be implementable, and the maximum achievable surplus may be affected by the distribution of power among principals. When the friction is sufficiently large, principals use threats exclusively and the equilibrium policy set is the same as in our baseline (threats-only) setting.

4.1 Model

We now augment our baseline model to accommodate transfers. Each principal can both impose a punishment $p_i \in [0, \Pi_i]$ and make a transfer $m_i^P \in [0, M_i^P]$ to the agent, where $M_i^P > 0$ is an exogenous budget constraint. The agent can also make transfers to each principal. Formally, we augment the "material" policy set X with

an additional choice $\boldsymbol{m}^A = (m_1^A, ..., m_N^A) \in \mathbb{R}_+^N$, which the agent selects alongside the material policy $x \in X$. The quantity m_i^A records the transfer from the agent to principal i. The agent faces an exogenous aggregate budget constraint $M^A > 0$, so that feasible payment profiles satisfy $\sum_{i=1}^N m_i^A \leq M^A$.[10] For clarity, we will refer to transfers from the principals as *subsidies* and transfers from the agent as *contributions*.

Given a policy choice x, contributions $\boldsymbol{m}^A$, punishments $\boldsymbol{p}$, and subsidies $\boldsymbol{m}^P$, each principal's total payoff is:

$$g_i(x) - c_i(p_i) + (1 - \tau^A)m_i^A - m_i^P,$$

where $\tau^A \in [0, 1]$ is a transaction cost on contributions. Meanwhile, the agent's total loss is:

$$\ell(x) + \sum_{i=1}^N \Big(p_i + m_i^A - (1 - \tau_i^P)m_i^P\Big)$$

where $\tau_i^P \in [0, 1]$ is a transaction cost on subsidies from principal i.

Each principal simultaneously commits to a (l.s.c.) *influence schedule* $\iota_i = (\pi_i, \mu_i^P)$, which conditions punishments and subsidies on both the policy x and contributions $\boldsymbol{m}^A$. (For consistency with our treatment of general policy spaces, we allow each principal to condition on the contributions to other principals.) After observing the profile of influence schedules, the agent selects a policy and contribution, following which punishments and subsidies are enacted according to $\boldsymbol{\iota}$.

An equilibrium consists of a profile of influence schedules $\boldsymbol{\iota}$, along with response functions $(\chi, \boldsymbol{\mu}^A)$ specifying the agent's chosen policy and contributions as a function of the posted threats and subsidies. The optimality conditions for all principals and the agent are analogous to the baseline model.

As in the baseline model, on-path punishments can be ignored. The outcome of an equilibrium is therefore a triple $(x^*, \boldsymbol{m}^{A*}, \boldsymbol{m}^{P*})$, where x^* is the implemented policy, $\boldsymbol{m}^{A*}$ are the realized contributions, and $\boldsymbol{m}^{P*}$ are the realized subsidies. As in the baseline model, an equilibrium is said to implement a policy x^* if its policy outcome is x^*. A policy is implementable if some equilibrium implements it. Unlike the baseline model, the policy outcome does not fully describe the equilibrium payoffs of the various actors. Two different equilibria may implement the same policy while

[10] This augmentation yields a grand policy space satisfying all the requirements of our baseline model, and it is therefore a special case of that framework. We introduce distinct notation for transfers from the agent solely for expositional simplicity.

involving different contributions or subsidies, and therefore distinct payoff vectors.

4.2 Frictionless Transfers: One Principal

We begin our analysis by characterizing equilibrium outcomes with frictionless transfers and $N = 1$ principal. The goal of this subsection is to gain intuition for how transfers impact the use of threats and equilibrium outcomes, before moving to the more complex case with multiple principals.

The following proposition characterizes the set of equilibrium policy choices and net transfers.

Proposition 6. *Suppose that $N = 1$. If there are no transaction costs ($\tau_1^P = \tau^A = 0$) and M_1^P and M^A are sufficiently large, then $x^* \in X$ is implementable iff it maximizes aggregate surplus:*

$$x^* \in \arg\max_{x \in X} g_1(x) - \ell(x)$$

Furthermore, every equilibrium implementing such a policy satisfies

$$m^{P*} - m^{A*} = \ell(x^*) - \Pi_1,$$

where m^{P*} and m^{A*} are the equilibrium-path subsidy and contribution.

This result demonstrates that the principal's power controls the equilibrium division of surplus but has no impact on the set of implementable policies. In Coasian fashion, the equilibrium policy must maximize joint surplus, since otherwise the principal could profitably induce the agent to switch policies by promising to split the extra surplus. In equilibrium, the principal captures the entire joint surplus, plus an additional transfer Π_1 which she can extract from the agent by threatening punishment.

If the principal's punishment power is insufficient to make the agent choose such a policy, then the principal demands no contributions and offers a subsidy equaling the difference between the agent's loss from choosing x^* and the maximal punishment: $m^{P*} = \ell(x^*) - \Pi_1$. On the other hand, if her punishment power is larger than required to incentivize an efficient policy choice, the principal makes no subsidies and demands a contribution equal to her spare punishment power: $m^{A*} = \Pi_1 - \ell(x^*)$. The principal's power determines the size (and direction) of the resulting net transfer.

4.3 Frictionless Transfers: Multiple Principals

We now extend our analysis to a multi-principal ($N \geq 2$) setting. This section bridges the threats-only setting of our baseline model and the transfers-only setting studied by Bernheim and Whinston (1986) and Grossman and Helpman (1994).

We first establish an *orthogonalization* result when transfers are frictionless. Consider a *no-coercion* benchmark where the principals can offer subsidies but not punishments and cannot extract contributions from the agent. (This setting corresponds to the usual menu auction setting studied in the literature.) We will call an equilibrium in this setting an NC-equilibrium. We show that all policies which can be implemented by NC-equilibria are also implementable when punishments and contributions are possible. Furthermore, they can be implemented using influence schedules which do not condition punishments on the policy and do not condition subsidies on contributions.

Proposition 7. *Suppose that $x^* \in X$ is implemented by NC-equilibrium $(\hat{\boldsymbol{\mu}}^P, \hat{\chi})$. Then there exists an equilibrium $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^{P*}, \chi^*, \boldsymbol{\mu}^{A*})$ implementing x^* where:*

(i) Each principal extracts rent from her threat as payment from the agent:

$$\pi_i^*(x, \boldsymbol{m}^A) = \begin{cases} 0, & m_i^A \geq \Pi_i \\ \Pi_i, & \textit{otherwise} \end{cases}$$

(ii) Each principal offers the NC-equilibrium subsidies, and in particular subsidies do not depend on contributions:

$$\boldsymbol{\mu}^{P*}(x, \boldsymbol{m}^A) = \hat{\boldsymbol{\mu}}^P(x).$$

In particular, Bernheim and Whinston (1986) establish that every outcome which maximizes joint surplus

$$\sum_{i=1}^{N} g_i(x) - \ell(x)$$

can be implemented by an NC-equilibrium. The following corollary is immediate:

Corollary 2. *Every surplus-maximizing outcome x^* can be implemented using the influence schedules described in Proposition 7.*

In the *orthogonalized* equilibria of Proposition 7, the agent's policy choice is incentivized solely through subsidies, as in a menu auction. Threats are used solely to extort contributions. Intuitively, if some principal could monetize its threats more efficiently through a change in policy than a direct contribution, then the policy change must grow the joint surplus of the principal and agent. But in that case, the principal would have already found it profitable to induce that policy change through an additional payment in the no-coercion setting. By assumption, no such profitable deviations exist. Hence, each principal prefers to use its punishment power to extract payments rather than policy concessions.

