Monitoring Corruption and Overcoming the Collective Action Problem: A Unified Model*

Torben Behmer† Michael Denly‡

May 15, 2026

Abstract

Principal-agent and collective-action theories anchor the two most common approaches to corruption, but scholars often treat them as competing explanations. We integrate principal-agent, collective-action, and functionalist insights by embedding monitoring into a multiple-equilibria, collective-action model of citizen behavior. It captures the limits of top-down supervision under systemic corruption, while shifting collective-action accounts from high-level equilibria toward citizen-level beliefs and payoffs. The model does so by disaggregating citizen-level monitoring benefits into collective and private types. Beyond identifying the relevant equilibria, we show how their selection and stability depend on citizens' beliefs about strategic reasoning and others' willingness to take action against corruption. Our model enables collective-action accounts of corruption to explain incremental change and heterogeneous outcomes within countries or societies. The distinction between collective and private monitoring benefits also provides a heuristic to guide policy.

*Denly thanks the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program (grant ANR-17-EUR-0010) for financial support. For advice and feedback, we thank Piret Avila, Peter Bayer, Terry Chapman, Theodore Charm, Lisa Faessler, Megan Farrell, Mike Findley, Michael Gibbs, Pedro Hemsley, Nate Jensen, Juan Sebastian Lozano, Xiaobo Lü, Luis Fernando Medina, Mobin Piracha, Pablo Querubin, Nimrah Siddiqui, Adi Tantravahi, Scott Wolford, and participants at the French Evolutionary Society for Human Sciences Annual Conference. All errors are those of the authors. All opinions, findings, and conclusions expressed in this paper are those of the authors alone and do not represent the official views or positions of the Office of the Governor of Texas.

†Researcher, Office of the Governor of Texas ✉ tbehmer@gmail.com

‡Assistant Professor, Texas A&M University ✉ mdenly@tamu.edu

Each year, corruption costs the world economy about USD 3.6 trillion or 5% of world gross domestic product (GDP), and bribes account for about USD 1 trillion of that prodigious sum of money (United Nations, 2018). Research is also clear that the most susceptible to paying bribes are the politically powerless and the poor, especially given their high levels of interaction with the state (Justesen and Bjørnskov, 2014; Peiffer and Rose, 2018; Robinson and Seim, 2018). Corruption is thus a costly and regressive phenomenon, afflicting nearly every country and sector of the world economy. What, then, reduces corruption?

In this paper, we examine how a canonical solution, monitoring, can reduce corruption. Since at least the 1990s, a large part of anti-corruption efforts has involved monitoring in line with the principal-agent model (Naím, 1995; Fukuyama and Recanatini, 2021). At its core, the principal-agent model suggests that it is possible to mitigate corruption through monitoring and sanctioning. However, influential research questions principal-agent approaches, suggesting that there is a shortage of people with the willingness to monitor and sanction corruption when it is systemic (e.g., Persson, Rothstein and Teorell, 2013).[1] This shortage of what Peiffer and Alvarez (2016) call "principled principals" is mainly not due to resource constraints but informal institutions, such as norms, expectations, and beliefs. These informal institutions tend to be sticky, making it difficult for even the best-designed anti-corruption reforms to change trajectories in corrupt societies through monitoring. That is a key reason why many scholars suggest that corruption is more of a collective action problem (e.g., Mungiu-Pippidi, 2015).[2]

Although the literature often implies that norms-based collective action approaches and top-down principal-agent approaches are antithetical, we unify them using a game-theoretic model. Our key contribution entails embedding monitoring into a collective-action model that takes into account both the drawbacks of principal-agent approaches and benefits of

[1]For related earlier work, see Andvig and Moene (1990), Aidt (2003), and Kingston (2008). For related empirical work documenting how rule-breaking is contagious and generally higher in corrupt societies, see Gächter and Schulz (2016) and Shalvi (2016).

[2]In more technical terms, we are referring to a second-order collective-action problem. See Ostrom (1998), Rothstein (2011*a*,*b*) and Persson, Rothstein and Teorell (2013).

incremental approaches (see Levy, 2014; Bersch, 2016; Taylor, 2018). In the process, we also address some functionalist critiques (e.g., Marquette and Peiffer, 2018).

The model's setup builds directly on Persson, Rothstein and Teorell's (2013, 457) interpretation of corruption as a collective-action problem that resembles an assurance game.[3] As such, multiple outcomes are possible, and the game does not devolve into a pre-determined, single-equilibrium prisoner's dilemma or game of harmony.[4] Beyond deriving payoffs and proving the existence of multiple equilibria, we also examine equilibrium selection: why citizens may converge on high- or low-corruption outcomes when both remain possible. To do so, we incorporate Schelling's (1978) fundamental insight from focal points and tipping games: collective action depends on contingent behavior (Fisman and Golden, 2017, 5). We operationalize contingent behavior, and how it varies across individuals, through beliefs about strategic reasoning and others' willingness to take costly action against corruption. Then, we apply Medina's (2007) variant of Harsanyi and Selten's (1988) tracing procedure to show how monitoring can shift the threshold separating high- and low-corruption equilibria.

Especially given that prominent collective-action accounts emphasize rare, "big-bang" reforms (e.g., Rothstein, 2011*b*), our model assists with understanding the role of incremental monitoring. Like Stephenson (2020), our model implies that big-bang reforms may be unnecessary to reduce corruption, but our solution is different. Whereas Stephenson (2020) questions whether self-reinforcing corruption necessarily entails multiple equilibria, we retain the multiple-equilibria structure emphasized by the collective-action canon. In doing so, we build on Marquette and Peiffer (2018, 2019), Persson, Rothstein and Teorell (2019), and Rothstein (2021), who argue that monitoring and collective action can be complementary.

Substantively, a key contribution of our model entails distinguishing between citizen-level monitoring benefits that are collective versus private. Collective benefits accrue broadly when monitoring raises the expected detection or sanctioning of corruption, thereby encouraging more citizens to take action against corruption. Nevertheless, collective benefits remain

[3]See also Fisman and Golden (2017, 6).
[4]See Appendix D for an overview of basic collective action games.

highly vulnerable to "unprincipled principals" when corruption is systemic (Persson, Rothstein and Teorell, 2013; Brierley, 2020). By contrast, private benefits, such as whistleblower protections and other support from diagonal and horizontal accountability actors, only accrue directly to citizens who take action against corruption. To be clear, external actors may face their own "unprincipled principals" and credibility problems (Ostrom, 1990, 17; Rothstein, 2021, Ch. 7). Still, they need not share the same norms, incentives, or constraints as citizens and local officials embedded in systemic corruption (e.g., Bersch, Praça and Taylor, 2017). By making this distinction, we address existing collective-actions accounts' focus on society-level social contracts and phenomena over ones at a more local level. That allows our model to better capture functionalist approaches as well as successful auditing and oversight studies more commonly associated with principal-agent logic (e.g., Olken, 2007; Ferraz and Finan, 2008; Marquette and Peiffer, 2018; Bersch, 2019; Lagunes, 2021). Additionally, our distinction between collective and private monitoring benefits provides a heuristic for policy design.

1. Corruption as a Principal-Agent Problem

The most common definition of corruption is the "misuse of public office [or entrusted power] for private gain" (e.g. Treisman, 2000; Transparency International, 2009).[5] The principal-agent model, along with most monitoring efforts, follows directly from that definition (Ugur and Dasgupta, 2011; Rose-Ackerman and Palifka, 2016, 9).

Under the first manifestation of the principal-agent model, a politician or highly-ranked bureaucrat (i.e., the principal) is entrusted with power to perform certain tasks, assuming that he/she will not misuse her power for private gain. Because the principal cannot ac-

[5]As Dixit (2016) explains, this definition of corruption maps well onto bribery, which entails a supply side (i.e., those providing the bribes—the private sector) and a demand side (i.e., those accepting/requesting the bribes—the public sector). In reality, though, corruption entails much more than bribery. For example, corruption entails kickbacks, coercion/extortion, nepotism, cronyism, financial fraud, electoral fraud, collusion, obstruction, and patronage (Søreide, 2014). For more on the definition of corruption, refer to Rose-Ackerman and Palifka (2016) and Fisman and Golden (2017).

complish all of the tasks by herself, the principal delegates at least some of these tasks to lower-ranking bureaucrats (i.e., the agents). To ensure that the latter do not misuse their monopoly and discretionary power for private gain, principals need to monitor the agents (Klitgaard, 1988).

Under the second manifestation of the principal-agent model, the politician or high-ranking bureaucrats assume the role of agent, and the public becomes the principal (Vaubel, 2006; Marquette and Peiffer, 2018). In democracies, the public can sanction corrupt agents by voting them out of office. In dictatorships, the public might rebel against the corrupt politicians or bureaucrats in whatever way possible, such as via protests.

The main issue with either manifestation of the principal-agent model is that it requires "benevolent" or "principled" principals: that is, bureaucrats, politicians, or a public with the will and capacity to monitor and sanction corrupt actors (Aidt, 2003; Peiffer and Alvarez, 2016). In societies where corruption is the predominant equilibrium behavior, "principled principals" with such qualities are in short supply. In most instances, shared norms, expectations, beliefs, and other informal institutions tend to be stronger than any formal institutions designed to control corruption (Collier, 2000; Fisman and Golden, 2017). That is why, according to an influential article by Persson, Rothstein and Teorell (2013), most monitoring-based anti-corruption reforms and programs fail when corruption is systemic.

2. Corruption as a Collective Action Problem

A collective action problem "arise[s] when the individual pursuit of self-interest generates socially undesirable outcomes" (Ferguson, 2013, 4). Corruption is a collective action problem because most people in corrupt societies would benefit from having less corruption. By the same token, resisting corruption—and contributing to the public good—is not in most people's individual self-interest. That is not just the case for the select people who financially benefit from corruption but also for its victims. Taking any type of action against corruption, such as refusing to pay a bribe, can carry potential costs such as intimidation,

violence, inability to obtain government services, and being put on a blacklist (e.g., Kingston, 2008; Wrong, 2009). Beyond that, it is challenging to engender large-scale contributions to a public good such as reducing corruption. Notably, it is difficult for citizens to trust that the intervention can overcome prevailing informal institutions (Rothstein, 2011*b*).

Among the standard collective action games (see Appendix D), the prisoner's dilemma and assurance game provide the most compelling macrostructural manifestations of corruption as a collective action problem (Yap, 2013; Dixit, 2018).[6] In their most basic forms, both the prisoner's dilemma and the assurance game involve two players. Each player's essential task is to (simultaneously) make a decision about how to behave without being able to coordinate with the other side. In the case of corruption, the preferred collective decision is to reject it, but it becomes privately optimal to accede and tolerate corruption when the other side has strong incentives to do so as well.

Although the prisoner's dilemma elucidates how citizens may want to free-ride on others' actions against corruption, the model has a significant flaw that makes it less useful for describing corruption than an assurance game. Because the prisoner's dilemma necessarily entails a dominant strategy and unique equilibrium at defection (see Appendix D), it cannot explain variation in the outcome of whether citizens take action against corruption (Persson, Rothstein and Teorell, 2013). By contrast, the assurance game can explain such variation because of its multiple equilibria: one pure strategy Nash equilibrium at both parties taking action ($PSNE_{Cooperate}$), another at doing nothing ($PSNE_{Defect}$), and a mixed-strategy Nash equilibrium (MSNE).[7]

The stock assurance game, however, is insufficient to explain how citizens' calculations about whether to take action against corruption can change. Fisman and Golden (2017, 5)

[6]Ostensibly, the game of harmony does not describe corruption well in corrupt countries, because incentives to defect and thus free-ride on the contributions of others are still present; otherwise, corruption would not be so prevalent across the world. Deadlock is also inappropriate, since the payoffs of defecting are not that high for all citizens; if so, there would be no reason to mitigate corruption. The game of chicken is similarly uncompelling for describing corruption: taking action is the challenge, and chicken assumes that someone always takes action. Notably, there are two pure strategy Nash equilibria in Chicken. See Appendix D.

[7]For a gentle introduction to basic game theory, see Dixit, Skeath and Reiley (2014) and Humphreys (2017).

correctly suggest that it relates to Schelling's (1978) fundamental insight from focal points, tipping games, and threshold models on contingent behavior: that is, a person's decision to take action against corruption depends on her belief about whether others will reciprocate (Dong, Dulleck and Torgler, 2012; Lee and Guven, 2013; Banerjee, 2016; Gneezy, Saccardo and van Veldhuizen, 2019; Sundström, 2019).[8]

A main drawback of focal points and tipping games is that expectations and other determinants of contingent behavior vary with events in the external environment (Medina, 2007, 59-61). For example, even social trust, which numerous authors argue is the most important determinant of corruption (e.g., Rothstein, 2021), changes according to a new leader taking power, repression, scandals, etc. By construction, therefore, prediction is difficult when focal points and tipping games have multiple equilibria (Medina, 2007, 60; Medina, 2018, 47).

To address these challenges, we shift from identifying possible equilibria to analyzing equilibrium selection and stability. Our approach uses a variant of Harsanyi and Selten's (1988) tracing procedure to vary the weight citizens place on strategic reasoning versus initial beliefs about others' likely behavior. By doing so, we can show when monitoring-generated benefits make costly action against corruption more likely without turning the assurance game into one of harmony or eliminating the high-corruption equilibrium altogether. In this way, we follow Medina's (2007, 7) lay person's theory of collective action:

> "When individuals can achieve some beneficial result by coordinating in a group, they are likely to coordinate. As the potential benefits of coordination increase (or the costs decrease), these individuals are more likely to coordinate and, conversely, as the potential benefits decrease (or the costs increase), they are less likely to do so."

[8]For a threshold model of revolution, see Kuran (1989).

