Essays on Economics of Information and Incentives

Jakub Redlicki
Hertford College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Economics

Trinity Term 2017

Abstract

The first chapter addresses a common presumption in organisational design that employees should not be given discretion about performance measures when offered performance pay. The concern is that they would make a self-serving choice, for example, one that allows them to boost their apparent performance by working on tasks which they find easy but bring little benefit to the company. I investigate this problem in a model in which the principal decides whether to delegate the choice of performance measure to an agent who is privately informed about the degree of substitutability of his effort on different tasks. I show that when the principal is using menus of contracts as a screening device, allowing the agent to choose his performance measure privately—and possibly in a self-serving way—can alleviate the problem of hidden information. Consequently, delegating the choice of performance measure can be complementary to provision of incentives and may increase the principal's payoff.

The second chapter analyses the incentives of authoritarian regimes to manipulate information by adding noise to the citizens' information, which is a tactic that is increasingly common in the real world. I consider a global games model in which a regime controls the amount of noise in the citizens' private information and is overthrown if enough citizens attack it. The analysis sheds light on recent findings in political science which show that the Chinese regime uses censorship to prevent collective action rather than criticism of the state *per se*, and that it employs social media commentators to distract the citizens rather than to persuade them. I show that the better the citizens are informed about the regime, the more its incentives to add noise are driven by the criticism of the state rather than by the need to minimise the size of collective attacks. Furthermore, the incentives to add noise may become stronger if citizens' better coordination comes at the cost of impeded information aggregation.

The third chapter, which is co-authored with Bartosz Redlicki (University of Cambridge), studies a game between a biased sender (an interest group) and a decision maker (a policy maker) where the former can falsify scientific evidence at a cost. The sender observes scientific evidence and knows that it will also be observed by the decision maker unless he falsifies it. If he falsifies, then there is a chance that the decision maker observes the falsified evidence rather than the true scientific evidence. First, we investigate the decision maker's incentives to privately acquire independent evidence, which not only provides additional information to her but can also strengthen or weaken the sender's falsification effort. We characterise the circumstances under which the benefit from the additional information is boosted, unaffected, dampened, and fully offset by the adjustment in the sender's falsification strategy. Second, we analyse the decision maker's incentives to acquire information from the sender. We show that she may be better off by committing to pay less than full attention to the sender as this can discourage falsification.

Acknowledgements

First, I would like to express my immense gratitude to my supervisor, Margaret Meyer, for her guidance and support throughout my M. Phil. and D. Phil. degrees. Meg's feedback, patience, and encouragement have been invaluable over the course of my work on this thesis.

I would also like to thank my economics tutors from my undergraduate years at St Edmund Hall, Outi Aarnio and Martin Slater, for giving me the chance to come to Oxford nine years ago and for providing me with a good grounding in economics. Conversations with my peers at the Department of Economics, in particular Rustu Duran, Paweł Gola, and Claudia Herresthal, helped improve the quality of my dissertation. I also very much enjoyed and benefited from discussions on economics with fellow graduate students at the 25th Jerusalem School in Economic Theory at the Hebrew University, the 3rd Summer School of the Econometric Society at the University of Tokyo, and during my short academic visit at the Center of Mathematical Sciences and Applications at Harvard University.

I am grateful to the Economic and Social Research Council and to the Department of Economics for providing me with funding throughout my postgraduate years at Oxford. Hertford College and the George Webb-Medley Fund at the Department of Economics gave me financial support to attend conferences and workshops.

I also want to thank my friends at the Oxford University Polish Society and the Oxford University Volleyball Club, spending time with whom was a welcome distraction after the hours spent in the department: Joanna Bagniewska, Andreas Iskra, Gytis Jankevicius, Mateusz Mazzini, Stefan Nekovar, Radosław Nowak, Nicolas Stone-Villani, Mateusz Ujma, Maria Wilczek, Andrzej Wolniewicz, and many others. My particular thanks go to Krzysztof Bar for the experiences and conversations shared on and off the volleyball court throughout the eight years spent together in Oxford; and to Maciej Lisik for the conversations on economics and other topics, many of which took place on the University Parks tennis courts.

Finally, and most importantly, I would like to thank my parents and my grandparents for their love and support, and for helping me pursue my dreams.