A key implication of Proposition 7 and its corollary are that the maximum implementable joint surplus is independent of the distribution of punishment power. Changes in power may impact the distribution of surplus, but not the amount that can be generated through the interaction.

4.4 Transfers with a Transaction Cost

We have shown that, when transfers are frictionless, surplus-maximizing outcomes can be implemented. We now show that this result does not hold universally when transfers face a transaction cost. Furthermore, the maximal surplus that can be obtained in equilibrium may depend on the magnitude of principals' power.

Remark 1. *Suppose that transaction costs are nonzero. Then:*

1. *There need not exist a surplus-maximizing equilibrium,*
2. *The maximum implementable surplus may depend on the distribution of power,*
3. *For sufficiently high transaction costs, equilibria may feature no transfers.*

We establish this claim via a simple 1-principal example. Suppose that the policy space is $X = \mathbb{R}$ and payoff functions are $\ell(x) = x^2$ and $g_1(x) = 10x$. The principal's power is $\Pi_1 < 25$, and each side faces a transaction cost $\tau_1^A = \tau_1^P = \tau \in (0, 1]$ on transfers. In other words, a dollar transferred in either direction is worth only $1 - \tau$ dollars to the receiver. In this environment, total surplus is

$$S(x) = 10x - x^2,$$

and the unique surplus-maximizing policy is $x^{FB} = 5$, generating total surplus $S(5) = 25$.

If there is no transaction cost ($\tau = 0$), then x^{FB} is the unique implementable policy. In particular, suppose that the principal promises subsidy $m_1^P = \$25 - \Pi_1$ if the agent chooses policy $x = 5$, and maximal punishment and no subsidy for any other choice. Under this influence schedule, the agent is indifferent between cooperating and selecting $x = 0$, both of which yield a net loss of Π_1. Hence, cooperation is a best response.

Under this influence schedule, surplus is maximized and the principal holds the agent down to his individual rationality constraint. She therefore receives a payoff at least as high as under any other influence schedule and agent best response, and so she has no profitable deviations supposing that the agent cooperates. In other words, the surplus-maximizing outcome is implementable. (In fact, it is the unique implementable outcome, since if the agent does not cooperate in response to the influence schedule just outlined, then the principal has no optimal influence schedule.)

With positive transaction costs, the principal must burn surplus in order to incentivize policy choices above the level $\underline{x} = \sqrt{\Pi_1} < x^{FB} = 5$ which can be incentivized with threats alone. In particular, in order to implement some policy $\hat{x} > \underline{x}$, the principal must offer the subsidy

$$m_1^P(\hat{x}) = \frac{\ell(\hat{x}) - \Pi_1}{1 - \tau}$$

in return for choosing $x = \hat{x}$, with maximal punishment following any other policy choice. Among all such policies, the one maximizing the principal's payoff solves

$$\max_{\hat{x} \geq \underline{x}} 10\hat{x} - \frac{\hat{x}^2 - \Pi_1}{1 - \tau}.$$

This problem has unique solution $x^* = \max\{5(1 - \tau), \underline{x}\}$, which is also the unique implementable policy. (It is never optimal for the principal to implement a policy $\hat{x} < \underline{x}$ and extract surplus with a contribution, since doing so yields an even lower aggregate surplus than implementing $\hat{x} = \underline{x}$ without extracting any additional rent from the agent.)

Several features of this optimum are noteworthy. First, $x^* < x^{FB}$, so the maximum implementable surplus is below the first-best level. Second, when $\tau \geq \overline{\tau} \equiv 1 - \underline{x}/5$, no

transfers are made and the principal implements $x^* = \underline{x}$ with threats alone. Third, the surplus achieved in equilibrium varies with the principal's power. When $\tau \leq \overline{\tau}$, the implemented policy is independent of Π_1 while subsidies are decreasing in Π_1. As a result, total surplus (which is reduced by subsidies due to transaction cost) is increasing in Π_1 for small transaction costs. And when $\tau \geq \overline{\tau}$, there are no transfers and the implemented policy $x^* = \underline{x} = \sqrt{\Pi_1}$ is increasing in Π_1, hence so is total surplus. As a result, total surplus is increasing in Π_1 for any positive transaction cost and all $\Pi_1 < 25$.

A Proofs

A.1 Proof of Lemma 1

Let $\boldsymbol{\pi}'(x) = \boldsymbol{\pi}^*(x)$ for all $x \neq x^*$ and $\boldsymbol{\pi}'(x^*) = \mathbf{0}$. And let $\chi'(\boldsymbol{\pi}) = x^*$ whenever $x^* \in \arg\min_{x \in X} L(x, \boldsymbol{\pi})$. Otherwise, if $\boldsymbol{\pi}_{-i} = \boldsymbol{\pi}'_{-i}$ for some i, then let $\chi'(\boldsymbol{\pi}) = \chi(\pi_i, \boldsymbol{\pi}^*_{-i})$. And for all remaining threat profiles, let $\chi'(\boldsymbol{\pi}) = \chi(\boldsymbol{\pi})$.

We first check that χ' is a best response function for the agent. Since $\chi(\boldsymbol{\pi})$ is by hypothesis a best response to each $\boldsymbol{\pi}$, it follows that $\chi'(\boldsymbol{\pi})$ is as well whenever $\chi'(\boldsymbol{\pi}) = \chi(\boldsymbol{\pi})$. And trivially $\chi'(\boldsymbol{\pi})$ is a best response whenever $x^* \in \arg\min_{x \in X} L(x, \boldsymbol{\pi})$. So consider the remaining case: x^* is not a best response to $L(x, \boldsymbol{\pi})$ and $\boldsymbol{\pi}_{-i} = \boldsymbol{\pi}'_{-i}$ for some i. By construction, $L(x, \boldsymbol{\pi}) \leq L(x, (\pi_i, \boldsymbol{\pi}^*_{-i}))$ for all x, with equality whenever $x \neq x^*$. Then since x^* is not a minimizer of $L(x, \boldsymbol{\pi})$, it must also not be a minimizer of $L(x, (\pi_i, \boldsymbol{\pi}^*_{-i}))$, and the set of minimizers of these two functions coincide. In particular, $\chi(\pi_i, \boldsymbol{\pi}^*_{-i})$, which minimizes the latter function by hypothesis, must also minimize the former. So χ' is a best response function.

We next establish that $\chi'(\boldsymbol{\pi}') = x^*$. Observe that $L(x, \boldsymbol{\pi}') \leq L(x, \boldsymbol{\pi}^*)$, with equality whenever $x \neq x^*$. If x^* were not a minimizer of $L(x, \boldsymbol{\pi}')$, it could then not be a minimizer of $L(x, \boldsymbol{\pi}^*)$, a contradiction of the hypothesis that χ is a best response function and $\chi(\boldsymbol{\pi}^*) = x^*$. So it must be that x^* minimizes $L(x, \boldsymbol{\pi}')$ and therefore, by construction, $\chi'(\boldsymbol{\pi}') = x^*$.