3. A Model of Corruption as a Collective Action Problem

We model collective action against corruption as a one-shot assurance game between two strategic and representative citizens, i and j, as well as an unstrategic bureaucrat, who always solicits a bribe. Let Γ^A denote the baseline game that we use as a point of departure for subsequent games Γ^B, Γ^C, and Γ^D. All of these games follow Persson, Rothstein and Teorell's (2013) account of systemic corruption. Citizens prefer a low-corruption outcome if others also cooperate. However, prevailing norms make corruption the default and unilateral cooperation costly. By contrast, defection has none of these costs but contributes nothing to the public good. For the sake of simplicity, we represent the costly actions as refusing to pay a bribe or reporting a corrupt bureaucrat, but our model can apply to other anti-corruption efforts as well. Because we aim to isolate how monitoring changes payoffs when citizens act simultaneously and cannot observe or coordinate with one another, our model does not incorporate group-size thresholds, iteration, or multiplayer structures.[9] Against this backdrop, our model does not constitute a hidden-action or hidden-information principal-agent game, and does not feature signaling or updating of beliefs.

3.1. The Basic Setup: Collective Action without Monitoring

Keeping with Medina's (2007) notation for collective action games, we denote payoffs for player i as W_1 when both sides cooperate; W_2 where player i defects but player j cooperates; W_3 where player i cooperates but player j defects; and, finally, W_4 where both players defect. Consistent with traditional assurance game payoffs,[10] and keeping all component parameters non-negative in line with Equation (B.2), we define W_1 through W_4 such that:

[9] Substantively, corruption is a clandestine behavior that provides a very low signaling environment to facilitate citizens' short-term preference falsification (see Kuran, 1991). Additionally, group size thresholds for overcoming corruption are context-dependent, and relevant collective action scholarship offers no firm guidance as well (Sandler, 2015; Weimann et al., 2019).

[10] Readers unfamiliar with payoff structures of basic collective action games may refer to Appendix D.

- W_1: If and only if citizen i refuses to pay the bribe (i.e., she "cooperates"), and another citizen j joins her in doing so. By refusing to pay the bribe, she (most likely) forgoes the opportunity to obtain the demanded service, but provides a public good. As citizen j matches her behavior, the benefits from the public good become more pronounced ($u_{public\ good} = t$). She does, however, face the risk of retaliation from the bureaucrat, for which she pays a cost, r.[11] Mathematically, her payoff is: $W_1 \equiv u_i(C_i|C_j) = t - r > 0$.

- W_2: If and only if citizen i pays the bribe (i.e., she "defects"), but citizen j opts to cooperate, thus providing for a corruption-fighting public good that citizen i can free-ride on. By citizen i not supplying the public good herself, its benefit is reduced ($u_{public\ good} = s, t > s > 0$). By paying the bribe, however, citizen i virtually guarantees receipt of the service, which she values at g. The (immediate financial) cost of her bribe is denoted by b. In line with functionalist scholarship on corruption (Marquette and Peiffer, 2018), she also avoids the problem of retaliation from the bureaucrat or anyone else by paying the bribe. Hence, citizen i's payoff is: $W_2 \equiv u_i(D_i|C_j) = s + g - b > 0$.

- W_3: If and only if citizen i cooperates by refusing to pay the bribe, thus investing into the public good, but her efforts are not matched by citizen j. As previously noted, refusal to pay results in foregone reception of services and possible retaliation from the bureaucrat requesting the bribe. By singularly investing in the public good, her benefits from public good are small (s). In total, the action yields: $W_3 \equiv u_i(C_i|D_j) = s - r$.

- W_4: If and only if neither citizen i nor citizen j cooperate, no investment in the public good occurs. More concretely, both citizens receive their services in return for paying the bribe. To center the game at zero, we assume a functionalist status-quo outcome that produces no individual net gain or loss for either citizen i or others represented by j, where the bureaucrat can extract the exact value of the good: $W_4 \equiv u_i(D_i|D_j) =$

[11] Beyond refusing to pay the bribe, the citizen could further choose to report the bureaucrat, thus producing a perhaps greater public good ($t' \geq t$), while also incurring larger costs from potential retaliation than refusing to pay a bribe ($r' > r$). As long as $t' > r'$, the assurance game setup is retained. Comparative statics on t and r provide the information needed to further study the impact of this change (see Appendix C).

$g - b = 0$.[12]

For version Γ^A to be an assurance game in line with Persson, Rothstein and Teorell (2013), first, the potential costs of retaliation must exceed those of singular cooperation (i.e., $r > s$). Second, despite the threat of retaliation, the societal benefits of successful collective action must eclipse the payoff from singular defection (i.e., $W_1 > W_2$, or $t - r > s + g - b$).[13] Taken together, any of the games covered here are assurance-type games if and only if $W_1 > W_2 > W_4 > W_3$. As a result, Γ^A through Γ^D must produce three equilibria: Two pure strategy Nash equilibria ($PSNE_{Cooperate}$ and $PSNE_{Defect}$), and one mixed-strategy Nash equilibrium (MSNE, denoted by $\alpha^\star$). As we show in Appendix B.2, the existence of the PSNEs is contingent on these aforementioned assumptions. For version Γ^A, as shown in Appendix B.3, the MSNE identifies the cooperation threshold at which citizens are indifferent between taking costly action and doing nothing, denoted by α_i (or α_j), where:

$$\alpha_i^{\star,A} = \frac{g - b + r - s}{t - 2s} \tag{1}$$

As Medina (2007, 147) shows, we can represent the MSNE more generally as:

$$\alpha_i^\star = \frac{W_3 - W_4}{W_2 - W_4 + W_3 - W_1} \tag{2}$$

Once again we can assume that the (unstrategic) bureaucrat is capable of extorting the full

[12] Any potential bargain over the size of the bribe amount given the value of the provided good depicts a hierarchical divide-the-dollar type of interaction between *strategic* bureaucrats and citizens. As stated earlier, the bureaucrats in this game are nonstrategic. This paper thus simply assumes that the bureaucrat can extort the full value of the public good, i.e., $g - b = 0$. From a substantive perspective, centering does not imply that bribery is harmless but normalizes the status quo payoff in the face of systemic corruption.

[13] For some, this assumption might be considered a strong one. Indeed, forgone services might be at the forefront of many citizens' thoughts as they contemplate the decision to refuse a bribe, despite the undeniable citizen benefits of systemically fighting corruption. However, without this assumption, the game defaults to a prisoner's dilemma. As we emphasize earlier, the prisoner's dilemma does not adequately describe the problem of corruption due to a lack of variation in the dependent variable.

value of the public good (i.e., $g - b = 0$), so the MSNE is:

$$\alpha_i^{\star,A} = \frac{r - s}{t - 2s} \tag{3}$$

3.2. The Augmented Setup: Collective Action with Monitoring

Building on the basic setup above, we introduce two functionally distinct sets of monitoring benefits: collective and private. Both can alter citizens' decision calculus regarding whether or not to take costly action against corruption, but they enter citizens' payoffs differently.

3.2.1. Collective Benefits of Monitoring

The collective benefits of monitoring (c) accrue to all citizens so long as at least one of them cooperates by investing in the public good of taking costly action against corruption. For example, a report of corruption may become more credible when monitoring creates an evidentiary record, thereby increasing the chance of sanctioning and citizens' willingness to take action against corruption. However, citizens may perceive collective benefits to be small, as they remain vulnerable to "unprincipled principals" and other professionalism concerns about bureaucrats. In this sense, our operationalization of the collective benefits of monitoring aims to capture both the specifics of the setting and how the larger social contract feeds into them (see Rothstein, 2021).

Formally, let Γ^B denote the game introducing c (see Appendix B.4 for the normal form game). In Γ^B, c increases W_1, W_2, and W_3, whereas W_4 remains constant due to lack of citizen cooperation. Maintaining the assurance-game structure requires $s - r + c < 0$,[14] and keeping with the centering of $g - b = 0$, the MSNE becomes (see Appendices B.5 and B.6 for proofs):

[14]Note that the assumption of $s - r + c < 0$ is more stringent than $s - r < 0$, because $c \in (0, \infty)$.

$$\alpha_i^{\star,B} = \frac{r - s - c}{t - 2s - c}, \quad \text{where } c > 0, \alpha_i^{\star,B} < \alpha_i^{\star,A} \tag{4}$$

3.2.2. Private Benefits of Monitoring

In contrast to collective benefits, the private benefits of monitoring (p) only accrue to citizens who actually invest in the public good at personal cost. Examples of such benefits include those facilitated by horizontal and diagonal accountability:[15] whistleblower protections, insulated reporting or resolution channels (e.g., ombudsmen), and support from external actors (e.g., NGOs, international organizations). All of these monitoring benefits are private because they can only accrue to the cooperator. Of course, external actors facilitating private monitoring benefits may face their own "unprincipled principals" and credibility problems (Ostrom, 1990, 17; Rothstein, 2021, Ch. 7). Even so, their norms, incentives, and constraints often differ from those prevailing in the institution where the corrupt transaction occurs. In making this level-of-analysis distinction (see Gingerich, 2013), we connect society-level collective action accounts to functionalist ones as well as auditing and oversight studies more commonly associated with principal-agent logic (e.g., Olken, 2007; Ferraz and Finan, 2008; Marquette and Peiffer, 2018; Bersch, 2019; Lagunes, 2021).

Formally, let Γ^C represent the game introducing p, which only increases W_1 and W_3 (see Appendix B.7 for the normal form game). Consistent with Γ^B, the inclusion of private benefits to those cooperating in Γ^C alters W_3, while holding W_4 constant. As with collective benefits, we do not assume that private protection fully solves corruption or eliminates the high-corruption equilibrium. If citizens perceive these protections as non-credible, p approaches zero, and when citizens view them as credible, p lowers the cost of potential retaliation. Maintaining the structure of the assurance game, and centering around $W_4 = 0$, $s - r + p < 0$, the MSNE is (see Appendices B.8 and B.9 for proofs):

[15] See Lührmann, Marquardt and Mechkova (2020) for distinctions between vertical, horizontal, and diagonal accountability.

$$\alpha_i^{\star,C} = \frac{r-s-p}{t-2s}, \quad \text{where } p > 0, \alpha_i^{\star,C} < \alpha_i^{\star,A} \tag{5}$$

3.2.3. Collective and Private Benefits of Monitoring

Finally, consider the scenario in which there are both collective and private monitoring benefits. In most real-world settings, both forms of monitoring benefits exist: bureaucracies usually contain supervisory hierarchies, and most countries have some form of horizontal or diagonal accountability. The key question is not whether such mechanisms exist on paper, but whether citizens perceive c and p as large enough to alter their expected payoffs. For purposes of generality, we do not claim that private monitoring benefits are greater than collective ones or vice-versa. While we consider such scenarios in Appendix B.14, we maintain that neither is *a priori* greater than the other given the context-specific nature of corruption.

Formally, let Γ^D denote the game with both collective and private monitoring benefits that accrue to citizens if cooperation takes place (see Appendix B.10). Under Γ^D, W_1 and W_3 increase by $c+p$, whereas W_2 increases by c, and W_4 remains unchanged when compared to Γ^A. To maintain assurance game structure, we assume $r-s > c+p$ (see Appendix B.12). Because W_1 increases by $c+p$, and W_2 increases by c, the basic structure of the game remains in place. Appendix B.11 provides the proofs for the pure strategy equilibria. Solving for the MSNE (see Appendix B.12), and assuming a game centered around $W_4 = g - b = 0$, we obtain:

$$\alpha_i^{\star,D} = \frac{r-s-p-c}{t-2s-c}, \quad \text{where } \forall\ p, c > 0\text{: } \ \alpha_i^{\star,D} < \alpha_i^{\star,B} < \alpha_i^{\star,A} \text{ and } \alpha_i^{\star,D} < \alpha_i^{\star,C} < \alpha_i^{\star,A} \tag{6}$$

4. Comparative Statics and Best Responses

The preceding subsections derive each game's threshold MSNE, α_i^*, which in this section we use to compute comparative statics for the monitoring parameters, c and p. Appendix C

provides the proofs for the fully-augmented game with both types of monitoring, Γ^D. Given that Γ^A, Γ^B, and Γ^C reflect special cases of Γ^D, in which c and/or p are zero, we do not repeat the relevant proofs for the other games.

Deriving $\alpha_i^\star$ with respect to the collective monitoring benefits, c, we obtain:

$$\frac{\partial \alpha_i^\star}{\partial c} = \frac{r + s - p - t}{(t - 2s - c)^2}. \tag{7}$$

Because the assurance-game assumptions imply $t > r+s$, and given that p and c are positive, the partial derivative of $\alpha_i^\star$ must be negative. Hence, an increase in collective benefits emanating from monitoring coincides with a shrinking MSNE. In other words, collective benefits lower the threshold for taking costly action against corruption.