Contents

Chapter 1

Delegation, Provision of Incentives, and Gaming of Performance Measures

1.1 Introduction

A common presumption in organisational design is that employees should not be given discretion regarding performance measures when their pay is performance-based. This is because they would otherwise have strong incentives to manipulate these measures for their own benefit, and may devote their attention and effort to tasks that boost apparent performance but are of little value to the company. For instance, if a manager whose compensation is tied to the profits or the market value of a company had control over the accounting practices, she would have strong incentives to manipulate the financial accounts or to make self-serving accounting choices.[1] It is therefore argued

[1]Fields, Lys, and Vincent (2001) point out that, given some degree of discretion allowed by the generally accepted accounting principles (GAAP), managers may opportunistically choose accounting methods that increase the stock price just before their stock options expire.

that companies are better off not delegating the decisions on accounting methods to managers (e.g., Watts and Zimmerman, 1986). Alternatively, imagine a teacher whose compensation is strongly based on her students' exam results. It would be natural not to give her much authority in setting the exam as it could encourage her to write easy questions or to teach to the test rather than convey deeper knowledge to the students.

In this paper, I analyse how delegating the choice of performance measure to a privately informed agent can allow the principal to offer contracts that screen the agents more effectively. I demonstrate that, by alleviating the problem of hidden information, delegation may bring benefits to the principal that outweigh the negative effects of gaming that it induces. Furthermore, the analysis of this trade-off identifies a set of environments where delegating the choice of performance measure is complementary to provision of incentives. For ease of exposition, I present the model in an education context with a school and a teacher in the roles of the principal and the agent. However, the results apply more generally to settings where (i) employees receive performance pay, and (ii) companies or organisations can choose how much discretion they give to employees about the choice of performance measure.

In the model, a principal (e.g., a school) hires an agent (e.g., a teacher) who can only be rewarded based on observed performance (e.g., students' exam results). The agent is privately informed about the degree of substitutability of his effort on two tasks. He is either "good", in which case he finds efforts on the two tasks to be perfect substitutes, or he is "bad", in which case they are not perfectly substitutable. From the principal's perspective, only effort on one of the tasks is valuable ("teaching deeper knowledge" as opposed to "teaching to the test").

Before offering a performance contract to the agent, the principal decides whether to delegate the choice of performance measure to the agent or to centralise it. Under delegation, the agent privately chooses either a high-quality performance measure ("a

difficult exam") or a low-quality one ("an easy exam"). Under centralisation, the principal ensures that the measure is of high quality but pays a fixed monitoring cost. With a low-quality measure, the observed performance may be improved with effort on either task; however, this is not possible with a high-quality measure, in which case only effort on the productive task is effective.

Whenever given discretion, a bad agent always engages in "gaming": he chooses a low-quality measure (i.e. an easy exam) so that he can improve his observed performance (i.e. students' exam results) by exerting effort on the unproductive task (i.e. by teaching to the test). He thus exerts positive effort on both tasks. On the other hand, a good agent perceives effort on the two tasks as perfect substitutes, and therefore—whether he selects a low- or a high-quality measure—he is always willing to exert all his effort on the productive task (i.e. on teaching deeper knowledge).

Although the principal is uninformed about the agent's type, she is allowed to use menus of contracts as a screening device. Under both delegation and centralisation, the optimal contract is a menu in which the compensation scheme designed for the good type of agent provides him with the same incentives as in the first best benchmark (i.e. no gaming and no hidden information). On the other hand, the compensation scheme designed for the bad agent is less "high-powered" than first best.[2]

The model offers two main results. First, the optimal incentives provided to the bad type of agent may be stronger when the principal allows the agent to privately choose the quality of his own performance measure than they are under centralisation. Given that the incentives provided to the good agent are the same in either case, this shows that provision of incentives can be complementary to giving discretion about performance measures. Second, even in the absence of any monitoring cost under centralisation, the principal's payoff may be higher when she delegates the choice of performance measure

[2]Incentives are more "high-powered" the steeper the slope of a linear, performance-contingent compensation scheme, e.g., see Holmström and Milgrom (1991).

to the agent than when she centralises it.

Both results arise because the impact of delegation is in fact twofold. On the one hand, delegation allows the bad agent to select a low-quality measure and then to improve apparent performance by exerting effort on the unproductive task. This gaming behaviour is costly to the principal because it boosts the agent's compensation but does not bring any direct benefit. It also discourages the principal from offering a high-powered compensation scheme to the bad type, since stronger incentives only exacerbate gaming.