Finally, we check that no principal wishes to unilaterally deviate from the threat profile $\boldsymbol{\pi}'$ given response function χ'. Fix a principal i and a deviation $\pi''_i \neq \pi'_i$. Let $\boldsymbol{\pi}'' = (\pi''_i, \pi'_{-i})$ and $\boldsymbol{\pi}^\dagger = (\pi''_i, \pi^*_{-i})$. If x^* is a best response to $\boldsymbol{\pi}''$, then by construction $\chi'(\boldsymbol{\pi}'') = x^*$, in which case π''_i is not a profitable deviation. So suppose that x^* is

not a best response. Since $L(x, \boldsymbol{\pi}'') \leq L(x, \boldsymbol{\pi}^\dagger)$ for all x, with equality for $x \neq x^*$, it follows that x^* is also not a best response to $\boldsymbol{\pi}^\dagger$. But then $\chi'(\boldsymbol{\pi}'') = \chi(\boldsymbol{\pi}^\dagger)$. By hypothesis, π_i'' is not a profitable deviation for i from $\boldsymbol{\pi}^*$ given response function χ. Therefore, π_i'' is also not a profitable deviation for i from $\boldsymbol{\pi}'$ given response function χ'. We conclude that $(\boldsymbol{\pi}', \chi')$ is an equilibrium with policy outcome x^*.

Assume further that costs are positive for each principal whenever inflicting a positive punishment. Suppose by way of contradiction that $\pi_i^*(x^*) > 0$ for some principal i. Consider a deviation to the policy π_i' given by $\pi_i'(x) = \pi_i(x)$ for all $x \neq x^*$ and $\pi_i'(x^*) = 0$. Let $\boldsymbol{\pi}' = (\pi_i', \boldsymbol{\pi}_{-i}^*)$. Note that $L(x, \boldsymbol{\pi}') = L(x, \boldsymbol{\pi}^*)$ for all $x \neq x$. Meanwhile, $L(x^*, \boldsymbol{\pi}') < L(x^*, \boldsymbol{\pi}^*)$ given that $\pi_i^*(x^*) > \pi_i'(x^*)$. By hypothesis, x^* is a best response by the agent to $\boldsymbol{\pi}^*$. Hence it must be a *unique* best response to $\boldsymbol{\pi}'$, meaning that $\chi(\boldsymbol{\pi}') = x^*$. Given that $c_i(\pi_i^*(x^*)) > 0$, we therefore have

$$\begin{aligned} &G_i(\chi(\boldsymbol{\pi}'), \pi_i'(\chi(\boldsymbol{\pi}'))) = G_i(x^*, \pi_i'(x^*)) = g_i(x^*) \\ &> g_i(x^*) - c_i(\pi_i^*(x^*)) = G_i(\chi(\boldsymbol{\pi}^*), \pi_i^*(\chi(\boldsymbol{\pi}^*))). \end{aligned}$$

In other words, π_i' is a profitable deviation for principal i.

A.2 Proof of Theorem 1

We begin by characterizing a useful class of admissible threats, which will be sufficient to construct equilibria supporting all implementable policies.

Lemma A.1. *Let Y be a closed subset of X. Suppose $\pi_i : X \to [0, \Pi_i]$ is lower semicontinuous when restricted to Y and satisfies $\pi_i(x) = \Pi_i$ for $x \notin Y$. Then $\pi_i \in \mathcal{P}_i$.*

Proof. Let E be the epigraph of π_i, and consider any limit point (x, p) of E. Then there exists a sequence $(x_n, p_n) \in E$ converging to (x, p), which must satisfy $p_n \geq \pi_i(x_n)$ for each n. Suppose first that only finitely many elements of the sequence x_n are in Y. Then eventually $\pi_i(x_n) = \Pi_i$, implying that $p = \lim_n p_n = \Pi_i \geq \pi_i(x)$ and so $(x, p) \in E$. So suppose instead that infinitely many elements of the sequence are in Y. Passing to this subsequence of $\{(x_n, p_n)\}_{n \in \mathbb{N}}$, we have by closedness of Y that $x \in Y$ and by lower semicontinuity of π_i on Y that $\liminf_n \pi_i(x_n) \geq \pi_i(x)$. Hence $p = \lim_n p_n \geq \liminf_n \pi_i(x_n) \geq \pi_i(x)$, i.e., $(x, p) \in E$. So E must contain all its limit points and is a closed set. Equivalently, $\pi_i \in \mathcal{P}_i$. □

In particular, whenever Y is a finite set, threats which inflict maximal punishment everywhere outside Y are admissible.

Lemma A.2. *Fix $x^* \in X$. If $\ell(x^*) > \Pi_1 + \Pi_2$, then x^* is not implementable.*

Proof. Suppose not. Let $(\boldsymbol{\pi}^*, \chi)$ be an equilibrium implementing policy x^*. If $\ell(x^*) > \Pi_1 + \Pi_2$, then

$$L(x^*, \boldsymbol{\pi}^*) > \Pi_1 + \Pi_2 \geq L(0, \boldsymbol{\pi}^*),$$

contradicting the hypothesis that $\chi(\boldsymbol{\pi}^*) = x^* \in \arg\min_{x \in X} L(x^*, \boldsymbol{\pi}^*)$. □

Lemma A.3. *Fix $x^* \in X$. If x^* is implementable, then $g_1(x^*) \geq \overline{g}_1(\Delta\Pi)$ and $g_2(x^*) \geq \underline{g}_2(\Delta\Pi)$.*

Proof. Let x^* be implemented by equilibrium $(\boldsymbol{\pi}^*, r)$. Suppose by way of contradiction that $g_1(x^*) < \overline{g}_1(\Delta\Pi)$. Then there exists some $x' \in X$ such that $\ell(x') < \Delta\Pi$ and $g_1(x') > g_1(x^*)$. In that case,

$$\ell(x') + \pi_2^*(x') < \Delta\Pi + \pi_2^*(x') \leq \Pi_1 \leq \ell(x) + \Pi_1 + \pi_2^*(x)$$

for every $x \in X$. So define the threat π_1' by $\pi_1'(x) = \Pi_1 \mathbf{1}\{x \neq x'\}$ and let $\boldsymbol{\pi}' = (\pi_1', \pi_2^*)$. Then

$$\arg\min_{x \in X} L(x, \boldsymbol{\pi}') = x'$$

and therefore $\chi(\boldsymbol{\pi}') = x'$. Since 1 strictly prefers x' to x^*, it cannot be that $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium, which is the desired contradiction.