The comparative statics for private monitoring benefits follow a similar logic. Recall that $t - 2s - c > 0$ is necessary to achieve a valid mixing strategy and, in turn, maintenance of all three assurance game equilibria for games Γ^B and Γ^D (see Appendices B.6 and B.9). Consequently, we can derive α_i^* with respect to p, obtaining:

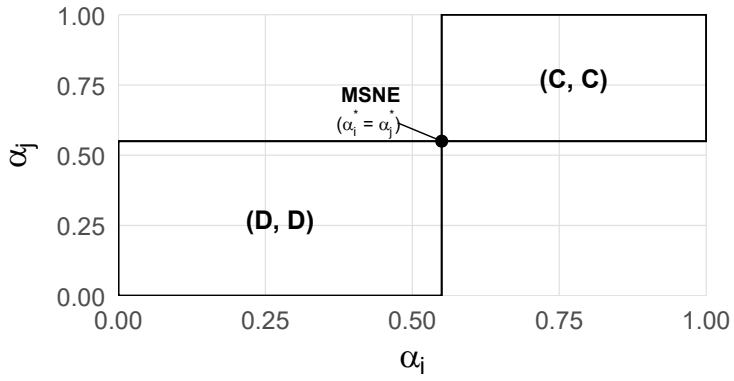
$$\frac{\partial \alpha_i^\star}{\partial p} = -\frac{1}{t - 2s - c} \tag{8}$$

Again, private benefits emanating from monitoring coincide with a shrinking mixing strategy under the MSNE. Thus, both collective and private monitoring benefits lower the threshold that separates cooperative and defecting basins of attraction. In other words, the result is not that monitoring eliminates the high-corruption equilibrium. It is that credible monitoring benefits reduce how optimistic citizens must be about others' cooperation before taking action themselves.

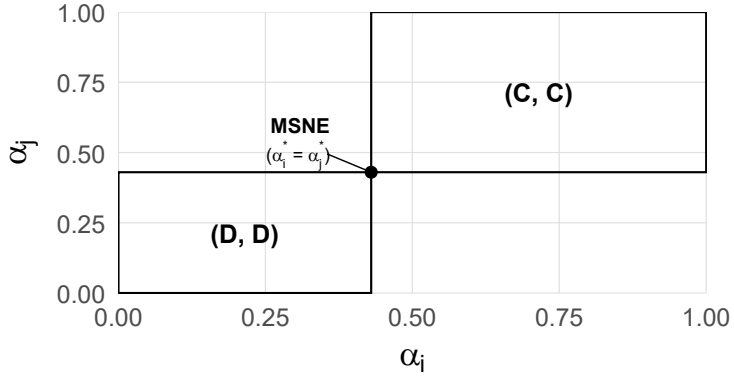
The comparative statics also yield a simple best-response interpretation (see Figure 1). For each citizen, the MSNE marks the threshold at which costly action against corruption becomes optimal. Below this threshold, defection is the best response; above it, cooperation

Figure 1: Simulated Best Response Plots

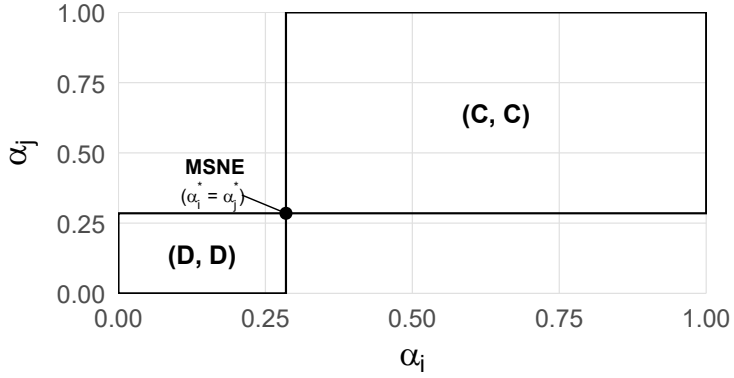
(a) Γ^A: No Monitoring Benefits



(b) Γ^B/Γ^C: Only Collective or Private Monitoring Benefits



(c) Γ^D: Collective and Private Monitoring Benefits



Note: Each panel plots the best-response regions for citizens i and j. The lower-left region corresponds to mutual defection (D, D), while the upper-right region corresponds to mutual cooperation (C, C). The amounts of mutual cooperation and mutual defection in Γ^A are not the same in every context, so Figure 1a merely represents a simulated baseline with which to compare the other games. The relative effects of c and p depend on how citizens value each form of monitoring, which is context-dependent, so we do not distinguish between Γ^B and Γ^C and represent them in Figure 1b. In this sense, the figure is illustrative rather than calibrated. Regardless of the sizes c and p, their combination in Γ^D exceeds the benefits of other games (see Figure 1c).

is the best response. In the joint strategy space, the two citizens' cutoffs divide the unit square into regions of mutual defection, mutual cooperation, and asymmetric best responses. The lower-left and upper-right regions foreshadow the high- and low-corruption basins of attraction that the next section formalizes with stability sets, whereas the unlabeled off-diagonal regions do not define basins on their own. As monitoring benefits lower the MSNE in Figures 1b and 1c, they shrink the mutually-reinforcing defection region and expand the mutually-reinforcing cooperation region.

5. Equilibrium Selection and Basins of Attraction

While the preceding section shows that monitoring can shift the threshold between basins of attraction without transforming the assurance game into one of harmony, this section shows why these incremental shifts matter for equilibrium selection. Our approach uses stability sets (Medina, 2007). Like global games (Carlsson and van Damme, 1993), stability sets evaluate equilibria by examining their robustness to alternative starting points for how players reason. Global games rely on noisy signals and study how equilibria behave as signal noise vanishes. Stability sets instead begin with players' initial beliefs about others' behavior and then vary the weight players place on strategic reasoning. Given a game's payoffs and equilibria, stability sets use a variant of Harsanyi and Selten's (1988) tracing procedure to map those initial beliefs into basins of attraction. Such an approach is useful for studying anti-corruption reforms given that they usually do not immediately change preferences or eliminate formal equilibria. Instead, reforms can change which equilibrium citizens expect others to help sustain.

As a baseline to understand the stability set, consider the standard utility function for any game Γ^0 with multiple equilibria under full strategic rationality or, more precisely,

common-knowledge rationality ($\lambda = 1$):[16]

$$u_i^{0,\lambda=1}(\alpha_i, \alpha_j) = \alpha_i(\alpha_j W_1 + (1-\alpha_j)W_3) + (1-\alpha_i)(\alpha_j W_2 + (1-\alpha_j)W_4) \tag{9}$$

Under Equation (9), citizen i's expected utility depends on citizen j's actual mixed strategy, α_j. However, that assumption is often too strong for understanding the collective action problem in the context of corruption. Citizens confronting corruption usually do not observe others' decisions, cannot coordinate with them, and cannot know with certainty whether others will resist or accommodate corruption.

To address these real-world constraints, we introduce beliefs about others' behavior. Formally, let β_j denote citizen i's initial belief that citizen j will cooperate, which background norms or contact with bureaucratic agencies may shape. These initial beliefs and monitoring can be interrelated in practice. However, to keep the analysis focused on the mechanism of interest, our model treats beliefs as exogenous and examines whether monitoring alone is sufficient to shift equilibrium selection. Because the game remains simultaneous, and citizens do not observe one another's choices, β_j also is not a Bayesian posterior. Instead, β_j is a continuous random variable bounded between 0 and 1 with probability density function $f(\beta_j)$ and cumulative distribution function F_{β_j}.

Following Medina's (2007) adaptation of Harsanyi and Selten's (1988) tracing procedure, we can now define the stability set of a strategy (σ) as the region of initial beliefs for which that strategy is a best response (BR). The stability set of cooperation is:

$$\sigma_i(Cooperate_i) = \{\beta_j \in [0,1] : Cooperate_i \in BR_i(\beta_j)\}. \tag{10}$$

Similarly, the stability set of defection is:

$$\sigma_i(Defect_i) = \{\beta_j \in [0,1] : Defect_i \in BR_i(\beta_j)\}. \tag{11}$$

[16] In game theory, common-knowledge rationality means that citizen i believes citizen j is rational; citizen j believes that citizen i is rational; citizen i believes that citizen j believes that citizen i is rational, etc.

To capture how initial beliefs shape equilibrium stability, now consider citizen i's weighted expected-utility function (see Appendix B.13):

$$\begin{aligned} u_i^{0,\lambda}(\alpha_i, \alpha_j, \beta_j) = \lambda \left[\alpha_i(\alpha_j W_1 + (1-\alpha_j)W_3) + (1-\alpha_i)(\alpha_j W_2 + (1-\alpha_j)W_4)\right] \\ + (1-\lambda)\left[\alpha_i(\beta_j W_1 + (1-\beta_j)W_3) + (1-\alpha_i)(\beta_j W_2 + (1-\beta_j)W_4)\right] \end{aligned} \tag{12}$$

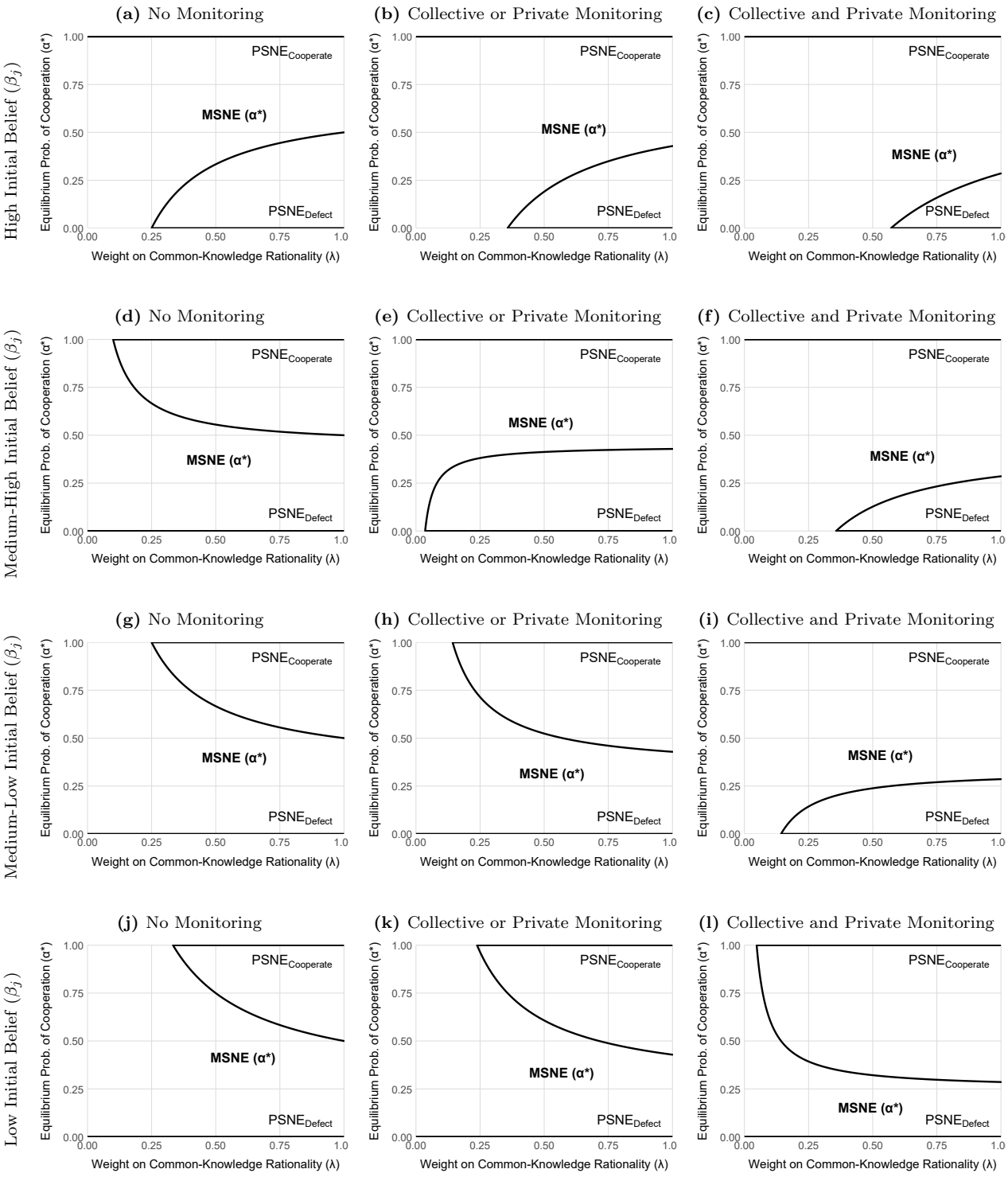
where $\lambda, \beta_j \in [0, 1]$. Equation (12) weighs two components. The first is the standard expected-utility calculation under fully strategic play. The second replaces citizen j's mixed strategy, α_j, with citizen i's initial belief that citizen j will cooperate, β_j. As common-knowledge rationality declines ($\lambda \to 0$), citizen i's expected utility depends more on initial beliefs, β_j.

The stylized trace plots in Figure 2 assist with understanding how all the parameters interact with each other. For each combination of β_j and monitoring level, under full common-knowledge rationality (i.e., $\lambda = 1$) we can verify the presence of the three assurance game equilibria: $PSNE_{Cooperate}$, $PSNE_{Defect}$, and $MSNE$ (α^*). As common-knowledge rationality declines ($\lambda \to 0$), the MSNE/indifference threshold (α^*) either decreases or increases until it approaches 0 or 1, respectively, where it becomes a PSNE. Beyond that point, common-knowledge rationality (λ) is sufficiently small for citizen i to prefer one PSNE over another due to her beliefs about citizen j's probability of cooperation, β_j.

Each row of Figure 2 illustrates how different initial beliefs about others' cooperation levels (β_j) affect coordination outcomes. Across all of its components 2a-2l, blank values for either $PSNE_{Cooperate}$ or $PSNE_{Defect}$ indicate the equilibrium is not sustained for all values of λ (conditional on β_j), thereby excluding it from the stability set.[17] Under high values of β_j in Figures 2a-2c, the low-corruption equilibrium strategy ($PSNE_{Cooperate}$) makes up the stability set regardless of the level of monitoring. The opposite is true for low values

[17] Medina (2007, 113) argues that an equilibrium is in the stability set when it provides a continuous feasible path for all values of λ. As common knowledge of rationality is phased in (rather than relaxed), players will need substantial impetus to depart from their original (optimal) strategy and opt for another optimal strategy. Although we find this logic to be most intuitive, we opted to gradually phase out for beliefs as full knowledge of rationality marks the starting point of our model (see Equation (9)).