On the other hand, by increasing the bad agent's expected utility from any given contract, delegation also allows the principal to decrease the fixed part of his compensation and still keep him willing to participate. This in turn relaxes the self-selection constraint for the good agent: the principal can lower the fixed part of the compensation for the good type and still make him prefer his contract to the one designed for the bad type. As a result, a lower rent is paid to the good agent for any given incentives provided to the bad type. Delegation therefore has a positive impact on the principal's payoff in that it alleviates the problem of hidden information. Furthermore, the fact the good agent's rent is reduced for any given incentives provided to the bad agent encourages the principal to offer the bad type a more high-powered contract.

If the agent is more likely to be bad than good, then the first effect dominates, which yields the *a priori* more natural results that (i) delegating the choice of performance measures to agents should be accompanied with provision of weaker incentives in the contract, and (ii) centralisation is preferred to delegation as long as monitoring is sufficiently cheap (and is always preferred when monitoring is costless). However, if it is more likely that the agent is of the good type, then the second effect dominates: giving discretion about performance measure to the agent allows the principal to better screen good and bad agents via performance contracts, and this positive effect outweighs the

negative effects of induced gaming. As a result, (i) delegation of performance measures and provision of incentives are complements, and (ii) delegation is preferable to centralisation even when monitoring is costless.

In fact, the second effect might be so strong that the good agent's expected rent from the optimal contract may be higher under delegation than under centralisation. Thus, an agent who does not engage in gaming may still benefit from being given discretion about the choice of performance measure. The reason for this is that while delegation reduces the good agent's rent for any given incentives provided to the bad agent, it also prompts the principal to offer stronger incentives to the bad agent as long as the agent is unlikely to be bad. If gaming is difficult enough and the probability that the agent is bad is sufficiently low, the latter effect dominates over the former, and therefore the good agent is better off when the principal delegates the choice of performance measure than when she centralises it. I use these observations to identify the circumstances under which the principal's and the teacher's preferences regarding delegation and centralisation are aligned: (i) delegation turns out to be Pareto optimal when the agent is very likely to be good and gaming is sufficiently difficult, and (ii) centralisation turns out to be Pareto optimal when the agent is very likely to be bad and gaming is sufficiently easy.

In Section 1.7, I provide further discussion of the assumptions of the model. I demonstrate that delegation does not alleviate the problem of hidden information in a knife-edge case where gaming is so easy that the principal can no longer use a menu of contracts as a screening device (Section 1.7.1), or if the teacher has limited liability (Section 1.7.2). In the appendix, I investigate—as a potential alternative to centralisation—the possibility of hiring an external monitor to select the performance measure, and investigate the necessary features of the monitor's contract.[3]

[3]The issues that are touched upon here are thus related to Alchian and Demsetz (1972), who discussed the need to provide monitors with appropriate incentives, and to Rahman (2012) who analyses

Related Literature

The paper is related to the broader theoretical literature on the interaction between provision of incentives and delegation of authority. Most of this literature suggests that the two are complementary in the majority of situations, i.e. the stronger the incentives provided to the employees in their contracts, the more authority they should be given when performing their tasks, and vice versa. One reason for this, as Milgrom and Roberts (1992) argue, comes from the incentive intensity principle, i.e. when a manager controls several divisions of a company, she has plenty of scope to improve performance of the company and therefore should be given strong financial incentives. Prendergast (2002) points to the role of uncertainty and demonstrates that more risky environments lead firms to delegate and offer output-based pay rather than assign tasks to employees and monitor their inputs.

Nevertheless, there are a few recent papers which try to identify settings where incentives and delegation can be substitutes, in particular Lando (2004), Bester and Krähmer (2008), De Varo and Prasad (2015), and Rohlfing-Bastian and Schöttner (2017).