Next, suppose by way of contradiction that $g_2(x^*) < \underline{g}_2$. Then surely $\ell(x^*) > \Delta\Pi$. Consider the threat π_2' defined by $\pi_2'(x) = \Pi_2 \mathbf{1}\{\ell(x) > \Delta\Pi\}$. (Since X is a closed set and ℓ is continuous, $\{\ell(x) \leq \Delta\Pi\}$ is also closed and so π_2' is l.s.c.) Let $\boldsymbol{\pi}' = (\pi_1^*, \pi_2')$. Then for every x' such that $\ell(x') > \Delta\Pi$, we have

$$\begin{aligned} &\min_{x \in X} L(x, \boldsymbol{\pi}') \leq L(0, \boldsymbol{\pi}') \leq \Pi_1 = \Delta\Pi + \Pi_2 \\ &< \ell(x') + \Pi_2 \leq L(x', \boldsymbol{\pi}'). \end{aligned}$$

Hence $\ell(x') \leq \Delta\Pi$ for every $x' \in \arg\min_{x \in X} L(x, \boldsymbol{\pi}')$, and in particular $\ell(\chi(\boldsymbol{\pi}')) \leq \Delta\Pi$. But then $g_2(\chi(\boldsymbol{\pi}')) \geq \underline{g}_2(\Delta\Pi) > g_2(x^*)$, meaning that 2 has a profitable deviation from π_2^* to π_2', a contradiction. □

Lemma A.4. *Fix x^* such that $\ell(x^*) \leq \Pi_1$ and $g_1(x^*) \geq \overline{g}_1(\Delta\Pi)$ and $g_2(x^*) \geq \underline{g}_2(\Delta\Pi)$. Then x^* is implementable.*

Proof. Choose $\underline{x} \in \arg\min_{\{x:\ell(x)\leq\Delta\Pi\}} g_2(x)$. (Since X is compact and ℓ and g_2 are continuous, such a minimizer exists.) By the richness condition, this minimizer can be chosen so that $\underline{x} \neq 0$, which we assume going forward. Let $\pi_2^*(x) = \Pi_2 \mathbf{1}\{x \neq x^*, 0\}$. If $\ell(x^*) \leq \Delta\Pi$, then let $\pi_1^*(x) = \Pi_1 \mathbf{1}\{x \neq x^*\}$, and otherwise let

$$\pi_1^*(x) = \begin{cases} 0, & x = x^* \\ \Delta\Pi - \ell(\underline{x}), & x = \underline{x} \\ \Pi_1, & \text{o.w.} \end{cases}$$

Since $\ell(\underline{x}) \leq \Delta\Pi$, these punishments are feasible. Note also that $\underline{x} \neq x^*$ whenever $\ell(x^*) > \Delta\Pi$. We construct a response function χ such that $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium implementing x^*.

Given punishments $\boldsymbol{\pi}^*$, the agent incurs loss $\ell(x^*) \leq \Pi_1$ from choosing x^* and loss $L(x, \boldsymbol{\pi}^*) \geq \Pi_1$ from any $x \neq x^*, \underline{x}$. Meanwhile, if $\ell(x^*) > \Delta\Pi$, then by choosing policy $\underline{x}$ the agent incurs loss $L(\underline{x}, \boldsymbol{\pi}^*) = \Pi_1$. So x^* is a best response by the agent to $\boldsymbol{\pi}^*$, and we select $\chi(\boldsymbol{\pi}^*) = x^*$.

Next, consider any threat $\pi_1' \neq \pi_1^*$ for 1. Let $\boldsymbol{\pi}' = (\pi_1', \pi_2^*)$. If x^* is a best response to $\boldsymbol{\pi}'$, we select $\chi(\boldsymbol{\pi}') = x^*$, in which case a deviation to π_1' is not profitable for 1. So suppose that x^* is not a best response. If 0 is a best response to $\boldsymbol{\pi}'$, we select $\chi(\boldsymbol{\pi}') = 0$, in which case again a deviation to π_1' is not profitable for 1 given that $g_1(0) \leq \overline{g}_1(\Delta\Pi) \leq g_1(x^*)$. Finally, suppose that neither x^* nor 0 are best responses. For any $x' \neq x^*$ such that $\ell(x') \geq \Delta\Pi$ we have

$$L(x', \boldsymbol{\pi}') \geq \Delta\Pi + \Pi_2 = \Pi_1 \geq L(0, \boldsymbol{\pi}').$$

Then since 0 is not a best response, no $x' \neq x^*$ such that $\ell(x') \geq \Delta\Pi$ can be. Since x^* is also not a best response, every best response x'' must therefore satisfy $\ell(x'') < \Delta\Pi$. So let $\chi(\boldsymbol{\pi}')$ be an arbitrary best response. Then $g_1(\chi(\boldsymbol{\pi}')) \leq \overline{g}_1(\Delta\Pi) \leq g_1(x^*)$. So π_1' is not a profitable deviation for 1. Under this selection of best responses, π_1^* is therefore a best response to π_2^*.

Finally, consider any threat $\pi_2' \neq \pi_2^*$ for 2. Let $\boldsymbol{\pi}' = (\pi_1^*, \pi_2')$. If $\ell(x^*) \leq \Delta\Pi$, then

for all $x' \neq x^*$ we have

$$L(x', \boldsymbol{\pi}') \geq \Pi_1 = \Delta\Pi + \Pi_2 \geq L(x^*, \boldsymbol{\pi}').$$

In other words, x^* is a best response to $\boldsymbol{\pi}'$, and we select $\chi(\boldsymbol{\pi}') = x^*$. Then χ_2' is not a profitable deviation for 2, meaning π_2^* is a best response to π_1^* under this choice of χ.

Suppose instead that $\ell(x^*) > \Delta\Pi$. If x^* is a best response to $\boldsymbol{\pi}'$, we select $\chi(\boldsymbol{\pi}') = x^*$, in which case a deviation to π_2' is not profitable for 2. So suppose that x^* is not a best response. Then for any $x' \neq x^*, \underline{x}$ we have

$$L(x', \boldsymbol{\pi}') \geq \Pi_1 = \ell(\underline{x}) + \Delta\Pi - \ell(\underline{x}) + \Pi_2 \geq L(\underline{x}, \boldsymbol{\pi}').$$

Hence, $\underline{x}$ must be a best response to $\boldsymbol{\pi}'$, and we select $\chi(\boldsymbol{\pi}') = \underline{x}$. In that case, $g_2(\chi(\boldsymbol{\pi}')) = \underline{g}_2(\Delta\Pi) \leq g_2(x^*)$. So π_2' is not a profitable deviation for 2, meaning π_2^* is a best response to π_1^* under this choice of χ. □

Lemma A.5. *Fix $x^* \in X$ such that $\ell(x^*) > \Pi_1$. If x^* is implementable, then $g_1(x^*) \geq \overline{g}_1(\ell(x^*) - \Pi_2)$ and $g_2(x^*) \geq \overline{g}_2(\ell(x^*) - \Pi_1)$.*

Proof. Let x^* be implemented by equilibrium $(\boldsymbol{\pi}^*, \chi)$. Suppose by way of contradiction that $g_1(x^*) < \overline{g}_1(\ell(x^*) - \Pi_2)$. Then there exists some $x' \in X$ such that $\ell(x') < \ell(x^*) - \Pi_2$ and $g_1(x') > g_1(x^*)$. In that case, for every $x \in X$ we have

$$\ell(x') + \pi_2^*(x') < \ell(x^*) = L(x^*, \boldsymbol{\pi}^*) \leq L(x, \boldsymbol{\pi}^*) \leq \ell(x) + \Pi_1 + \pi_2^*(x)$$

given that x^* is a best response to $\boldsymbol{\pi}^*$. So define the threat π_1' by $\pi_1'(x) = \Pi_1 \mathbf{1}\{x \neq x'\}$, and let $\boldsymbol{\pi}' = (\pi_1', \pi_2^*)$. Then

$$\arg\min_{x \in X} L(x, \boldsymbol{\pi}') = x'$$

and therefore $\chi(\boldsymbol{\pi}') = x'$. Since 1 strictly prefers x' to x^*, it cannot be that $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium, which is the desired contradiction.