Figure 2: Trace Plots



Note: Each panel traces the MSNE cutoff, $\alpha^{\star}$, as the weight on common-knowledge rationality, λ, varies. Blank regions indicate values outside the relevant stability set. The middle column collapses the collective-only and private-only monitoring cases into a single monitoring-benefit category.

of β_j. In Figures 2j-2l, introducing various forms of monitoring does not change the fact that only the high-corruption equilibrium is in the stability set. For these select cases where initial beliefs in others' cooperation are either exceedingly high or low, monitoring does not change the central outcome of determining which equilibrium makes up the stability set. Nonetheless, monitoring can substantively matter in those cases, too. To the extent that initial beliefs are sticky but responsive to policy, monitoring can facilitate an easier transition to a low-corruption equilibrium or avoid a retreat to a high-corruption equilibrium.

Meanwhile, the second and third rows of Figure 2 most markedly show the effect of monitoring. These rows depict sufficiently low values of β_j for the high-corruption equilibrium to be in the stability set when no monitoring is present (see Figures 2d and 2g). However, with the addition of monitoring-derived collective and/or private benefits, the low-corruption equilibrium can replace its high-corruption counterpart in the stability set (see Figures 2e, 2f, and 2i).

Let us now show the visual logic from the trace plots probabilistically. As before, the MSNE ($\alpha_i^\star$) denotes the threshold separating the high- and low-corruption equilibria. If initial beliefs (β_j) fall below this threshold, they lie in the basin of attraction of the high-corruption equilibrium, which we can calculate as follows:

$$Pr(High\ Corruption) = F_{\beta_j}(\alpha_i^\star) = \int_0^{\alpha_i^\star} f(\beta_j) d\beta_j \tag{13}$$

Alternatively, if initial beliefs are higher than $\alpha_i^\star$, they lie in the basin of attraction of the low-corruption equilibrium, which we can calculate as follows:

$$Pr(Low\ Corruption) = 1 - F_{\beta_j}(\alpha_i^\star) = \int_{\alpha_i^\star}^1 f(\beta_j) d\beta_j \tag{14}$$

Conditional on β_j having a probability distribution bounded between 0 and 1, we can also calculate the change in probability of the low-corruption equilibrium occurring across different versions of the game. For instance, moving from the baseline game, Γ^A, to the one

where citizens receive both collective and private monitoring benefits, Γ^D, we obtain:

$$\begin{aligned} &Pr(Low\ Corruption\ \Gamma^D) - Pr(Low\ Corruption\ \Gamma^A) \\ &\quad = (1 - F_{\beta_j}(\alpha_i^{\star,D})) - (1 - F_{\beta_j}(\alpha_i^{\star,A})) \\ &\quad = \int_{\alpha_i^{\star,D}}^{1} f(\beta_j)d\beta_j - \int_{\alpha_i^{\star,A}}^{1} f(\beta_j)d\beta_j \geq 0 \end{aligned} \tag{15}$$

Because monitoring lowers the mixed-strategy cutoff, it expands the set of initial beliefs for which the low-corruption equilibrium is the best response. Thus, for a given distribution F_{β_j}, monitoring (weakly) increases the probability of the low-corruption equilibrium relative to the baseline game.[18]

6. Discussion

Recently, the corruption literature has undergone a recalibration away from top-down, principal-agent approaches. In turn, numerous scholars have suggested that corruption is an equilibrium outcome (e.g., Fisman and Golden, 2017), implying that corruption is more of a collective action problem. Exploring corruption through this alternative lens has improved understanding of the basic nature of corruption and its control (Persson, Rothstein and Teorell, 2019). While big-bang-style reforms are perhaps the most well-known solutions from a collective action standpoint (e.g., Rothstein, 2011*a*), they are rare, take time, and are not the only ones available (see Appendix A). For example, the common-pool resource solutions of increasing the supply of institutions, credible commitments, and mutual monitoring can also control corruption through collective action (Ostrom, 1990; Rothstein, 2021).

Despite the clear strengths of the collective action approach, existing accounts tend to locate the relevant equilibrium at an overly high level of aggregation. By emphasizing

[18]Note that the difference of cumulative density functions of β_j is only zero when the distribution of β does not have any mass within the range of $\beta \in (\alpha_i^{\star,D}, \alpha_i^{\star,A})$. For $\beta_j \sim \text{Beta}(\alpha, \beta)$, where α and β are the distribution's shape parameters, and not parameters in our model, this is a possible outcome. Conversely, if $\beta_j \sim \text{U}(0, 1)$, all values for β_j between 0 and 1 have non-zero probability, in which case Equation (15) is strictly positive.

society-wide norms and generalized trust, existing collective action accounts are more proficient at explaining persistence than incremental change and outcome heterogeneity. In this respect, they face a level-of-analysis problem similar to the one that afflicts perception-based corruption indicators (see Gingerich, 2013). As Bersch (2016, 2019) underscores, different institutions within the same country can have very different levels of norms, constraints, and autonomy, and that matters for corruption control. Marquette and Peiffer's (2018, 2019) functionalist critique reinforces the point: corruption persists not only because citizens distrust one another, but also because it solves concrete local problems.

Our model expands upon the collective action literature's insight that corruption can operate as a multiple-equilibria assurance game. We operationalize the model and derive its implications by focusing on the micro-level beliefs and payoffs that shape equilibrium selection and stability. While doing so, we adapt an implication of Ostrom's (1990) theory from the common-pool resource setting: monitoring can generate both collective and private benefits. In our corruption setting, collective benefits are subject to "unprincipled principals" but can immediately benefit everyone. By contrast, private benefits from diagonal and horizontal accountability institutions, such as whistleblower protections, only immediately benefit cooperators. Given that corruption is always context-specific, both forms of monitoring can theoretically lower the threshold for taking costly action against corruption, α_i^*. The efficacy of each form of monitoring, however, depends on initial beliefs about other citizens' likelihood of taking action against corruption (β_j) and common-knowledge rationality (λ) in each setting. These distinctions allow the model to explain when monitoring interventions can produce incremental and heterogeneous gains, even while corruption remains entrenched elsewhere. By extension, the model can explain why some monitoring interventions move low-corruption equilibrium behavior into the stability set in some offices, regions, or bureaucratic encounters. Over time, these types of interventions can break the cycles of corruption (Levy, 2014; Bersch, 2016, 2019).

The distinction between collective (c) and private (p) benefits from monitoring also

provides a useful heuristic for policy design. Of course, both c and p may be low for *some* interventions in *some* settings, but c and p are not necessarily always both zero and/or equal to each other. Their actual size—and which one reformers should prioritize—depends less on whether an intervention copies external best practices than on whether it incorporates local knowledge about implementation bottlenecks (Andrews, Pritchett and Woolcock, 2017). In this sense, effective reforms may need to be "second-best": not theoretically ideal, but targeted to the relevant constraint in the relevant context (Rodrik, 2007). Consistent with the political settlements framework, efforts to increase c and p must confront political economy and power dynamics (Khan, 2018). Reforms that directly threaten elite pacts without building sufficient societal or institutional backing are unlikely to succeed (Dercon, 2022).

7. Conclusion

What reduces corruption? A predominant answer and policy response is top-down monitoring in line with the principal-agent model, but a now well-established collective action literature rightfully asks: what happens when the principals are unprincipled?

We devise a multiple-equilibria assurance model that integrates collective-action, principal-agent, and functionalist approaches to corruption. Our model internalizes the "unprincipled principals" critique while addressing the tendency of collective-action approaches to locate the relevant equilibria at overly high levels of aggregation. Our solution to this level-of-analysis problem is two-fold. First, we distinguish between citizen-level monitoring benefits that are collective and private. Second, our model moves beyond merely identifying equilibria to examining their selection and stability. As a result, our model can explain how monitoring and other incremental interventions can be successful without a "big-bang" reform.

Going forward, future research can build not just on the distinctions between collective and private monitoring benefits but also their underpinnings. If relevant interventions do not empirically move citizens' beliefs about whether it is worthwhile to take costly action against corruption, then perhaps another intervention may be more appropriate. In deciding

upon these interventions, the literatures on problem-driven iterative adaptation and political settlements provide useful clues: reforms must be local and take into account elite interests (Andrews, Pritchett and Woolcock, 2017; Khan, 2018). While the constraints can be difficult, and corruption can be endemic in some places, there is always within-country heterogeneity to exploit for progress (Levy, 2014; Bersch, 2016). With enough interventions along these lines, it may even be possible to attain some context-specific virtuous cycles (see Mungiu-Pippidi and Johnston, 2017).

References

Acemoglu, Daron and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Crown Business.

Aidt, Toke S. 2003. "Economic Analysis of Corruption: A Survey." *Economic Journal* 113(491):F632–F652.

Al Athmay, Alaa Aldin Abdul Rahim A. 2015. "Demographic Factors as Determinants of E-Governance Adoption." *Transforming Government: People, Process and Policy* 9(2):159–180.

Andrews, Matt, Lant Pritchett and Michael Woolcock. 2017. *Building State Capability: Evidence, Analysis, Action*. New Haven, Connecticut: Oxford University Press.

Andvig, Jens Chr and Karl Moene. 1990. "How Corruption May Corrupt." *Journal of Economic Behavior and Organization* 13(1):63–76.

Arroyo Chacón, Jennifer Isabel. 2015. "Primeros frutos en la lucha contra la corrupción en Costa Rica: condenatoria en sede penal por delitos tipificados por la Ley contra la corrupción." *Derecho Penal y Criminología* 37(101):51–86.

Baker, Bruce. 2009. "Cape Verde: Marketing Good Governance." *Africa Spectrum* 44(2):135–147.

Banerjee, Ritwik. 2016. "Corruption, Norm Violation and Decay in Social Capital." *Journal of Public Economics* 137:14–27.

Bersch, Katherine. 2016. "The Merits of Problem-Solving over Powering Governance Reforms in Brazil and Argentina." *Comparative Politics* 48(2):205–225.

Bersch, Katherine. 2019. *When Democracies Deliver: Governance Reform in Latin America*. Cambridge: Cambridge University Press.

Bersch, Katherine, Sérgio Praça and Matthew M. Taylor. 2017. Bureaucratic Capacity and Political Autonomy within National States: Mapping the Archipelago of Excellence in Brazil. In *States in the Developing World*, ed. Miguel Angel Centeno, Atul Kohli and Deborah Yashar. Cambridge: Cambridge University Press chapter 6, pp. 157–183.

Brierley, Sarah. 2020. "Unprincipled Principals: Co-opted Bureaucrats and Corruption in Ghana." *American Journal of Political Science* 64(2):209–222.

Camp, Edwin, Avinash Dixit and Susan C. Stokes. 2014. "Catalyst or Cause? Legislation and the Demise of Machine Politics in Britain and the United States." *Legislative Studies Quarterly* 39(4):559–592.

Carlsson, Hans and Eric van Damme. 1993. "Global Games and Equilibrium Selection." *Econometrica* 61(5):989–1018.

Collier, Paul. 2000. "How to Reduce Corruption." *African Development Review* 12(2):191–205.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Haakon Gjerløw, Adam N. Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Moa Olin, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Luca J. Uberti, Yi Ting Wang, Tore Wig and Daniel Ziblatt. 2018. Varieties of Democracy Codebook (V8). Technical report University of Gothenburg, V-Dem Institute Gothenburg, Sweden: .

Dercon, Stefan. 2022. *Gambling on Development: Why Some Countries Win and Others Lose*. Hurst Publishers.

Dixit, Avinash. 2016. Corruption: Supply-Side and Demand-Side Solutions. In *Development in India: Micro and Macro Perspectives*, ed. S. Mahendra Dev and P. G. Babu. New Dehli, India: Springer chapter 4, pp. 57–69.

Dixit, Avinash. 2018. Anti-Corruption Institutions: Some History and Theory. In *Institutions, Governance, and the Control of Corruption*, ed. Kaushik Basu and Tito Cordella. Cham, Switzerland: Palgrave Macmillan chapter 2, pp. 15–50.

Dixit, Avinash, Susan Skeath and David Reiley. 2014. *Games of Strategy*. New York: W.W. Norton and Company.

Dong, Bin, Uwe Dulleck and Benno Torgler. 2012. "Conditional Corruption." *Journal of Economic Psychology* 33(3):609–627.

Ferguson, William. 2013. *Collective Action and Exchange: A Game-Theoretic Approach to Contemporary Political Economy*. Stanford, California: Stanford University Press.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123(2):703–745.

Fisman, Raymond and Miriam A. Golden. 2017. *Corruption: What Everyone Needs to Know*. Oxford: Oxford University Press.

Fukuyama, Francis. 2018. Corruption as a Political Phenomenon. In *Institutions, Governance, and the Control of Corruption*, ed. Kaushik Basu and Tito Cordella. Cham, Switzerland: Palgrave Macmillan chapter 3, pp. 51–74.

Fukuyama, Francis and Francesca Recanatini. 2021. Corruption, Elites, and Power: An Overview of International Policy Efforts to Improve the Quality of Government. In *Oxford Handbook of the Quality of Government*, ed. Andreas Bågenholm, Monika Bauhr, Marcia Grimes and Bo Rothstein. Oxford, UK: Oxford University Press chapter 22, pp. 471–494.

Gächter, Simon and Jonathan F. Schulz. 2016. "Intrinsic Honesty and the Prevalence of Rule Violations across Societies." *Nature* 531(7595):496–499.

Gingerich, Daniel W. 2013. "Governance Indicators and the Level of Analysis Problem: Empirical Findings from South America." *British Journal of Political Science* 43(July 2013):505–540.