Lando (2004) and De Varo and Prasad (2015) show that when effort and task selection are induced with a single performance measure as an instrument, giving high-powered incentives to a risk-averse agent may lead him to choose tasks with too low risk and too low return. De Varo and Prasad (2015) use their theoretical model to support their empirical analysis of data on British workplaces, which indicates that whether there is complementarity between delegation and incentives depends on the complexity of a job. My model bears some resemblance to Lando (2004) and De Varo and Prasad (2015) in that the principal can use only a single instrument to influence the

the problem of "monitoring the monitor" in a more general game-theoretic setting. There is also a distinct literature which studies optimal incentive contracts when performance evaluation is delegated to a reviewer who is possibly lenient towards the agent, e.g., Prendergast and Topel (1996), Giebe and Gürtler (2012), and Letina, Liu, and Netzer (2016).

agent's choice of performance measure and his allocation of efforts; however, I propose a different mechanism which determines whether delegation and incentives are substitutes or complements. Bester and Krähmer (2008) demonstrate that delegating the choice of a project to an agent who has different preferences over projects is less likely to be optimal when effort is non-observable and hence needs to be incentivised. Rohlfing-Bastian and Schöttner (2017) show that delegation and incentives may be substitutes when the available performance measures are "incongruent", i.e. they are too sensitive to effort on certain tasks relative to how these tasks contribute to actual productivity.

The empirical evidence on the complementarity between delegation and the provision of incentives is mixed. Most studies show a positive relationship between the two instruments, e.g., MacLeod and Parent (1999) for U.S. workers who are paid commissions, Nagar (2002) for bank managers, Colombo and Delmastro (2004) for plant managers in the Italian metalworking sector, Foss and Laursen (2005) for firms in Denmark, Wulf (2007) for divisional managers, De Varo and Kurtulus (2010) for British establishments, and Lo, Dessein, Ghosh, and Lafontaine (2016) for sales people who are constrained in terms of the maximum discount they can give to their customers. At the same time, De Varo and Prasad (2015), as stated earlier, and Jia and van Veen-Dirks (2015) present evidence that delegation and the provision of incentives may be substitutes. Hong, Kueng, and Yang (2016) demonstrate that stronger incentives lead to centralisation of decision-making from non-managerial to managerial employees.

The above literature focuses predominantly on the interaction between provision of incentives and delegation of authority about the choice of tasks that are to be performed. Little attention has been so far devoted to the study of delegation of authority on the *choice of performance measure*, which is the focus of my model. At the same time, it is worth noting that, in my model, the choice of performance measure also indirectly affects the agent's selection of tasks, so the paper does also contribute to the wider

literature on the provision of incentives and delegation.

Finally, the paper is also related to the literature on gaming, which has been defined by Ederer, Holden, and Meyer (2017) as "exploitation of an incentive scheme by an agent for his own self-interest to the detriment of the objectives of the incentive designer." Before formally investigating the benefits of deliberately opaque incentive schemes, they also list a number of specific forms which gaming can take. The type of gaming analysed in this paper has common features with two of them: "diversion of effort away from activities that are socially valuable but difficult to measure and reward, towards activities that are easily measured and rewarded," and "exploitation of the rules of classification to improve apparent, though not actual, performance."

1.2 The Model

A principal (e.g., a school or a university) hires a teacher to perform a job for her. The teacher can exert effort on two distinct tasks, A and B, which can be interpreted as "teaching to the test" and "teaching deeper knowledge", respectively.[4] The teacher's efforts, denoted by e_A and e_B, are not observable to the principal. The teacher is also privately informed about his cost of effort function. With probability $p \in (0,1)$, he is "bad" and his cost of effort is

$$c_{bad}(e_A, e_B) = (e_A)^2 + (e_B)^2, \tag{1.1}$$

which means that he perceives the efforts on the tasks as neither substitutes nor complements; otherwise, he is "good" and his cost function is

$$c_{good}(e_A, e_B) = \frac{1}{2}(e_A + e_B)^2, \tag{1.2}$$

[4]The naming of tasks follows Kerr (1975), "On the Folly of Rewarding A, While Hoping for B."

which means that he perceives the efforts on the tasks as perfect substitutes. In the next section, I discuss the rationale behind these cost of effort functions.

The principal's payoff is $V = R$ if the teacher succeeds in teaching deeper knowledge ($\mathcal{K} = 1$), and $V = 0$ if he does not succeed ($\mathcal{K} = 0$). The probability that the teacher succeeds in teaching deeper knowledge depends only on his effort on task B:

$$\Pr(\mathcal{K} = 1 \mid e_A, e_B) = e_B \quad \text{and} \quad \Pr(\mathcal{K} = 0 \mid e_A, e_B) = 1 - e_B. \tag{1.3}$$

The principal's payoff function reflects the idea that she cares only about the students obtaining deeper knowledge. I assume that $R < 1$ in order to restrict attention to interior solutions. Both the principal and the teacher are risk neutral. The teacher has no limited liability and his reservation utility is normalised to zero.