Next, suppose by way of contradiction that $g_2(x^*) < \overline{g}_2(\ell(x^*) - \Pi_1)$. Then there exists some $x' \in X$ such that $\ell(x') < \ell(x^*) - \Pi_1$ and $g_2(x') > g_2(x^*)$. In that case, for every $x \in X$ we have

$$\ell(x') + \pi_1^*(x') < \ell(x^*) = L(x^*, \boldsymbol{\pi}^*) \leq L(x, \boldsymbol{\pi}^*) \leq \ell(x) + \pi_1^*(x) + \Pi_2$$

given that x^* is a best response to $\boldsymbol{\pi}^*$. So define the threat π_2' by $\pi_2'(x) = \Pi_2 \mathbf{1}\{x \neq x'\}$, and let $\boldsymbol{\pi}' = (\pi_1^*, \pi_2')$. Then

$$\arg\min_{x \in X} L(x, \boldsymbol{\pi}') = x'$$

and therefore $\chi(\boldsymbol{\pi}') = x'$. Since 2 strictly prefers x' to x^*, it cannot be that $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium, which is the desired contradiction. □

Lemma A.6. *Fix $x^* \in X$ such that $\Pi_1 < \ell(x^*) \leq \Pi_1 + \Pi_2$. If $g_1(x^*) \geq \overline{g}_1(\ell(x^*) - \Pi_2)$ and $g_2(x^*) \geq \overline{g}_2(\ell(x^*) - \Pi_1)$, then x^* is implementable.*

Proof. The hypothesis $\ell(x^*) > \Pi_1$ implies that $x^* \neq 0$. So let

$$\pi_2^*(x) = \begin{cases} 0, & x = x^* \\ \ell(x^*) - \Pi_1, & x = 0 \\ \Pi_2, & \text{otherwise} \end{cases}$$

Choose any $\underline{x} \neq 0$ such that $\ell(\underline{x}) < \ell(x^*) - \Pi_1$. Under our richness assumption, such a choice is possible. Note that $\underline{x} \neq x^*$. Let

$$\pi_1^*(x) = \begin{cases} 0, & x = x^* \\ \ell(x^*) - \ell(\underline{x}) - \Pi_2, & x = \underline{x} \\ \Pi_1, & \text{o.w.} \end{cases}$$

Since $\Delta\Pi < \ell(x^*) - \ell(\underline{x}) - \Pi_2 \leq \Pi_1$ and $0 < \ell(x^*) - \Pi_1 \leq \Pi_2$, these punishments are feasible. We construct a response function χ for which $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium implementing x^*.

Given punishments $\boldsymbol{\pi}^*$, the agent incurs loss $\ell(x^*) \leq \Pi_1 + \Pi_2$ from choosing x^* and loss $L(x, \boldsymbol{\pi}^*) \geq \Pi_1 + \Pi_2$ from any $x \neq x^*, \underline{x}, 0$. Meanwhile, by choosing policy 0, the agent incurs loss $L(0, \boldsymbol{\pi}^*) = \ell(x^*)$. Finally, by choosing policy $\underline{x}$ the agent incurs loss $L(\underline{x}, \boldsymbol{\pi}^*) = \ell(x^*)$. So x^* is a best response by the agent to $\boldsymbol{\pi}^*$, and we select $\chi(\boldsymbol{\pi}^*) = x^*$.

Next, consider any threat $\pi_1' \neq \pi_1^*$ for 1, and let $\boldsymbol{\pi}' = (\pi_1', \pi_2^*)$. If x^* is a best response to $\boldsymbol{\pi}'$, we select $\chi(\boldsymbol{\pi}') = x^*$, in which case a deviation to π_1' is not profitable for 1. So suppose that x^* is not a best response. If 0 is a best response, we select $\chi(\boldsymbol{\pi}') = 0$, in which case a deviation is again not profitable given that $g_1(0) \leq \overline{g}_1(\ell(x^*) - \Pi_2) \leq g_1(x^*)$. Finally, suppose that neither x^* nor 0 are best responses.

For any $x' \neq x^*$ such that $\ell(x') \geq \ell(x^*) - \Pi_2$ we have

$$L(x', \boldsymbol{\pi}') \geq \ell(x^*) = \ell(x^*) - \Pi_1 + \Pi_1 \geq L(0, \boldsymbol{\pi}').$$

Since 0 is not a best response, neither is any $x' \neq x^*$ such that $\ell(x') \geq \ell(x^*) - \Pi_2$. And since x^* is not a best response, every every best response x'' must satisfy $\ell(x'') < \ell(x^*) - \Pi_2$. So let $\chi(\boldsymbol{\pi}')$ be an arbitrary best response. Then $g_1(r(\pi_1', \pi_2^*)) \leq \overline{g}_1(\ell(x^*) - \Pi_2) \leq g_1(x^*)$, meaning that π_1' is not a profitable deviation for 1. Under this selection of best responses, π_1^* is therefore a best response to π_2^*.

Finally, consider any threat $\pi_2' \neq \pi_2^*$ for 2, and let $\boldsymbol{\pi}' = (\pi_1^*, \pi_2')$. If $\underline{x}$ is a best response to $\boldsymbol{\pi}'$, select $\chi(\boldsymbol{\pi}') = \underline{x}$. Since $\ell(\underline{x}) < \ell(x^*) - \Pi_1$, we have $g_2(\chi(\boldsymbol{\pi}')) \leq \overline{g}_2(\ell(x^*) - \Pi_2) \leq g_2(x^*)$, implying that π_2' is not a profitable deviation for 2. Suppose instead that $\underline{x}$ is not a best response. For all $x' \neq x^*$ such that $\ell(x') \geq \ell(x^*) - \Pi_1$ we have $x' \neq \underline{x}$ and therefore

$$L(x', \boldsymbol{\pi}') \geq \ell(x^*) = \ell(\underline{x}) + \ell(x^*) - \ell(\underline{x}) - \Pi_2 + \Pi_2 \geq L(\underline{x}, \boldsymbol{\pi}').$$

Since $\underline{x}$ is not a best response, neither is any $x' \neq x^*$ such that $\ell(x') \geq \ell(x^*) - \Pi_1$. Then as x^* is also not a best response, every best response x'' satisfies $\ell(x'') < \ell(x^*) - \Pi_1$. So let $\chi(\boldsymbol{\pi}')$ be an arbitrary best response. Then $g_2(\chi(\boldsymbol{\pi}')) \leq \overline{g}_2(\ell(x^*) - \Pi_2) \leq g_2(x^*)$, implying that π_2' is not a profitable deviation for 2. Under this selection of best responses, π_2^* is therefore a best response to π_1^*. □

A.3 Proof of Proposition 1

$\ell(x^*) > \Pi_1'$ ensures that x^* requires collaboration to implement under both sets of power levels. Then since each $\overline{g}_i$ is nondecreasing, both principals' reservation values are nonincreasing in $\boldsymbol{\Pi}$ within the collaborative regime. Hence, incentive compatibility for the principals at power level $\boldsymbol{\Pi}$ implies the same at power levels $\boldsymbol{\Pi}'$. Finally, implementability under power level $\boldsymbol{\Pi}$ implies that $\ell(x^*) \leq \Pi_1 + \Pi_2 \leq \Pi_1' + \Pi_2'$, so x^* is incentive compatible for the agent at power levels $\boldsymbol{\Pi}'$. In other words, x^* is implementable at those power levels.