Glaeser, Edward L. and Claudia Goldin, eds. 2007. *Corruption and Reform: Lessons from America's Economic History*. Chicago: University of Chicago Press.

Gneezy, Uri, Silvia Saccardo and Roel van Veldhuizen. 2019. "Bribery: Behavioral Drivers of Distorted Decisions." *Journal of the European Economic Association* 17(3):917–946.

Grindle, Merilee S. 2012. *Jobs for the Boys: Patronage and the State in Comparative Perspective*. Cambridge, Massachusetts: Harvard University Press.

Harsanyi, John C. and Reinhart Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, Massachusetts: MIT Press.

Humphreys, Macartan. 2017. *Political Games: Mathematical Insights on Fighting, Voting, Lying and Other Affairs of State*. New York: W.W. Norton and Company.

Justesen, Mogens K. and Christian Bjørnskov. 2014. "Exploiting the Poor: Bureaucratic Corruption and Poverty in Africa." *World Development* 58:106–115.

Khan, Mushtaq H. 2018. "Introduction: Political Settlements and the Analysis of Institutions." *African Affairs* 117(469):636–655.

Kingston, Christopher. 2008. "Social Structure and Cultures of Corruption." *Journal of Economic Behavior and Organization* 67(1):90–102.

Klitgaard, Robert. 1988. *Controlling Corruption*. Berkeley, California: University of California Press.

Kuran, Timur. 1989. "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice* 61(1):41–74.

Kuran, Timur. 1991. "The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(01):7–48.

Lagunes, Paul. 2021. *The Eye and Whip: Corruption Control in the Americas*. New York: Oxford University Press.

Lee, Wang-Sheng and Cahit Guven. 2013. "Engaging in Corruption: The Influence of Cultural Values and Contagion Effects at the Microlevel." *Journal of Economic Psychology* 39:287–300.

Levy, Brian. 2014. *Working with the Grain: Integrating Governance and Growth in Development Strategies*. New York: Oxford University Press.

Lizzeri, Alessandro and Nicola Persico. 2004. "Why Did the Elites Extend the Suffrage? Democracy and the Scope of Government, with an Application to Britain's Age of Reform." *Quarterly Journal of Economics* 119(2):707–765.

Lührmann, Anna, Kyle L. Marquardt and Valeriya Mechkova. 2020. "Constraining Governments: New Indices of Vertical, Horizontal, and Diagonal Accountability." *American Political Science Review* 114(3):811–820.

Marquette, Heather and Caryn Peiffer. 2018. "Grappling with the Real Politics of Systemic Corruption: Theoretical Debates Versus Real-World Functions." *Governance* 31(3):499–514.

Marquette, Heather and Caryn Peiffer. 2019. "Thinking Politically about Corruption as Problem-Solving: A Reply to Persson, Rothstein, and Teorell." *Governance* 32(4):811–820.

Masoud, Tarek. 2018. "Why Tunisia?" *Journal of Democracy* 29(4):166–175.

McCarthy, Colm. 2003. Corruption in Public Office in Ireland: Policy Design as a Countermeasure. In *Tribunals of Inquiry*. Dublin: pp. 1–15.

Medina, Luis Fernando. 2007. *A Unified Theory of Collective Action and Social Change*. Ann Arbor, Michigan: University of Michigan Press.

Medina, Luis Fernando. 2018. *Beyond the Turnout Paradox: The Political Economy of Electoral Participation*. Cham, Switzerland: Springer.

Mungiu-Pippidi, Alina. 2015. *The Quest for Good Governance: How Societies Develop Control of Corruption*. New York: Cambridge University Press.

Mungiu-Pippidi, Alina. 2016. "Learning from Virtuous Circles." *Journal of Democracy* 27(1):95–109.

Mungiu-Pippidi, Alina and Michael Johnston, eds. 2017. *Transitions to Good Governance: Creating Virtuous Circles of Anti-corruption*. Cheltenham, United Kingdom: Elgar.

Naím, Moisés. 1995. "The Corruption Eruption." *Brown Journal of World Affairs* 2(2):245–262.

Nyblade, Benjamin and Steven R Reed. 2008. "Who Cheats? Who Loots? Political Competition and Corruption in Japan, 1947-1993." *American Journal of Political Science* 52(4):926–941.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(21):200–249.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address." *American Political Science Review* 92(1):1–22.

Peiffer, Caryn and Linda Alvarez. 2016. "Who Will Be the Principled-Principals? Perceptions of Corruption and Willingness to Engage in Anticorruption Activism." *Governance* 29(3):351–369.

Peiffer, Caryn and Richard Rose. 2018. "Why Are the Poor More Vulnerable to Bribery in Africa? The Institutional Effects of Services." *Journal of Development Studies* 54(1):18–29.

Persson, Anna, Bo Rothstein and Jan Teorell. 2013. "Why Anticorruption Reforms Fail–Systemic Corruption as a Collective Action Problem." *Governance* 26(3):449–471.

Persson, Anna, Bo Rothstein and Jan Teorell. 2019. "Getting the Basic Nature of Systemic Corruption Right: A Reply to Marquette and Peiffer." *Governance* 32(4):799–810.

Polus, Andrzej, Dominik Kopinski and Wojciech Tycholiz. 2015. "Ready or Not: Namibia as a Potentially Successful Oil Producer." *Africa Spectrum* 50(2):31–55.

Robinson, Amanda Lea and Brigitte Seim. 2018. "Who is Targeted in Corruption? Disentangling the Effects of Wealth and Power on Exposure to Bribery." *Quarterly Journal of Political Science* 13:313–331.

Rodrik, Dani. 2007. *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth*. Princeton, NJ: Princeton University Press.

Rose-Ackerman, Susan and Bonnie Palifka. 2016. *Corruption and Government: Causes, Consequences, and Reform*. Second ed. New York: Cambridge University Press.

Rothstein, Bo. 2011*a*. "Anti-Corruption: The Indirect Big Bang Approach." *Review of International Political Economy* 18(2):228–250.

Rothstein, Bo. 2011*b*. *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective*. Chicago: University of Chicago Press.

Rothstein, Bo. 2021. *Controlling Corruption: The Social Contract Approach*. Oxford: Oxford University Press.

Rothstein, Bo and Jan Teorell. 2015. "Getting to Sweden, Part II: Breaking with Corruption in the Nineteenth Century." *Scandinavian Political Studies* 38(3):238–254.

Sandler, Todd. 2015. "Collective Action: Fifty Years Later." *Public Choice* 164(3-4):195–216.

Schelling, Thomas. 1978. *Micromotives and Macrobehavior*. New York: W.W. Norton and Company.

Shalvi, Shaul. 2016. "Corruption Corrupts." *Nature* 531(7595):456–457.

Søreide, Tina. 2014. *Drivers of Corruption: A Brief Review*. Washington, DC: World Bank.

Soto, Raimundo and Ilham Haouas. 2016. Has the UAE Escaped the Oil Curse? In *Understanding and Avoiding the Oil Curse in Resource-Rich Arab Economies*, ed. Ibrahim Elbadawi and Hoda Selim. Cambridge: Cambridge University Press chapter 12, pp. 373–420.

Stephenson, Matthew C. 2020. "Corruption as a Self-Reinforcing Trap: Implications for Reform Strategy." *World Bank Research Observer* 35(2):192–226.

Sundström, Aksel. 2019. "Why Do People Pay Bribes? A Survey Experiment with Resource Users." *Social Science Quarterly* 100(3):725–735.

Taylor, Matthew M. 2018. "Getting to Accountability: A Framework for Planning and Implementing Anticorruption Strategies." *Daedalus* 147(3):63–82.

Teorell, Jan and Bo Rothstein. 2015. "Getting to Sweden, Part I: War and Malfeasance, 1720-1850." *Scandinavian Political Studies* 38(3):217–237.

Therialt, Sean M. 2003. "Patronage, the Pendleton Act, and Power of the People." *Journal of Politics* 65(1):485–486.

Transparency International. 2009. The Anti-Corruption Plain Language Guide. Technical report Transparency International Berlin: .

Treisman, Daniel. 2000. "The Causes of Corruption: A Cross-National Study." *Journal of Public Economics* 76(3):399–457.

Ugur, Mehmet and Nandini Dasgupta. 2011. Evidence on the Economic Growth Impacts of Corruption in Low-Income Countries and Beyond. Technical Report August EPPI-Centre, Social Science Research Unit, Institute of Education, University of London London: .

United Nations. 2018. "Global Cost of Corruption at Least 5 Per Cent of World Gross Domestic Product, Secretary-General Tells Security Council, Citing World Economic Forum Data.".
URL: *https://www.un.org/press/en/2018/sc13493.doc.htm*

Vaubel, Roland. 2006. "Principal-Agent Problems in International Organizations." *Review of International Organizations* 1(2):125–138.

Weimann, Joachim, Jeannette Brosig-Koch, Timo Heinrich, Heike Hennig-Schmidt and Claudia Keser. 2019. "Public Good Provision by Large Groups – The Logic of Collective Action Revisited." *European Economic Review* 118:348–363.

Weyland, Kurt. 2012. "The Arab Spring: Why the Surprising Similarities with the Revolutionary Wave of 1848?" *Perspectives on Politics* 10(4):917–934.

Wilson, Bruce and Evelyn Villarreal. 2017. Costa Rica: Tipping Points and An Incomplete Journey. In *Transitions to Good Governance: Creating Virtuous Circles of Anti-corruption*, ed. Alina Mungiu-Pippidi and Michael Johnston. Cheltenham, United Kingdom: Elgar chapter 8, pp. 184–212.

Wrong, Michaela. 2009. *It's Our Turn to Eat: A Story of a Kenyan Whistleblower*. London: Fourth Estate.

Yap, O. Fiona. 2013. "When do Citizens Demand Punishment of Corruption?" *Australian Journal of Political Science* 48(1):57–70.

Appendix A Illustrative Pathways to Lower-Corruption Outcomes

Table A.1: Country-Level Paths to Lower-Corruption Outcomes

Country	**Critical Period(s)**	**How the Country Shifted to a Lower-Corruption Path**	**Maintained?**
Denmark	1658-1665, 1814, 1849	Loss in wars against Sweden; top-down reform initiated by kings; drafting of a new constitution following demonstrations	Yes
Sweden	1810-1850	Losing the 1808-1809 war against Russia, followed by a series of reforms	Yes
Great Britain	1780-1883	Civil service reform; legislation; a secret ballot; suffrage reform, resulting in the decline of clientelism and more funds for public services	Yes
France	1791-1975	The French Revolution; gradual decline of patronage appointments; construction of impartial institutions	Yes
Germany	1919, 1945	End of World War I and World War II.	Yes
Spain	1975-1978	End of the dictatorship of Francisco Franco.	Yes
Portugal	1974	End of the Estado Novo dictatorship, fueled by the Carnation Revolution	Yes
Ireland	1995-2007, 2012	Privatization, market reforms, corruption law reforms introduced (2012), the rise of the Celtic Tiger ended up increasing corruption, Council of Europe's Greco initiative (anti-corruption program)	Yes
Italy	1992-1996	The Clean Hands scandal, prompted by the arrest of one well-connected individual, who provided information that led to the arrest of hundreds and changed the party system	No
Estonia	1990-1995	Tax reform; e-governance; procurement reform; privatization	Yes

Continued on next page

Table A.1: Country-Level Paths to Lower-Corruption Outcomes – *continued*

Country	**Critical Period(s)**	**How the Country Shifted to a Lower-Corruption Path**	**Maintained?**
Georgia	2004-2008	The Georgian Transition, followed by a "big bang" approach from President Mikheil Saakashvili (i.e., large-scale dismissal of civil servants, televised arrests, and e-governance)	Mixed, with creeping authoritarianism and human rights issues.
Tunisia	2011-2014	Citizen demonstrations over autocratic rule fueled the Arab Spring	Mostly, though some patronage remains a challenge
Botswana	1966-present	Excellent natural resource management; protection of property rights; transparent policy-making; management of potential ethnic tensions. Strong and independent judiciary, and an exceptional rule of law.	Regular scandals imperil progress. The discovery of rich resource deposits has also led to an increase in corruption.
United States	1870-1920	The regulation of patronage appointments through the Pendleton Act; the press; the Progressivist movement; successful prosecutions.	Yes, though the role of money in politics is significant
Hong Kong	1974-1977	Egregious malfeasance by the head of police, which prompted the creation of an independent anti-corruption agency and many subsequent arrests	Yes
Taiwan	1992-	Civil service reform; high-level corruption initiatives; legislation; party system change	Yes
Singapore	1959-1990	Authoritarian leader Lee Kuan Yew pushed through a series of reforms	Yes
South Korea	1961-2003	Education; import-substitution industrialization that fueled economic growth; market reforms; legislation; protests	Yes
Japan	1945-1993,	Loss of World War 2; MacArthur Plan; resolution of a series of corruption scandals involving kickbacks and vote-buying.	Yes

Continued on next page

Table A.1: Country-Level Paths to Lower-Corruption Outcomes – *continued*

Country	**Critical Period(s)**	**How the Country Shifted to a Lower-Corruption Path**	**Maintained?**
Chile	1984-1990	Economic liberalization; privatization; loss of natural resource rents; democratic and authoritarian legacies from previous periods	Yes
Uruguay	1984	Fiscal/tax system consolidation; privatization; a democratic history; an educated and active citizenry; loss of patronage funds.	Yes
United Arab Emirates	1989-2016	E-Governance; institutional development; avoided the misuse of oil rents by investing in physical capital and institutional fabric, establishment of the Prosecution of Public Funds Office to prosecute corruption, adopting a federal penal code anchored on UN Convention Against Corruption (UNCAC)	Yes
Namibia	2000-2010	Effective legislation (Anti-Corruption Act, 2003, under the ambit of the Constitution, Electoral Act, 2014) and enforcement bodies (e.g. the Anti-Corruption Commission)	Yes
Cape Verde	1990-present	A wave of democratization, introduction of parties, e-governance	Yes
Costa Rica	1948-present	civil war of 1948, followed by new regime and electoral institute to deter election fraud; improvement in professionalism of judicial branch staff; audit and constitutional improvements following scandals; signing of OECD anti-bribery measures	Mostly