Teacher's contract. The principal offers a menu of compensation schemes denoted by (f, w), where f is the fixed (lump-sum) part of the compensation while w is its variable part. However, the teacher's compensation cannot be contingent on whether the students obtain deeper knowledge ($\mathcal{K} = 1$ or $\mathcal{K} = 0$); it can only be contingent on whether students pass the exam ($\mathcal{E} = 1$ or $\mathcal{E} = 0$). If the students fail the exam ($\mathcal{E} = 0$), the teacher receives only f, while if they pass the exam ($\mathcal{E} = 1$), he receives $f + w$.

There are two types of exam: easy and difficult. If the exam is easy, the probability that students pass the exam is

$$\Pr(\mathcal{E} = 1 \mid e_A, e_B) = \gamma e_A + e_B. \tag{1.4}$$

On the other hand, when the exam is difficult, the probability that students pass the

exam is

$$\Pr\left(\mathcal{E}=1 \mid e_A, e_B\right) = e_B. \tag{1.5}$$

Thus, teaching deeper knowledge is equally effective for both types of exam, while teaching to the test is effective only for an easy exam. The parameter $\gamma \in [0,1)$ can be interpreted as a measure of how easy gaming is.

Delegation vs. centralization. Before offering a contract to the teacher, the principal decides whether to delegate the choice of the exam to the teacher or to centralise it. Under delegation, the teacher privately chooses the type of exam. Under centralisation, the principal ensures that a difficult exam is set but pays a fixed monitoring cost of $c \geq 0$.

Timeline. The game proceeds as follows:

1. The principal (P) chooses whether to delegate the choice of the exam to the teacher (T) or to centralise it at cost c.
2. P offers a menu of contracts to T.[5]
3. T accepts or rejects the menu; if he has accepted, he selects one contract from the menu.
4. T chooses the type of exam (if the choice was delegated by P).
5. T allocates his teaching efforts, i.e. he chooses his efforts on tasks A and B.
6. The teaching outcome ($\mathcal{K}=0$ or $\mathcal{K}=1$), the exam outcome ($\mathcal{E}=0$ or $\mathcal{E}=1$), and the payoffs of P and T are realised.

[5]Including the delegation vs. centralisation dimension in the menu would not bring additional screening benefits to the principal here. In other words, the principal would not do any better in terms of her expected payoff if, instead of choosing between delegation and centralisation herself in stage 1 of the game, she were allowed to offer menus of contracts of the form (f, w, d), where $d \in \{0,1\}$ and $d=1$ ($d=0$) denotes that the choice of exam is (not) delegated to the teacher.

Other applications of the model. I focus here on the relationship between a school and a teacher for expositional reasons; nonetheless, the model admits other interpretations. The more general interpretation of the agent's choice of the type of exam is that the agent has discretion over the quality of the performance measure (which is not observable to the principal) according to which he is going to be compensated. If the agent chooses a low-quality performance measure, he can increase his compensation by exerting effort on the task that is worthless from the principal's point of view. For instance, the model could also apply to an environment in which a principal decides whether to allow a manager to have influence over the accounting rules used—with the rules being either lenient (easy exam) or strict (difficult exam), and the manager being rewarded based on his financial performance. Giving discretion to the manager would allow a bad one to choose lenient rules and then engage in actions that improve his apparent financial performance without increasing the value of the company.

1.3 Preliminaries

1.3.1 Agent's Expected Utility from a Contract

When a good teacher sets an easy exam, the utility that he expects to get from a contract (f, w) is $f + (\gamma e_A + e_B)\, w - \frac{1}{2}\left(e_A + e_B\right)^2$. The optimal effort choices are then $(e_A, e_B) = (0, w)$. Hence, a good teacher's expected utility with an easy exam is

$$U_{good,easy}\left(f, w\right) = f + \frac{1}{2}w^2. \tag{1.6}$$

When a good teacher sets a difficult exam, his expected utility from a contract (f, w) is $f + e_B w - \frac{1}{2}\left(e_A + e_B\right)^2$. This yields exactly the same effort choices as with an easy

exam, i.e. $(e_A, e_B) = (0, w)$, and his expected utility with a difficult exam is therefore

$$U_{good,diff}(f, w) = f + \frac{1}{2}w^2. \tag{1.7}$$

It follows that a good teacher is indifferent between setting an easy and a difficult exam, and that he always exerts a strictly positive effort only on teaching deeper knowledge.