A.4 Proof of Proposition 2

Consider first a rise in principal 1's power. Suppose that x^* is not implementable when principal 1's power level is Π_1'. The hypothesis that x^* is implementable under power level Π_1 implies that $g_2(x^*) \geq \underline{g}_2(\Pi_1 - \Pi_2) \geq \underline{g}_2(\Pi_1' - \Pi_2)$. Then since $\ell(x^*) \leq \Pi_1'$ by hypothesis, lack of implementability under power level Π_1' must imply that $g_1(x^*) < \overline{g}_1(\Pi_1' - \Pi_2)$.

Let

$$M \equiv \underset{\{x \in X \,:\, \ell(x) < \Pi_1' - \Pi_2\}}{\arg\max} g_1(x).$$

If M is non-empty, let x^{**} be any member of M. Then $g_1(x^{**}) = \overline{g}_1(\Pi_1' - \Pi_2) > g_1(x^*)$ by construction. Also by construction, $\ell(x^{**}) < \Pi_1' - \Pi_2$, so automatically $\ell(x^{**}) \leq \Pi_1'$ and $g_2(x^{**}) \geq \underline{g}_2(\Pi_1' - \Pi_2)$. Hence x^{**} is implementable under power level Π_1' and is preferred by 1 to x^*.

Suppose instead that M is empty. Since the set $\{x \,:\, \ell(x) < \Pi_1' - \Pi_2\}$ is non-empty (in particular, it contains the element 0), there must exist a sequence $x^1, x^2, ...$ in X such that each $\ell(x^n) < \Pi_1' - \Pi_2$ and $g_1(x^n) \to \overline{g}_1(\Pi_1' - \Pi_2)$. Since X is a compact subset of a Euclidean space, there must exist a convergent subsequence of $x^1, x^2, ...$ with limit $x^{**} \in X$. Since g_1 is continuous, it must be that $g_1(x^{**}) = \overline{g}_1(\Pi_1' - \Pi_2) > g_1(x^*)$. Meanwhile, emptiness of M implies that $\ell(x^{**}) \geq \Pi_1' - \Pi_2$. Then since ℓ is continuous, it must be that $\ell(x^{**}) = \Pi_1' - \Pi_2$. Hence $\ell(x^{**}) \leq \Pi_1'$ and $g_2(x^{**}) \geq \underline{g}_2(\Pi_1' - \Pi_2)$. So x^{**} is implementable under power level Π_1' and is preferred by 1 to x^*.

Next, consider a rise in principal 1's power. Suppose that x^* is not implementable when principal 2's power level is Π_2'. The hypothesis that x^* is implementable under power level Π_2 implies that $g_1(x^*) \geq \overline{g}_1(\Pi_1 - \Pi_2) \geq \overline{g}_2(\Pi_1 - \Pi_2')$. Then since $\ell(x^*) \leq \Pi_1$ by hypothesis, lack of implementability under power level Π_2' must imply that $g_2(x^*) < \underline{g}_2(\Pi_1 - \Pi_2')$.

Let

$$M \equiv \underset{\{x \in X \,:\, \ell(x) \leq \Pi_1 - \Pi_2'\}}{\arg\max} g_1(x).$$

Since X is compact and ℓ is continuous, the feasible set $\{x \in X \,:\, \ell(x) \leq \Pi_1 - \Pi_2'\}$ is also compact. It is also non-empty, since in particular it contains the outcome $x = 0$. Then since g_1 is also continuous, M is non-empty. Let x^{**} be any member of M. Then $g_1(x^{**}) \geq \overline{g}_1(\Pi_1 - \Pi_2')$ by construction. (Recall that $\overline{g}_1$ maximizes over the smaller set of outcomes satisfying $\ell(x) < \Pi_1 - \Pi_2'$ so this inequality may be

strict.) Also by construction, $\ell(x^{**}) \leq \Pi_1 - \Pi_2'$, so automatically $\ell(x^{**}) \leq \Pi_1$ and $g_2(x^{**}) \geq \underline{g}_2(\Pi_1 - \Pi_2') > g_2(x^*)$. Hence x^{**} is implementable under power level Π_2 and is preferred by 2 to x^*.

A.5 Proof of Propositions 3 and 4

Suppose that x^* is implementable, and fix any equilibrium $(\boldsymbol{\pi}^*, \chi)$ implementing x^*. If $\ell(x^*) > \bar{\Pi}$, then the agent's loss from deviating to $x = 0$ must be strictly less than his loss under the putative equilibrium policy. So $\ell(x^*) \leq \bar{\Pi}$.

We now show that $g_i(x^*) \geq \underline{g}_i(\bar{\Pi}_{-i} - \Pi_i)$ for every principal i whose power satisfies $\Pi_i \leq \bar{\Pi}_{-i}$. Note in particular that every principal $i \geq 2$ has a power satisfying this bound. Suppose that such a principal deviates to the threat $\pi_i'(x) = \Pi_i \cdot \mathbf{1}\{\ell(x) > \bar{\Pi}_{-i} - \Pi_i\}$. Then by choosing the policy $x = 0$, the agent incurs a loss no larger than $\bar{\Pi}_{-i}$; whereas by choosing any policy x satisfying $\ell(x) > \bar{\Pi}_{-i} - \Pi_i$, the agent incurs a loss strictly larger than $\bar{\Pi}_{-i}$. As a result, no such policy can be optimal following principal i's deviation. In other words, the deviation ensures implementation of *some* outcome x' satisfying $\ell(x') \leq \bar{\Pi}_{-i} - \Pi_i$. Hence, x^* is not implementable unless $g_i(x^*)$ is at least as good as some such x', yielding the desired bound.

We next show that the stronger bound $g_i(x^*) \geq \overline{g}_i(\ell(x^*) - \bar{\Pi}_{-i})$ must hold for any principal i such that $\ell(x^*) > \bar{\Pi}_{-i}$. (Note in particular that $\ell(x^*) > \bar{\Pi}_{-i}$ implies $\ell(x^*) > \Pi_1$ whenever $i \geq 2$.) Let i be any such principal. Suppose that there existed an $x' \neq x^*$ such that $g_i(x') > g_i(x^*)$ and $\ell(x') < \ell(x^*) - \bar{\Pi}_{-i}$. In that case, for every $x \in X$ we have

$$\ell(x') + \sum_{j \neq i} \pi_i^*(x') < \ell(x^*) = L(x^*, \boldsymbol{\pi}^*) \leq L(x, \boldsymbol{\pi}^*) \leq \ell(x) + \Pi_i + \sum_{j \neq i} \pi_i^*(x)$$

given that x^* is a best response to $\boldsymbol{\pi}^*$. So define the threat π_i' by $\pi_i'(x) = \Pi_i \mathbf{1}\{x \neq x'\}$, and let $\boldsymbol{\pi}' = (\pi_i', \boldsymbol{\pi}_{-i}^*)$. Then

$$\arg\min_{x \in X} L(x, \boldsymbol{\pi}') = x'$$

and therefore $\chi(\boldsymbol{\pi}') = x'$. Since i strictly prefers x' to x^*, it cannot be that $(\boldsymbol{\pi}^*, \chi)$ is an equilibrium. So there cannot exist such an x', implying the desired bound.