Sources: McCarthy (2003), Therialt (2003), Lizzeri and Persico (2004), Glaeser and Goldin (2007), Baker (2009), Rothstein (2011*b*), Acemoglu and Robinson (2012), Grindle (2012), Weyland (2012), Camp, Dixit and Stokes (2014), Al Athmay (2015), Arroyo Chacón (2015), Mungiu-Pippidi (2015, 2016), Polus, Kopinski and Tycholiz (2015), Rothstein and Teorell (2015), Soto and Haouas (2016), Teorell and Rothstein (2015), Soto and Haouas (2016), Fisman and Golden (2017), Fukuyama (2018), Masoud (2018), Nyblade and Reed (2008), Coppedge et al. (2018), Wilson and Villarreal (2017)

Appendix B Mathematical Proofs for Theoretical Model

B.1 Basic Collective Action Game: Version Γ^A

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/Take Action*	$t-r$, $t-r$	$s-r$, $g-b+s$
	Defect/Do Nothing	$g-b+s$, $s-r$	$g-b$, $g-b$

Where we define:

$$\begin{aligned} W_1 &= t-r \\ W_2 &= g-b+s \\ W_3 &= s-r \\ W_4 &= g-b; \end{aligned} \tag{B.1}$$

We further define all parameters to be non-negative:

$$\{t, r, g, b, s\} \in (0, \infty) : \tag{B.2}$$

And, in order to maintain the characteristics of an assurance game, as well as meet the

assumptions of collective action games (Medina, 2007, 53)[19], we assume:

$$\begin{aligned} W_1 > W_4 > W_3, \\ W_2 > W_4 > W_3, \\ W_1 > W_2 \end{aligned} \tag{B.3}$$

B.2 Pure Strategy Nash Equilibria for Version Γ^A

$\oplus$ (C, C) is a PSNE if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} u_i(C_i|C_j) &> u_i(D_i|C_j) \\ &= t - r > g - b + s, \\ &= t > s + r \text{ (centering game at zero (g-b=0))} \end{aligned} \tag{B.4}$$

$\oplus$ (D, D) is a PSNE, if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} u_i(D_i|D_j) &> u_i(C_i|D_j) \\ &= g - b > s - r, \\ &= 0 > s - r \text{ (centering game at zero (g-b=0))} \\ &= r > s \end{aligned} \tag{B.5}$$

B.3 Mixed Strategy Nash Equilibrium for Version Γ^A

Player i chooses to play C [i.e. report the bureaucrat/refuse to pay the bribe] with probability α_i, such that player j $(i \neq j)$ is indifferent between playing C [reporting/refusing to pay the bribe] and playing D [paying the bribe]:

[19] Medina (2007) points out that for there to be any hope for cooperation, at least some members of society will need prefer mutual cooperation over unilateral defection (i.e., $W_{1i} > W_{2i}$). To indicate this *individual* preference, he uses the subscript i. Since our focus here lies on the feasibility of generating greater propensity to cooperate, rather than teasing out thresholds for sufficient cooperation, we have dropped the subscript for ease of exposition.

$$\begin{aligned} EU_j(C) &= EU_j(D) \\ \alpha_i(W_1) + (1-\alpha_i)(W_3) &= \alpha_i(W_2) + (1-\alpha_i)(W_4) \\ \alpha_i(t-r) + (1-\alpha_i)(s-r) &= \alpha_i(g-b+s) + (1-\alpha_i)(g-b) \\ \alpha_i(t-s) + (s-r) &= \alpha_i(s) + (g-b) \\ \alpha_i^{\star,A} &= \frac{g-b+r-s}{t-2s} \end{aligned} \tag{B.6}$$

Which is a valid mixing probability under the previously made assumptions when players are fully rational. Specifically, centering the game at zero again, $\alpha_i^{\star,A} = \frac{r-s}{t-2s}$, where $r-s>0$ as $r>s$ by assumption (see above); and $t-2s>0$, since $t>s+r$, and $r>s$.

In such a case, their best response is to choose α_i as follows:

$$\alpha_i^*(\alpha_j) = \begin{cases} 1 \text{ if } \alpha_j > \frac{g-b+r-s}{t-2s} \\ [0,1] \text{ if } \alpha_j = \frac{g-b+r-s}{t-2s} \\ 0 \text{ if } \alpha_j < \frac{g-b+r-s}{t-2s} \end{cases} \tag{B.7}$$

B.4 Partially Augmented Collective Action Game Version Γ^B

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/Take Action*	$t-r+c$, $t-r+c$	$s-r+c$, $g-b+s+c$
	Defect/Do Nothing	$g-b+s+c$, $s-r+c$	$g-b$, $g-b$

Where:

$$\begin{aligned} W_1 &= t - r + c \\ W_2 &= g - b + s + c \\ W_3 &= s - r + c \\ W_4 &= g - b; \end{aligned} \tag{B.8}$$

Again, we further define all parameters to be non-negative:

$$\{t, r, g, b, s, c\} \in (0, \infty) : \tag{B.9}$$

And maintain the assumptions made previously so as to sustain the assurance game, and collective action game, requirements (Eq. B.3).

B.5 Pure Strategy Nash Equilibria for Version Γ^B

$\oplus$ (C, C) is a PSNE if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(C_i|C_j) > u_i(D_i|C_j) \\ & = t - r + c > g - b + s + c, \\ & \quad = t > s + r \text{ (centering game at zero (g-b=0))} \end{aligned} \tag{B.10}$$

$\oplus$ (D, D) is a PSNE, if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(D_i|D_j) > u_i(C_i|D_j) \\ & = g - b > s - r + c, \\ & \quad = 0 > s - r + c \text{ (centering game at zero (g-b=0))} \\ & \quad = r > s + c \end{aligned} \tag{B.11}$$

B.6 Mixed Strategy Nash Equilibrium for Version Γ^B

Player i chooses to play C [i.e. report the bureaucrat/refuse to pay the bribe] with probability α_i, such that player j $(i \neq j)$ is indifferent between playing C [reporting/refusing to pay the bribe] and playing D [paying the bribe]:

$$\begin{aligned} EU_j(C) &= EU_j(D) \\ \alpha_i(W_1) + (1-\alpha_i)(W_3) &= \alpha_i(W_2) + (1-\alpha_i)(W_4) \\ \alpha_i(t-r+c) + (1-\alpha_i)(s-r+c) &= \alpha_i(g-b+s+c) + (1-\alpha_i)(g-b) \\ \alpha_i(t-s) + (s-r+c) &= \alpha_i(s+c) + (g-b) \\ \alpha_i^{\star,B} &= \frac{g-b+r-s-c}{t-2s-c} \end{aligned} \tag{B.12}$$

Which is a valid mixing probability under the previously made assumptions as long as $t-2s-c > 0$. Specifically, centering the game at zero again, $\alpha_i^{\star,B} = \frac{r-s-c}{t-2s-c}$, where $r-s-c > 0$ and $r > s+c$ by assumption (see above).

Now, their best response under full common-knowledge is to choose α_i as follows:

$$\alpha_i^*(\alpha_j) = \begin{cases} 1 \text{ if } \alpha_j > \frac{g-b+r-s-c}{t-2s-c} \\ [0,1] \text{ if } \alpha_j = \frac{g-b+r-s-c}{t-2s-c} \\ 0 \text{ if } \alpha_j < \frac{g-b+r-s-c}{t-2s-c} \end{cases} \tag{B.13}$$

B.7 Partially Augmented Collective Action Game Version Γ^C

		Citizen 2	
		Cooperate/ Take Action	*Defect/ Do Nothing*
Citizen 1	*Cooperate/Take Action*	$t-r+p$, $t-r+p$	$s-r+p$, $g-b+s$
	Defect/Do Nothing	$g-b+s$, $s-r+p$	$g-b$, $g-b$

Where:

$$\begin{aligned} W_1 &= t-r+p \\ W_2 &= g-b+s \\ W_3 &= s-r+p \\ W_4 &= g-b; \end{aligned} \tag{B.14}$$

Still, we further define all parameters to be non-negative:

$$\{t, r, g, b, s, p\} \in (0, \infty) : \tag{B.15}$$

And maintain the assumptions made previously (Eq. B.3) keeping with the necessary collective action and assurance game assumptions.

B.8 Pure Strategy Nash Equilibria for Version Γ^C

$\oplus$ (C, C) is a PSNE if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(C_i|C_j) > u_i(D_i|C_j) \\ & = t - r + p > g - b + s, \\ & \quad = t > s + r - p \text{ (centering game at zero (g-b=0))} \end{aligned} \tag{B.16}$$

$\oplus$ (D, D) is a PSNE, if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(D_i|D_j) > u_i(C_i|D_j) \\ & = g - b > s - r + p, \\ & \quad = 0 > s - r + p \text{ (centering game at zero (g-b=0))} \\ & \quad = r > s + p \end{aligned} \tag{B.17}$$

B.9 Mixed Strategy Nash Equilibrium for Version Γ^C

Player i chooses to play C [i.e. report the bureaucrat/refuse to pay the bribe] with probability α_i, such that player j $(i \neq j)$ is indifferent between playing C [reporting/refusing to pay the bribe] and playing D [paying the bribe]:

$$\begin{aligned} EU_j(C) &= EU_j(D) \\ \alpha_i(W_1) + (1 - \alpha_i)(W_3) &= \alpha_i(W_2) + (1 - \alpha_i)(W_4) \\ \alpha_i(t - r + p) + (1 - \alpha_i)(s - r + p) &= \alpha_i(g - b + s) + (1 - \alpha_i)(g - b) \\ \alpha_i(t - s) + (s - r + p) &= \alpha_i(s) + (g - b) \\ \alpha_i^{\star,C} &= \frac{g - b + r - s - p}{t - 2s} \end{aligned} \tag{B.18}$$

Which is a valid mixing probability under the previously made assumptions when players are fully rational. Specifically, centering the game at zero again, $\alpha_i^{\star,C} = \frac{r-s-p}{t-2s}$, where

$r - s - p > 0$ as $r > s + p$ by assumption (see above); and $t - 2s > 0$, since $t > s + r$, and $r > s$.

Now, their best response under full common-knowledge rationality is to choose α_i as follows:

$$\alpha_i^*(\alpha_j) = \begin{cases} 1 \text{ if } \alpha_j > \frac{g-b+r-s-p}{t-2s} \\ [0,1] \text{ if } \alpha_j = \frac{g-b+r-s-p}{t-2s} \\ 0 \text{ if } \alpha_j < \frac{g-b+r-s-p}{t-2s} \end{cases} \tag{B.19}$$

B.10 Fully Augmented Collective Action Game Version Γ^D

		Citizen 2	
		Cooperate/ Take Action	*Defect/ Do Nothing*
Citizen 1	*Cooperate/Take Action*	$t - r + p + c$, $t - r + p + c$	$s - r + p + c$, $g - b + s + c$
	Defect/Do Nothing	$g - b + s + c$, $s - r + p + c$	$g - b$, $g - b$

Where:

$$\begin{aligned} W_1 &= t - r + p + c \\ W_2 &= g - b + s + c \\ W_3 &= s - r + p + c \\ W_4 &= g - b \end{aligned} \tag{B.20}$$

Here, too, we further define all parameters to be non-negative:

$$\{t, r, g, b, s, p, c\} \in (0, \infty) : \tag{B.21}$$

And maintain the assumptions made previously (Eq. B.3).