Consider now a bad teacher. When a bad teacher sets an easy exam, his expected utility from a contract (f, w) is $f + (\gamma e_A + e_B) w - (e_A)^2 - (e_B)^2$. His optimal effort choices are then $(e_A, e_B) = \left(\frac{1}{2}\gamma w, \frac{1}{2}w\right)$, which means that his expected utility with an easy exam is

$$U_{bad,easy}(f, w) = f + \frac{1}{4}\left(1 + \gamma^2\right) w^2. \tag{1.8}$$

On the other hand, when a bad teacher sets a difficult exam, his expected utility from a contract (f, w) is $f + e_B w - (e_A)^2 - (e_B)^2$. Consequently, his optimal effort choices are $(e_A, e_B) = \left(0, \frac{1}{2}w\right)$, and hence the expected utility that he derives from a difficult exam is

$$U_{bad,diff}(f, w) = f + \frac{1}{4}w^2. \tag{1.9}$$

Note here that $U_{bad,easy}(f, w) \geq U_{bad,diff}(f, w)$ and that the inequality is strict whenever $\gamma > 0$. Thus, when $\gamma > 0$ and the choice of the exam is delegated, a bad teacher always strictly prefers to set an easy one and then to exert strictly positive effort on teaching to the test.

Since $U_{bad,easy}(f, w) \geq U_{bad,diff}(f, w)$ and $U_{good,easy}(f, w) = U_{good,diff}(f, w)$, it is only the bad teacher who benefits from delegation for a given contract (f, w). Furthermore, because $\gamma < 1$, we have that $U_{bad,easy}(f, w) < U_{good,easy}(f, w) = U_{good,diff}(f, w)$, which allows the principal to use menus of contracts as a screening device.

Rationale behind the teacher's cost of effort function. The cost of effort functions of the two types of teacher are constructed in such a way that if the ease of gaming were $\gamma = 1$, the maximum expected utility of a bad teacher given a contract (f, w) would be the same as that of a good teacher: $U_{bad,easy}(f, w) = U_{good,easy}(f, w) = U_{good,diff}(f, w)$. This is because, given optimal effort allocations of the two types of teacher, we would have that (i) the probability that the students pass the exam is the same for both of them, and (ii) they have the same level of the cost of effort. As a result, separating contracts would become infeasible.

In other words, the assumption of $\gamma < 1$ is equivalent to saying that the problem of gaming is not so acute as to prevent the principal from using menus of contracts as a screening device under hidden information. A more general cost of effort function would be $c(e_A, e_B) = \left(1 - \frac{\delta}{2}\right)\left((e_A)^2 + (e_B)^2\right) + \delta e_A e_B$, where δ measures the substitutability of effort across tasks.[6] The cost of effort functions of the bad type and the good type are its special cases with, respectively, $\delta = 0$ and $\delta = 1$.[7]

1.3.2 The No Hidden Information Benchmark

Suppose that the principal can observe the teacher's type but allows him to privately choose the type of exam. Henceforth, I will refer to this setting as the "no hidden information" (NHI) benchmark.

[6]This function is notably different from the more commonly seen $\frac{1}{2}\left(c_A(e_A)^2 + c_B(e_B)^2\right) + \delta e_A e_B$, where $\delta = \sqrt{c_A c_B}$ implies that efforts on the two tasks are perfect substitutes, while $\delta = 0$ means the agent's cost of effort is a sum of two convex (quadratic) functions. But an increase in δ has then also another effect in that it weakly increases the total cost of effort. My specification corrects for that and allows me to isolate the role of substitutability of effort across tasks.

[7]This contrasts with Holmström and Milgrom (1991) who show that when efforts are highly substitutable for the agent, the "effort substitution problem" may arise: e.g., if performance on teaching basic skills is more easily measured than that on teaching higher-thinking skills, a teacher may substitute effort away from the latter. Here, the principal is hurt by the presence of the bad teacher who perceives the efforts on the tasks as less substitutable than the good teacher does.