Finally, we derive the yet stronger bound $g_1(x^*) \geq \overline{g}_1(\Pi_1 - \bar{\Pi}_{-1})$ for principal 1

when $\Pi_1 > \bar{\Pi}_{-1}$ and $\ell(x^*) \leq \Pi_1$. Suppose $\ell(x^*) \leq \Pi_1$. Fix any $x' \neq x^*$ such that $\ell(x') < \Pi_1 - \bar{\Pi}_{-1}$, and suppose that principal 1 deviates to the threat $\pi_1'(x) = \Pi_1 \cdot \mathbf{1}\{x \neq x'\}$. Then the agent incurs a loss no greater than $\ell(x') + \bar{\Pi}_{-1}$ by choosing $x = x'$, and no less than $\Pi_1 > \ell(x') + \bar{\Pi}_{-1}$ from choosing any $x \neq x'$. This deviation therefore ensures implementation of any such outcome x'. Hence, x^* is not implementable unless $g_1(x^*) \geq g_1(x')$ for all such x', yielding the desired bound.

A.6 Proof of Proposition 6

Let $S(x) \equiv g_1(x) - \ell(x)$ be joint surplus. We first argue that no policy can be implemented which does not maximize S. By way of contradiction, suppose that some equilibrium implemented x^* satisfying $S(x') > S(x)$ for some alternative policy x'. We proceed by constructing a profitable deviation schedule for the principal which modifies her response following a single choice by the agent.

Suppose first that both the principal and agent (strictly) prefer x' to x^*. Consider a deviation schedule which modifies the response to (x', m^{A*}) so that it duplicates the response to (x^*, m^{A*}). By hypothesis, the agent strictly prefers (x', m^{A*}) to (x^*, m^{A*}) under the deviation schedule. Since (x^*, m^{A*}) was a best response to the original schedule, it must be at least as good as all choices other than (x', m^{A*}) under the deviation schedule. Hence (x', m^{A*}) must be a unique best response to the deviation schedule and must be chosen by the agent in the hypothesized equilibrium. By hypothesis, the principal strictly prefers x' to x^*, implying that this deviation schedule is profitable.

Suppose next that the principal (strictly) prefers x' to x^* but the agent does not. Fix $\epsilon \in (\ell(x') - \ell(x^*), g_1(x') - g_1(x^*))$ and consider a deviation schedule which, in response to a choice of (x', m^{A*}), promises the agent no punishment and subsidy $m^{P*} + \epsilon$. By hypothesis, the agent strictly prefers to choose (x', m^{A*}) over (x^*, m^{A*}) under this schedule. Since (x^*, m^{A*}) was a best response to the original schedule, it must be at least as good as all choices other than (x', m^{A*}) under the deviation schedule. Hence (x', m^{A*}) must be a unique best response to the deviation schedule and must be chosen by the agent in the hypothesized equilibrium. By hypothesis, the principal strictly prefers x' to x^*, implying that this deviation schedule is profitable.

Finally, suppose that the agent (strictly) prefers x' to x^* but the principal does not. Fix $\epsilon \in (g_1(x^*) - g_1(x'), \ell(x^*) - \ell(x'))$ and consider a deviation schedule which, in

response to a choice of $(x', m^{A*} + \epsilon)$, promises the agent no punishment and subsidy m^{P*}. By hypothesis, the agent strictly prefers choosing $(x', m^{A*}+\epsilon)$ to (x^*, m^{A*}) under this schedule. Since (x^*, m^{A*}) was a best response to the original schedule, it must be at least as good as all choices other than $(x', m^{A*} + \epsilon)$ under the deviation schedule. Hence $(x', m^{A*} + \epsilon)$ must be a unique best response to the deviation schedule and must be chosen by the agent in the hypothesized equilibrium. By hypothesis, the principal strictly prefers x' with an additional transfer of ϵ to x^*, implying that this deviation schedule is profitable.

Now, fix any $x^* \in X$ which maximizes S. We show that x^* is implementable, and further that every equilibrium implementing x^* satisfies $m^{P*} - m^{A*} = \ell(x^*) - \Pi_1$.

Case 1: $\ell(x^*) \leq \Pi_1$. Let $\mu_1^P(x, m_1^A) = 0$ for all x and m_1^A, and let

$$\pi_1(x, m_1^A) = \begin{cases} 0, & x = x^* \text{ and } m_1^A = \Pi_1 - \ell(x^*) \\ \Pi_1, & \text{otherwise} \end{cases}$$

Under this schedule, the agent incurs a loss of exactly Π_1 by choosing $(x^*, \Pi_1 - \ell(x^*))$, and a loss of at least Π_1 under any other choice. As a result, $(x^*, \Pi_1 - \ell(x^*))$ is a best response by the agent to this schedule. The principal cannot achieve a higher payoff by deviating to any other schedule under any best response by the agent, because by choosing (π_1, μ_1^P) total surplus is maximized and the agent receives his maxmin loss Π_1. Hence, there exists an equilibrium in which the principal chooses schedule (π_1, μ_1^P) and the agent chooses $(x^*, \Pi_1 - \ell(x^*))$ along the equilibrium path. This equilibrium satisfies $m^{P*} - m^{A*} = \ell(x^*) - \Pi_1$.

Suppose there were an equilibrium implementing x^* which satisfied $m^{P*} - m^{A*} < \ell(x^*) - \Pi_1$. Then the agent's equilibrium loss would exceed Π_1, while by deviating to $(x, m_1^A) = (0, 0)$ the agent could obtain a loss no greater than Π_1. So no such equilibrium can exist. In the other direction, suppose there were an equilibrium implementing x^* which satisfied $m^{P*} - m^{A*} > \ell(x^*) - \Pi_1$. Then the agent's on-path loss is $\Pi_1 - \epsilon$ for some $\epsilon > 0$, and the principal's equilibrium profit is $g_1(x^*) - \ell(x^*) + \Pi_1 - \epsilon$. Suppose the principal deviates to the schedule given by $\mu_1^P(x, m_1^A) = 0$ for all x and m_1^A and

$$\pi_1(x, m_1^A) = \begin{cases} 0, & x = x^* \text{ and } m_1^A = \Pi_1 - \ell(x^*) - \epsilon/2 \\ \Pi_1, & \text{otherwise} \end{cases}$$

Under this schedule, the agent incurs a loss of $\Pi_1 - \epsilon/2$ by choosing $(x^*, \Pi_1 - \ell(x^*))$, and a loss of at least Π_1 under any other choice. As a result, $(x^*, \Pi_1 - \ell(x^*))$ is the agent's unique best response to the deviation schedule and must be chosen in equilibrium. Under this deviation, the principal's profits are $g_1(x^*) - \ell(x^*) + \Pi_1 - \epsilon/2$, so this deviation is profitable. Hence, again, no such equilibrium can exist.

Case 2: $\ell(x^*) > \Pi_1$. Let

$$\pi_1(x, m_1^A) = \begin{cases} 0, & x = x^* \\ \Pi_1, & \text{otherwise} \end{cases}$$

and

$$\mu_1^P(x, m_1^A) = \begin{cases} \ell(x^*) - \Pi_1 & x = x^* \\ 0, & \text{otherwise} \end{cases}$$

Under this schedule, the agent incurs a loss of Π_1 by choosing $(x^*, 0)$, and a loss of at least Π_1 under any other choice. Hence, $(x^*, 0)$ is a best response by the agent to this schedule. As in the previous case, the principal cannot profitability deviate from this schedule anticipating any best response by the agent. Hence, there exists an equilibrium in which the principal chooses schedule (π_1, μ_1^P) and the agent chooses $(x^*, 0)$ along the equilibrium path. This equilibrium satisfies $m^{P*} - m^{A*} = \ell(x^*) - \Pi_1$. The proof that all equilibria satisfy this identity follows along nearly identical lines to the previous case.