B.11 Pure Strategy Nash Equilibria for Version Γ^D

$\oplus$ (C, C) is a PSNE, because for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(C_i|C_j) > u_i(D_i|C_j) \\ &= t - r + p + c > g - b + s + c \\ &= t - r + p + c > s + c \text{ (applying centering at g-b = 0)} \\ &\quad = t > s + r - p \text{ (True by base game's assumption, as } t > s + r) \end{aligned} \tag{B.22}$$

$\oplus$ (D, D) is a PSNE, if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(D_i|D_j) > u_i(C_i|D_j) \\ &= g - b > s - r + p + c \\ &\quad = 0 > s - r + p + c \text{ (applying centering at g-b = 0)} \\ &= r - p - c > s \end{aligned} \tag{B.23}$$

B.12 Mixed Strategy Nash Equilibrium for Version Γ^D

Again, player i chooses to play C with probability α_i such that player j is indifferent between playing C and D:

$$\begin{aligned} EU_j(C) &= EU_j(D) \\ \alpha_i(W_1) + (1 - \alpha_i)(W_3) &= \alpha_i(W_2) + (1 - \alpha_i)(W_4) \\ \alpha_i(t - r + p + c) + (1 - \alpha_i)(s - r + p + c) &= \alpha_i(g - b + s + c) + (1 - \alpha_i)(g - b) \\ \alpha_i(t - s) + (s - r + p + c) &= \alpha_i(s + c) + g - b \\ \alpha_i(t - 2s - c) &= g - b + r - s - p - c \\ \alpha_i^{\star,D} &= \frac{g - b + r - s - p - c}{t - 2s - c} \end{aligned} \tag{B.24}$$

Which is a valid mixing probability as long as $t-2s-c > 0$, and (as previously assumed) $r > s + p + c$. Again, under full common-knowledge rationality player i's best response is:

$$\alpha_i^*(\alpha_j) = \begin{cases} 1 \text{ if } \alpha_j > \frac{g-b+r-s-p-c}{t-2s-c} \\ [0,1] \text{ if } \alpha_j = \frac{g-b+r-s-p-c}{t-2s-c} \\ 0 \text{ if } \alpha_j < \frac{g-b+r-s-p-c}{t-2s-c} \end{cases} \tag{B.25}$$

B.13 Relaxing the Assumption of Common-Knowledge Rationality

Generally, if players are not *fully* rational, player i's utility functions can be re-written as a weighted average of her payoffs under rationality, and her payoffs conditional upon her beliefs about the likelihood of cooperation by the other player (j). Further, let $\lambda \in [0,1]$ denote the degree to which actors are fully rational:

$$\begin{aligned} u_i^{\lambda}(\alpha_i, \alpha_j, \beta_j) = \lambda[&\alpha_i(\alpha_j(W_1) + (1-\alpha_j)(W_3)) \\ &+ (1-\alpha_i)(\alpha_j(W_2) + (1-\alpha_j)(W_4))] \\ &+ (1-\lambda)[\alpha_i(\beta_j(W_1) + (1-\beta_j)(W_3)) \\ &+ (1-\alpha_i)(\beta_j(W_2) + (1-\beta_j)(W_4))] \end{aligned} \tag{B.26}$$

First, note that when $\lambda = 1$, Equation (B.26) simplifies to the optimal mixing probabilities for each player noted in the MSNE of a given game. Meanwhile, let $\lambda = 0$ so as to generate player i's utility function when strategies entirely depend on beliefs about the other player's behavior rather than rationality:

$$\begin{aligned} u_i^{\lambda=0}(\alpha_i, \beta_j) &= \alpha_i(\beta_j(W_1) + (1-\beta_j)(W_3)) + (1-\alpha_i)(\beta_j(W_2) + (1-\beta_j)(W_4)) \\ &= \alpha_i\beta_j W_1 - \alpha_i\beta_j W_3 - \alpha_i\beta_j W_2 + \alpha_i\beta_j W_4 + \alpha_i W_3 - \alpha_i W_4 + \beta_j W_2 - \beta_j W_4 + W_4 \end{aligned} \tag{B.27}$$

B.14 Private vs. Collective Benefits

In order to assess whether the provision of private or collective benefits is more likely to reduce the likelihood of corruption, we must ascertain the conditions under which the Mixed Strategy Nash Equilibria of Equations (4) and (5) exceed one another:[20]

$$\begin{aligned} \alpha_i^{\star\prime} &> \alpha_i^{\star\prime\prime} \\ \frac{r-s-c}{t-2s-c} &> \frac{r-s-p}{t-2s} \\ \frac{(r-s-c)(t-2s)}{(t-2s-c)(t-2s)} &> \frac{(r-s-p)(t-2s-c)}{(t-2s)(t-2s-c)} \\ \frac{rt-st-ct-2rs+2s^2+2cs}{(t-2s-c)(t-2s)} &> \frac{rt-2rs-rc-st+2s^2+cs-pt-2ps+pc}{(t-2s-c)(t-2s)} \end{aligned} \tag{B.28}$$

Upon simplifying and dropping the denominator—which is always positive, as assumed above so as to maintain the assurance game format of the game—we obtain:

$$\begin{aligned} cs-ct-cr-pc+pt-2ps &> 0 \\ (t-2s-c)*p &> (t-s-r)*c \\ p &> \frac{t-s-r}{t-2s-c}*c \end{aligned} \tag{B.29}$$

What we can discern from Equation (B.29) is that when private benefits (p) are sufficiently large relative to collective benefits (c), providing private benefits is more likely to be *effective* in reducing corruption than collective benefits. Further, we can show that the multiplier on c, $\frac{t-s-r}{t-2s-c}$, is always smaller than unity:

$$\begin{aligned} t-s-r &< t-2s-c \\ s+c &< r \end{aligned} \tag{B.30}$$

[20] Note that the greater the reduction in the mixing probability, the more likely a policy is to reduce corruption as we relax the rationality assumption. See proofs in Appendices B.6 and B.9.

This inequality follows from Equation B.11 and always holds due to the assumptions necessary to uphold the basic form of the augmented assurance game, Version Γ^B.

In summary, private benefits to individuals, such as whistleblower protections, can diminish the likelihood that citizens will acquiesce to bribe requests by corrupt bureaucrats. Similarly, collective benefits, such as increased bureaucrat monitoring, can also reduce this likelihood. When citizens value collective and private benefits equally, providing private benefits is relatively more likely to be effective than providing collective benefits. Collective and private benefits, however, are not substitutes. Even as the basic form of the assurance game is maintained, the effects of both collective and private benefits are additive in combating corruption.

B.15 Calculating Stability Sets: A Numerical Example

Using Equation (B.26), we can now compute the most likely (single) equilibria for each of the versions of our games we have specified above (Γ^A through Γ^D) and for different initial belief values.

First, we must choose values for all the parameters in the game, consistent with the assumptions we made (i.e. $W_1 > W_4 > W_3, W_2 > W_4 > W_3$, and $W_1 > W_2$), so as to identify the numerical equilibria for each game when $\lambda = 1$. Furthermore, we center the game at zero by letting $g = b$ (see Section 3).

Since $W_4 > W_3$ and $W_4 \equiv g - b = 0$, $W_3 \equiv s - r < 0$. Let $s = 1$, $r = 2$ to fulfil this. Given those choices, $W_2 \equiv g - b + s = 1$, and $W_2 > W_4 > W_3$ holds. Further, let $t = 4$, such that $W_1 \equiv t - r = 2$, and $W_1 > W_4 > W_3$ as well as $W_1 > W_2$ is met. Finally, let us define the additional payoffs from monitoring (p, c) such that $W_3 < W_4$ is maintained. For ease of exposition, let $p = c = 0.25$ and to meet these conditions.

Given these parameter values, we can calculate the payoffs for each version of the game as follows:

- Γ^A: $W_1 = 2, W_2 = 1, W_3 = -1, W_4 = 0$.
- Γ^B: $W_1 = 2.25, W_2 = 1.25, W_3 = -0.75, W_4 = 0$.
- Γ^C: $W_1 = 2.25, W_2 = 1, W_3 = -0.75, W_4 = 0$.
- Γ^D: $W_1 = 2.5, W_2 = 1.25, W_3 = -0.5, W_4 = 0$.

Note that the choice of parameter values is, of course, somewhat arbitrary. As a result of our choice, as we will show in what follows, the mixed strategy Nash Equilibrium will be moved closer towards the origin upon providing *private* benefits rather than *collective* ones. As we have shown in Appendix B.14, this does not have to be the case, but rather is an artifact of our choices for parameter values.

Equations (B.6), (B.12), (B.18), and (B.24) allow us to calculate the corresponding mixing probabilities pertaining to each of the MSNEs for the versions of our game given the numerical payoffs above.

- $\alpha_i^{*,A} = \frac{r-s}{t-2s} = \frac{2-1}{4-2(1)} = \frac{1}{2}$
- $\alpha_i^{*,B} = \frac{r-s-c}{t-2s-c} = \frac{2-1-0.25}{4-2(1)-0.25} = \frac{0.75}{1.75} = \frac{3}{7}$
- $\alpha_i^{*,C} = \frac{r-s-p}{t-2s} = \frac{2-1-0.25}{4-2(1)} = \frac{0.75}{2} = \frac{3}{8}$
- $\alpha_i^{*,D} = \frac{r-s-p-c}{t-2s-c} = \frac{2-1-0.25-0.25}{4-2(1)-0.25} = \frac{0.5}{1.75} = \frac{2}{7}$

Having identified these mixing probabilities, we must now choose an array of beliefs so as to see how optimal behavior varies when the assumption of $\lambda = 1$ is relaxed and beliefs about others' likelihood of cooperation begin to determine the outcome. To do this, we must choose values for these beliefs (β_j) so as to fall between the mixing probabilities identified above—in addition to values above (and below) the maximal (minimal) mixing probability, respectively—in order to demonstrate varying optimal behavior upon introducing monitoring parameters.

Recall Equation (B.27), and let $\beta_j = \frac{1}{4}$, such that $\beta_j < \alpha_i^{*,D} < \alpha_i^{*,C} < \alpha_i^{*,B} < \alpha_i^{*,A}$:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{1}{4}) = \frac{\alpha_i W_1}{4} - \frac{\alpha_i W_3}{4} - \frac{\alpha_i W_2}{4} + \frac{\alpha_i W_4}{4} + \alpha_i W_3 - \alpha_i W_4 + \frac{W_2}{4} - \frac{W_4}{4} + W_4 \tag{B.31}$$

Now apply payoffs from version Γ^A to determine the optimal strategy when $\lambda = 0$:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^A}(\alpha_i, \beta_j = \frac{1}{4}) &= \frac{2\alpha_i}{4} + \frac{\alpha_i}{4} - \frac{\alpha_i}{4} - \alpha_i + \frac{1}{4} \\ &= \frac{1 - 2\alpha_i}{4} \end{aligned} \tag{B.32}$$

Which is maximized when $\alpha_i = 0$. Symmetrical initial beliefs of $\beta = \frac{1}{4}$ thus correspond to an optimal strategy of always defecting, (D, D) or W_4, making it the only equilibrium.

Applying the payoffs from version Γ^B to Equation (B.31), we obtain:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^B}(\alpha_i, \beta_j = \frac{1}{4}) &= \frac{2.25\alpha_i}{4} + \frac{0.75\alpha_i}{4} - \frac{1.25\alpha_i}{4} - 0.75\alpha_i + \frac{1.25}{4} \\ &= \frac{1.25 - 1.25\alpha_i}{4} \end{aligned} \tag{B.33}$$

Which is also maximized when $\alpha_i = 0$, making (D, D) or W_4 the only equilibrium.

Continuing with version Γ^C, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^C}(\alpha_i, \beta_j = \frac{1}{4}) &= \frac{2.25\alpha_i}{4} + \frac{0.75\alpha_i}{4} - \frac{\alpha_i}{4} - 0.75\alpha_i + \frac{1}{4} \\ &= \frac{1 - \alpha_i}{4} \end{aligned} \tag{B.34}$$

Which again is maximized when $\alpha_i = 0$.

Turning to the fully augmented version of our game, Γ^D, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^D}(\alpha_i, \beta_j = \frac{1}{4}) &= \frac{2.5\alpha_i}{4} + \frac{0.5\alpha_i}{4} - \frac{1.25\alpha_i}{4} - 0.5\alpha_i + \frac{1.25}{4} \\ &= \frac{1.25 - 0.25\alpha_i}{4} \end{aligned} \tag{B.35}$$

As expected, since beliefs (β_j) are still smaller than the version's optimal mixing probabilities under the game's MSNE, this equation, too, is maximized when $\alpha_i = 0$. When the beliefs in others' cooperation are low (e.g., $\beta_j = \frac{1}{4}$), all versions point towards uniform defection by the actors when beliefs (and not rationality) inform the actors' choices.

Now, return to Equation (B.27), and let $\beta_j = \frac{1}{3}$, such that $\alpha_i^{*,D} < \beta_j < \alpha_i^{*,C} < \alpha_i^{*,B} < \alpha_i^{*,A}$:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{1}{3}) = \frac{\alpha_i W_1}{3} - \frac{\alpha_i W_3}{3} - \frac{\alpha_i W_2}{3} + \frac{\alpha_i W_4}{3} + \alpha_i W_3 - \alpha_i W_4 + \frac{W_2}{3} - \frac{W_4}{3} + W_4 \tag{B.36}$$

Again, we begin by plugging in the payoffs of version Γ^A to determine the optimal strategy when $\lambda = 0$:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^A}(\alpha_i, \beta_j = \frac{1}{3}) &= \frac{2\alpha_i}{3} + \frac{\alpha_i}{3} - \frac{\alpha_i}{3} - \alpha_i + \frac{1}{3} \\ &= \frac{1 - \alpha_i}{3} \end{aligned} \tag{B.37}$$

Which is maximized when $\alpha_i = 0$, thus yielding uniform defection as the optimal strategy for player i, and also for all players j due to payoff symmetry.

Applying the payoffs from version Γ^B to Equation (B.36), we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^B}(\alpha_i, \beta_j = \frac{1}{3}) &= \frac{2.25\alpha_i}{3} + \frac{0.75\alpha_i}{3} - \frac{1.25\alpha_i}{3} - 0.75\alpha_i + \frac{1.25}{3} \\ &= \frac{5 - 2\alpha_i}{12} \end{aligned} \tag{B.38}$$

Which is also maximized when $\alpha_i = 0$, and the logic above with respect to optimal strategy continues to apply.

Turning to version Γ^C, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^C}(\alpha_i, \beta_j = \frac{1}{3}) &= \frac{2.25\alpha_i}{3} + \frac{0.75\alpha_i}{3} - \frac{\alpha_i}{3} - 0.75\alpha_i + \frac{1}{3} \\ &= \frac{5 - \alpha_i}{12} \end{aligned} \tag{B.39}$$

Which again is maximized when $\alpha_i = 0$, yielding (D, D) or W_4 as the only equilibrium in the stability set when $\beta_j = \frac{1}{3}$.

Turning to the fully augmented version of our game, Γ^D, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^D}(\alpha_i, \beta_j = \frac{1}{3}) &= \frac{2.5\alpha_i}{3} + \frac{0.5\alpha_i}{3} - \frac{1.25\alpha_i}{3} - 0.5\alpha_i + \frac{1.25}{3} \\ &= \frac{5 + \alpha_i}{12} \end{aligned} \tag{B.40}$$

Which conversely is maximized when $\alpha_i = 1$. Consequently, an equilibrium of (C, C) or W_1 is in the stability set: when beliefs regarding the likelihood of cooperation of other actors are slightly larger (i.e. $\beta_j = \frac{1}{3}$ rather than $\beta_j = \frac{1}{4}$), we see varying equilibria in the stability sets for version Γ^D, holding all other factors constant.