Good teacher. Given a contract (f_g, w_g), a good teacher is indifferent between the two types of exam but with either of them he chooses an effort allocation $(e_A, e_B) = (0, w_g)$, which implies that $\Pr[\mathcal{K}=1] = \Pr[\mathcal{E}=1] = w_g$. The principal's maximisation problem is then

$$\begin{aligned} &\max_{f_g, w_g} w_g R - f_g - w_g^2 \text{ s.t.} \\ &\text{(IRg) } f_g + \frac{1}{2} w_g^2 \geq 0, \end{aligned} \tag{1.10}$$

where IRg denotes the individual rationality constraint of the good type of teacher. In the optimal contract for the good teacher, the IRg constraint is binding: $f_g = -\frac{1}{2} w_g^2$, and $f_g^{NHI} = -\frac{1}{2} R^2$ and $w_g^{NHI} = R$. Thus, the principal effectively "sells the project" for a fee that makes the good teacher indifferent between accepting and rejecting the contract offer.

Bad teacher. Given a contract (f_b, w_b), a bad teacher chooses an easy exam and an effort allocation $(e_A, e_B) = \left(\frac{1}{2}\gamma w_b, \frac{1}{2} w_b\right)$. The resulting probability that students obtain deeper knowledge is $\Pr[\mathcal{K}=1] = \frac{1}{2} w_b$, whereas the probability that they pass the exam is $\Pr[\mathcal{E}=1] = \frac{1}{2}(1+\gamma^2) w_b$. The principal's maximisation problem is then

$$\begin{aligned} &\max_{f_b, w_b} \frac{1}{2} w_b R - f_b - \frac{1}{2}\left(1+\gamma^2\right) w_b^2 \text{ s.t.} \\ &\text{(IRb) } f_b + \frac{1}{4}\left(1+\gamma^2\right) w_b^2 \geq 0, \end{aligned} \tag{1.11}$$

where IRb denotes the bad type's individual rationality constraint. In the optimal contract for the bad teacher, the IRb constraint is binding: $f_b = -\frac{1}{4}(1+\gamma^2) w_b^2$. Hence, for any given w_b, the easier the gaming is, the lower the optimal lump-sum compensation f_b. Given the binding IRb constraint, the optimal contract for the bad teacher is $f_b^{NHI} = -\frac{R^2}{4(1+\gamma^2)}$ and $w_b^{NHI} = \frac{R}{1+\gamma^2}$. As gaming becomes easier, the bad teacher's

compensation becomes less high-powered. This is because easier gaming boosts the bad teacher's expected compensation for any given incentives he is provided but does not bring any benefit to the principal.

To summarise:

Lemma 1.1. *When the principal can observe the teacher's type, the optimal contract is a menu of* $\left(f_b^{NHI}, w_b^{NHI}\right) = \left(-\frac{R^2}{4(1+\gamma^2)}, \frac{R}{1+\gamma^2}\right)$ *and* $\left(f_g^{NHI}, w_g^{NHI}\right) = \left(-\frac{1}{2}R^2, R\right)$. *The principal's expected payoff is*

$$\Pi_{NHI}^* = \frac{1}{2}R\left[\frac{1}{2}pw_b^{NHI} + (1-p)\, w_g^{NHI}\right]. \tag{1.12}$$

Thus, the optimal contract for the good teacher is more high-powered than the one for the bad teacher, $w_g^{NHI} > w_b^{NHI}$, and the difference is driven by how easy gaming is, γ. Lemma 1.1 also implies that the principal's maximised expected payoff is linear in $\mathbb{E}\left[V \mid w_b^{NHI}, w_g^{NHI}\right] = \frac{1}{2}pw_b^{NHI} + (1-p)\, w_g^{NHI}$, i.e. her expected reward from students obtaining deeper knowledge given w_b^{NHI} and w_g^{NHI}.

Finally, suppose that—in addition to the principal observing the teacher's type—we do not allow the teacher to engage in gaming, i.e. he cannot set an easy exam or exert any effort on teaching to the test. Henceforth, I will refer to this setting as the "first best" (FB) benchmark. To find the optimal contract for the teacher, we substitute $\gamma = 0$ into the optimal compensation schemes in the NHI benchmark.