A.7 Proof of Proposition 7

Let $\boldsymbol{\iota}^* = (\boldsymbol{\pi}^*, \boldsymbol{m}^{P*})$ be the purported equilibrium influence schedules. Since x^* is a best response to $\hat{\boldsymbol{\mu}}^P$ by hypothesis, $(x^*, \boldsymbol{\Pi})$ must be a best response to $\boldsymbol{\iota}^*$. We select this best response. To complete the proof, we establish that under any best responses for the agent for influence schedules other than $\boldsymbol{\iota}^*$, no principal has a profitable unilateral deviation from $\boldsymbol{\iota}^* = (\boldsymbol{\pi}^*, \boldsymbol{m}^{P*})$. Since the agent must have a best response to every profile of influence schedules given compactness of the action spaces and lower semicontinuity of the influence schedules, the claimed equilibrium exists.

Fix a principal i. Lemma 2(iv) of Bernheim and Whinston (1986) establishes

existence of a policy x^0 such that $\hat{\mu}_i^P(x^0) = 0$ and

$$\ell(x^0) - \sum_{j \neq i} \hat{\mu}_j^P(x^0) = \ell(x^*) - \sum_{j=1}^N \hat{\mu}_j^P(x^*).$$

Consider a deviation by the principal to schedule $\iota_i' = (\pi_i', \mu^{P'}) \neq (\pi_i^*, \mu^{P*})$, and let $(x', \boldsymbol{m}^{A'})$ be any best response by the agent to the influence schedules $(\iota_i', \boldsymbol{\iota}_{-i}^*)$. This choice must yield a loss to the agent no larger than the loss of choosing policy x^0 and making no contributions. This latter choice yields a loss no larger than

$$\ell(x^0) - \sum_{j \neq i} \hat{\mu}_j^P(x^0) + \sum_{j=1}^N \Pi_j = \ell(x^*) - \sum_{j=1}^N \hat{\mu}_j^P(x^*) + \sum_{j=1}^N \Pi_j.$$

It follows that $(x', \boldsymbol{m}^{A'})$ cannot be a best response unless its loss is bounded above as:

$$\begin{aligned} &\ell(x') + m_i^{A'} + \pi_i^*(x', \boldsymbol{m}^{A'}) - \mu_i^{P'}(x', \boldsymbol{m}^{A'}) + \sum_{j \neq i} \Big(\max(\Pi_j, m_j^{A'}) - \hat{\mu}_j^P(x') \Big) \\ \leq\ &\ell(x^*) - \sum_{j=1}^N \hat{\mu}_j^P(x^*) + \sum_{j=1}^N \Pi_j. \end{aligned}$$

This bound in turn implies the weaker inequality

$$\ell(x') + m_i^{A'} - \mu_i^{P'}(x', \boldsymbol{m}^{A'}) - \sum_{j \neq i} \hat{\mu}_j^P(x') \leq \ell(x^*) - \sum_{j=1}^N \hat{\mu}_j^P(x^*) + \Pi_i.$$

Meanwhile, Lemma 2(iii) of Bernheim and Whinston (1986) ensures that

$$\ell(x') - \ell(x^*) \geq g_i(x') - g_i(x^*) + \sum_{j \neq i} \big[\hat{\mu}_j^P(x') - \hat{\mu}_j^P(x^*) \big].$$

Combining these two inequalities yields the bound

$$g_i(x') - \mu_i^{P'}(x', \boldsymbol{m}^{A'}) + m_i^{A'} \leq g_i(x^*) - \hat{\mu}_i^P(x^*) + \Pi_i.$$

The left-hand side of this inequality is an upper bound on principal i's payoff under ι_i' and choice $(x', \boldsymbol{m}^{A'})$ by the agent. (It may exceed the true payoff if the principal

additionally inflicts a costly punishment.) Meanwhile, the right-hand side is her payoff under the schedule ι_i^* and choice $(x^*, \mathbf{\Pi})$ by the agent. Hence, the deviation is not profitable under the response $(x', \boldsymbol{m}^{A'})$. Since this argument held for an arbitrary best response of the agent, the desired claim holds.

References

Aidt, Toke S., Facundo Albornoz, and Esther Hauk (2018). "Foreign Influence and Domestic Policy". *Journal of the European Economic Association* 16 (1), pp. 1–45.

Antràs, Pol and Gerard Padró i Miquel (2025). "Exporting Ideology: The Right and Left of Foreign Influence". *Quarterly Journal of Political Science* 20 (1), pp. 33–70.

Bernheim, B. Douglas and Michael D. Whinston (1986). "Menu Auctions, Resource Allocation, and Economic Influence". *Quarterly Journal of Economics* 101 (1), pp. 1–31.

Broner, Fernando, Alberto Martin, Josefin Meyer, and Christoph Trebesch (2025). *Hegemonic Globalization and International Alignment*. Tech. rep. 1483. Centre for Economic Policy Research (CEPR) / Barcelona School of Economics Working Paper.

Chwe, Michael Suk-Young (1990). "Why Were Workers Whipped? Pain in a Principal–Agent Model". *The Economic Journal* 100 (403), pp. 1109–1121.

Clayton, Christopher, Matteo Maggiori, and Jesse Schreger (2025a). "A Framework for Geoeconomics". *Econometrica*, forthcoming.

Clayton, Christopher, Matteo Maggiori, and Jesse Schreger (2025b). "Putting Economics Back Into Geoeconomics". *NBER Macroeconomics Annual* 40.

Dal Bó, Ernesto, Pedro Dal Bó, and Rafael Di Tella (2006). "Plata o Plomo?: Bribe and Punishment in a Theory of Political Influence". *American Political Science Review* 100 (1), pp. 41–53.

Dixit, Avinash, Gene M. Grossman, and Elhanan Helpman (1997). "Common Agency and Coordination: General Theory and Application to Government Policy Making". *Journal of Political Economy* 105 (4), pp. 752–769.

Dutta, Rohan, David K. Levine, and Salvatore Modica (2018). "Damned if You Do and Damned if You Don't: Two Masters". *Journal of Economic Theory* 177, pp. 101–125.

Grossman, Gene M. and Elhanan Helpman (1994). "Protection for Sale". *American Economic Review* 84 (4), pp. 833–850.

Kleinman, Benny, Ernest Liu, and Stephen J. Redding (2024). "International Friends and Enemies". *American Economic Journal: Macroeconomics* 16 (4), pp. 350–385.

Konrad, Kai A. (2024). "Dominance and Technology War". *European Journal of Political Economy* 81, p. 102493.

Madsen, Erik, Basil Williams, and Andrzej Skrzypacz (2025). "Performance Incentives with Multiple Instruments". Working paper.

Miller, David A. and Kareen Rozen (2014). "Wasteful Sanctions, Underperformance, and Endogenous Supervision". *American Economic Journal: Microeconomics* 6 (4), pp. 326–361.

Prat, Andrea and Aldo Rustichini (2003). "Games Played Through Agents". *Econometrica* 71 (4), pp. 989–1026.

Thoenig, Mathias (2023). *Trade Policy in the Shadow of War: Quantitative Tools for Geoeconomics*. Discussion Paper DP18419. London and Paris: Centre for Economic Policy Research (CEPR).