Again returning to Equation (B.27), now let $\beta_j = \frac{2}{5}$, i.e., $\alpha_i^{*,D} < \alpha_i^{*,C} < \beta_j < \alpha_i^{*,B} < \alpha_i^{*,A}$:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{2}{5}) = \frac{2\alpha_i W_1}{5} - \frac{2\alpha_i W_3}{5} - \frac{2\alpha_i W_2}{5} + \frac{2\alpha_i W_4}{5} + \alpha_i W_3 - \alpha_i W_4 + \frac{2W_2}{5} - \frac{2W_4}{5} + W_4 \tag{B.41}$$

Now, apply payoffs from version Γ^A to determine the optimal strategy when $\lambda = 0$:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^A}(\alpha_i, \beta_j = \frac{2}{5}) &= \frac{4\alpha_i}{5} + \frac{2\alpha_i}{5} - \frac{2\alpha_i}{5} - \alpha_i + \frac{2}{5} \\ &= \frac{2 - \alpha_i}{5} \end{aligned} \tag{B.42}$$

Which is maximized when $\alpha_i = 0$, making (D, D) or W_4 in the stability set.

Applying the payoffs from version Γ^B to Equation (B.41), we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^B}(\alpha_i, \beta_j = \frac{2}{5}) &= \frac{4.5\alpha_i}{5} + \frac{1.5\alpha_i}{5} - \frac{2.5\alpha_i}{5} - 0.75\alpha_i + \frac{2.5}{5} \\ &= \frac{5 - 0.5\alpha_i}{10} \end{aligned} \tag{B.43}$$

Which is also still maximized when $\alpha_i = 0$.

As for version Γ^C, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^C}(\alpha_i, \beta_j = \frac{2}{5}) &= \frac{4.5\alpha_i}{5} + \frac{1.5\alpha_i}{5} - \frac{2\alpha_i}{5} - 0.75\alpha_i + \frac{2}{5} \\ &= \frac{4 + 0.5\alpha_i}{10} \end{aligned} \tag{B.44}$$

Which is now maximized when $\alpha_i = 1$, thus no longer yielding the uncooperative equilibrium in its stability set.

Similarly, for the fully augmented version of our game, Γ^D, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^D}(\alpha_i, \beta_j = \frac{2}{5}) &= \frac{5\alpha_i}{5} + \frac{\alpha_i}{5} - \frac{2.5\alpha_i}{5} - 0.5\alpha_i + \frac{2.5}{5} \\ &= \frac{2.5 + \alpha_i}{5} \end{aligned} \tag{B.45}$$

Again considering Equation (B.27), now let $\beta_j = \frac{4}{9}$, i.e., $\alpha_i^{*,D} < \alpha_i^{*,C} < \alpha_i^{*,B} < \beta_j < \alpha_i^{*,A}$:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{4}{9}) = \frac{4\alpha_i W_1}{9} - \frac{4\alpha_i W_3}{9} - \frac{4\alpha_i W_2}{9} + \frac{4\alpha_i W_4}{9} + \alpha_i W_3 - \alpha_i W_4 + \frac{4W_2}{9} - \frac{4W_4}{9} + W_4 \tag{B.46}$$

Now apply payoffs from version Γ^A to determine the optimal strategy when $\lambda = 0$:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^A}(\alpha_i, \beta_j = \frac{4}{9}) &= \frac{8\alpha_i}{9} + \frac{4\alpha_i}{9} - \frac{4\alpha_i}{9} - \alpha_i + \frac{4}{9} \\ &= \frac{4 - \alpha_i}{9} \end{aligned} \tag{B.47}$$

Which remains optimized at $\alpha_i = 0$, and thus only holds (D, D) or W_1 in the stability set.

Applying the payoffs from version Γ^B to Equation (B.46), we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^B}(\alpha_i, \beta_j = \frac{4}{9}) &= \frac{9\alpha_i}{9} + \frac{3\alpha_i}{9} - \frac{5\alpha_i}{9} - 0.75\alpha_i + \frac{5}{9} \\ &= \frac{20 + \alpha_i}{36} \end{aligned} \tag{B.48}$$

Which is now optimized at $\alpha_i = 1$, thus yielding the cooperative equilibrium in the stability set.

Continuing with version Γ^C, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^C}(\alpha_i, \beta_j = \frac{4}{9}) &= \frac{9\alpha_i}{9} + \frac{3\alpha_i}{9} - \frac{4\alpha_i}{9} - 0.75\alpha_i + \frac{4}{9} \\ &= \frac{16 + 5\alpha_i}{36} \end{aligned} \tag{B.49}$$

Which can be optimized when $\alpha_i = 1$.

Turning to the fully augmented version of our game, Γ^D, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^D}(\alpha_i, \beta_j = \frac{4}{9}) &= \frac{10\alpha_i}{9} + \frac{2\alpha_i}{9} - \frac{5\alpha_i}{9} - 0.5\alpha_i + \frac{5}{9} \\ &= \frac{10 + 5\alpha_i}{18} \end{aligned} \tag{B.50}$$

Which again yields the cooperative equilibrium (C, C) or W_1 as the sole equilibrium in the stability set.

Lastly, consider Equation (B.27), and let $\beta_j = \frac{2}{3}$, i.e., $\alpha_i^{*,D} < \alpha_i^{*,C} < \alpha_i^{*,B} < \alpha_i^{*,A} < \beta_j$:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{2}{3}) = \frac{2\alpha_i W_1}{3} - \frac{2\alpha_i W_3}{3} - \frac{2\alpha_i W_2}{3} + \frac{2\alpha_i W_4}{3} + \alpha_i W_3 - \alpha_i W_4 + \frac{2W_2}{3} - \frac{2W_4}{3} + W_4 \tag{B.51}$$

Now apply payoffs from version Γ^A to determine the optimal strategy when $\lambda = 0$:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^A}(\alpha_i, \beta_j = \frac{2}{3}) &= \frac{4\alpha_i}{3} + \frac{2\alpha_i}{3} - \frac{2\alpha_i}{3} - \alpha_i + \frac{2}{3} \\ &= \frac{2 + \alpha_i}{3} \end{aligned} \tag{B.52}$$

Which now holds the cooperative equilibrium in the stability set, as it is maximized when $\alpha_i = 1$.

Applying the payoffs from version Γ^B to Equation (B.51), we obtain:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^B}(\alpha_i, \beta_j = \frac{2}{3}) &= \frac{4.5\alpha_i}{3} + \frac{1.5\alpha_i}{3} - \frac{2.5\alpha_i}{3} - 0.75\alpha_i + \frac{2.25}{3} \\ &= \frac{9 + 5\alpha_i}{12} \end{aligned} \tag{B.53}$$

Which too is maximized when $\alpha_i = 1$.

Continuing with version Γ^C, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\ \Gamma^C}(\alpha_i, \beta_j = \frac{2}{3}) &= \frac{4.5\alpha_i}{3} + \frac{1.5\alpha_i}{3} - \frac{2\alpha_i}{3} - 0.75\alpha_i + \frac{2}{3} \\ &= \frac{8 + 7\alpha_i}{12} \end{aligned} \tag{B.54}$$

Finally, turning to the fully augmented version of our game, Γ^D, we obtain:

$$\begin{aligned} u_i^{\lambda=0,\,\Gamma^D}(\alpha_i, \beta_j = \frac{2}{3}) &= \frac{5\alpha_i}{3} + \frac{1\alpha_i}{3} - \frac{2.5\alpha_i}{3} - 0.5\alpha_i + \frac{2.5}{3} \\ &= \frac{2.5 + 2\alpha_i}{3} \end{aligned} \tag{B.55}$$

As we can see, when beliefs are sufficiently large, all versions of our games discussed here (i.e., Γ^A through Γ^D) hold the cooperative equilibrium (C, C) or W_1 in the stability set. If beliefs are a random variable, it becomes evident that versions that include monitoring or protection (i.e., Γ^B through Γ^D) have cooperative equilibria in the stability set for greater portion of the probability density function (pdf) of the beliefs. Thus, regardless of the shape of the pdf, cooperative outcomes such as reporting of corruption become more likely upon introducing monitoring (c) and/or protection (p) parameters.

Appendix C Comparative Statics

Recall from Appendix (B.12) that the augmented version of our game's (Γ^D) mixed strategy Nash Equilibrium is given by:

$$\alpha_i^* = \frac{g - b + r - s - c - p}{t - 2s - c}$$

Further recall that by virtue of the consistently applied assumption of the assurance game (or Stag Hunt) for versions Γ^A through Γ^D form:

$$t - r > g - b + s > g - b > s - r$$

and

$$t - r + c + p > g - b + s + c > g - b > s - r + p + c$$

and all of the above parameters are assumed non-negative. Against this backdrop, we can compute comparative statics.

C.1 Collective Benefits from Monitoring, *c*

$$\begin{aligned}
\frac{\partial \alpha_i^*}{\partial c} &= \frac{r - s - c - p}{(t - 2s - c)^2} - \frac{1}{t - 2s - c} \\
&= \frac{r - s - c - p}{(t - 2s - c)^2} - \frac{t - 2s - c}{(t - 2s - c)^2} \\
&= \frac{r - s - c - p - t + 2s + c}{(t - 2s - c)^2} \\
&= \frac{r - t + s - p}{(t - 2s - c)^2}
\end{aligned}$$

Which is negative, since $t - r + p > s$, as shown in Appendix B.11. Hence, an increase in collective benefits emanating from monitoring coincides with a shrinking MSNE, and thus

greater likelihood of cooperation for any given probability distribution of prior beliefs.

C.2 Private Benefits from Monitoring, p

$$\frac{\partial \alpha_i^*}{\partial p} = -\frac{1}{t - 2s - c}$$

The partial derivative of α_i^* is negative, as $t > 2s+c$, which is necessary for the existence of the MSNE (see Appendix B.12). Hence, an increase in the private benefits emanating from monitoring coincides with a shrinking MSNE, and thus greater likelihood of cooperation for any given probability distribution of prior beliefs.

C.3 Bureaucrat Retaliation, r

$$\frac{\partial \alpha_i^*}{\partial r} = \frac{1}{t - 2s - c}$$

Again, the partial derivative of α_i^* is positive, as $t > 2s + c$, which is necessary for the existence of the MSNE (see Appendix B.12). As the MSNE shifts towards the one, and given a probability distribution of prior beliefs, taking costly action against corruption becomes less likely.

As above, we assume that $t - r > g - b + s$, such that the cost from potential retaliation, r, cannot exceed the difference between joint cooperation and singular cooperation, $t - s$. If r did increase beyond that threshold, the game ceases to be an assurance game, and takes on a prisoner's dilemma-type of structure (see Table D.1 in Appendix D). As described above, such a move is theoretically and substantively illogical.

C.4 Benefits from Joint Cooperation, *t*

$$\frac{\partial \alpha_i^*}{\partial t} = -\frac{r - s - c - p}{(t - 2s - c)^2}$$

Which is negative, because $r > s + p + c$ by assumption, as shown in Appendix B.12. Consequently, increases in these benefits shift the MSNE towards zero, and under any given probability distribution of initial beliefs, coincide with taking more costly action against corruption.

C.5 Benefits from Individual Cooperation, *s*

$$\begin{aligned}
\frac{\partial \alpha_i^*}{\partial s} &= \frac{2(r - s - c - p)}{(t - 2s - c)^2} - \frac{1}{t - 2s - c} \\
&= \frac{2(r - s - c - p)}{(t - 2s - c)^2} - \frac{t - 2s - c}{(t - 2s - c)^2} \\
&= \frac{2r - 2s - 2c - 2p - t + 2s + c}{(t - 2s - c)^2} \\
&= \frac{2r - c - 2p - t}{(t - 2s - c)^2}
\end{aligned}$$

Which can be either positive if $(t > 2r - c - 2p)$ or negative otherwise. In other words, if benefits from joint cooperation (t) are sufficiently large, increases to s shift the MSNE towards one, thus rendering cooperation less likely—i.e., under a given probability distribution of initial beliefs. This might seem counterintuitive at first, but recall that these benefits accrue to free-riding actors as much as they do to individual cooperators.

Since we assume $r > s + p + c$ and $t > s + r + c$ to maintain the assurance game, as shown in Appendix B.12, s can only increase to a limited degree, *ceteris paribus*, before the game no longer resembles an assurance game.

Appendix D Basic Collective Action Games

Table D.1: Types of Collective Action Games

Prisoner's Dilemma

		Citizen 2: *Cooperate/ Take Action*	*Citizen* 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	2, 2	0, 3
	Defect/ Do Nothing	3, 0	1, 1

Assurance Game/Stag Hunt

		Citizen 2: *Cooperate/ Take Action*	*Citizen* 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	3, 3	0, 2
	Defect/ Do Nothing	2, 0	1, 1

Deadlock

		Citizen 2: *Cooperate/ Take Action*	*Citizen* 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	0, 0	1, 2
	Defect/ Do Nothing	2, 1	3, 3

Harmony

		Citizen 2: *Cooperate/ Take Action*	*Citizen* 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	3, 3	2, 1
	Defect/ Do Nothing	1, 2	0, 0

Chicken/Hawk-Dove

		Citizen 2: *Cooperate/ Take Action*	*Citizen* 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	2, 2	1, 3
	Defect/ Do Nothing	3, 1	0, 0

Note: The numeric payoffs denote the preference orderings in each game, which are not perfectly comparable across games. For more, see Dixit, Skeath and Reiley (2014) and Humphreys (2017).