Lemma 1.2. *When the teacher cannot engage in gaming (i.e. he cannot set an easy exam or exert any effort on task A) and the principal can observe his type, the optimal contract is a menu of* $\left(f_b^{FB}, w_b^{FB}\right) = \left(-\frac{1}{4}R^2, R\right)$ *and* $\left(f_g^{FB}, w_g^{FB}\right) = \left(-\frac{1}{2}R^2, R\right)$. *The principal's expected payoff is*

$$\Pi_{FB}^* = \frac{1}{2}R\left[\frac{1}{2}pw_b^{FB} + (1-p)\, w_g^{FB}\right]. \tag{1.13}$$

Similarly to the NHI benchmark, the principal's maximised expected payoff is linear in $\mathbb{E}\left[V \mid w_b^{FB}, w_g^{FB}\right]$. Since $w_b^{NHI} < w_b^{FB}$ and $w_g^{NHI} = w_g^{FB}$, the principal's maximised expected payoff in the NHI benchmark is—unsurprisingly—lower than the first best: $\Pi_{NHI}^* < \Pi_{FB}^*$. Thus, when there is no hidden information, the principal is always better off not delegating the choice of exam to the teacher (or, in other words, not permitting gaming). I now show how this conclusion may be reversed when there is hidden information.

1.4 The Optimal Contracting Scheme

1.4.1 Optimal Contract under Delegation

Let us now consider the scenario where the teacher is privately informed about his type and is allowed by the principal to privately choose the type of exam. A bad teacher's maximised expected utility given a contract (f, w) is then $f + \frac{1}{4}\left(1 + \gamma^2\right) w^2$, which—as long as $\gamma < 1$—is strictly less than a good teacher's maximised expected utility, $f + \frac{1}{2}w^2$. This difference allows the principal to use menus of contracts as a screening device under hidden information.

Consider a menu of $\left(\left(f_b, w_b\right), \left(f_g, w_g\right)\right)$ in which each type of teacher selects the compensation scheme designed for himself. The effort allocation of a bad teacher is then $(e_A, e_B) = \left(\frac{1}{2}\gamma w_b, \frac{1}{2}w_b\right)$, which means that $\Pr\left[\mathcal{K} = 1\right] = \frac{1}{2}w_b$ and $\Pr\left[\mathcal{E} = 1\right] = \frac{1}{2}\left(1 + \gamma^2\right) w_b$. Thus, for any given w_b, the probability that students pass the exam is higher than the probability that they obtain deeper knowledge, and the difference increases as gaming becomes easier. On the other hand, the effort allocation of a good teacher is $(e_A, e_B) = (0, w_g)$, which implies $\Pr\left[\mathcal{K} = 1\right] = \Pr\left[\mathcal{E} = 1\right] = w_g$.

The principal's maximisation problem is now

$$\begin{aligned}
&\max_{f_b,w_b,f_g,w_g} p\left[\frac{1}{2}w_bR - f_b - \frac{1}{2}\left(1+\gamma^2\right)w_b^2\right] + (1-p)\left[w_gR - f_g - w_g^2\right] \text{ s.t.} \\
&\text{(IRb) } f_b + \frac{1}{4}\left(1+\gamma^2\right)w_b^2 \geq 0 \\
&\text{(IRg) } f_g + \frac{1}{2}w_g^2 \geq 0 \\
&\text{(SSb) } f_b + \frac{1}{4}\left(1+\gamma^2\right)w_b^2 \geq f_g + \frac{1}{4}\left(1+\gamma^2\right)w_g^2 \\
&\text{(SSg) } f_g + \frac{1}{2}w_g^2 \geq f_b + \frac{1}{2}w_b^2, \qquad (1.14)
\end{aligned}$$

where IRb and IRg denote the individual rationality while SSb and SSg denote the self-selection constraints of the two types of teacher.

Let us start the analysis by assuming that the principal wants to attract both types of teacher and hence that the IRb constraint is binding: $f_b = -\frac{1}{4}\left(1+\gamma^2\right)w_b^2$. This means that the easier the gaming is, the higher the "fee" which the principal can charge to the bad teacher. Substituting this expression for f_b into the SSg constraint, we obtain

$$f_g + \frac{1}{2}w_g^2 \geq \frac{1}{4}\left(1-\gamma^2\right)w_b^2, \qquad (1.15)$$

which implies that a good teacher receives a strictly positive rent. In the optimal screening contract, the principal sets f_g as low as possible while satisfying the SSb and SSg constraints, so it must be that the SSg constraint binds. Thus, the good teacher's rent equals $\frac{1}{4}\left(1-\gamma^2\right)w_b^2$ and the principal faces the following trade-off when designing the menu of contracts: as she provides stronger incentives to the bad teacher, it becomes more difficult to construct a contract that a good teacher would prefer to the one designed for the bad teacher, and therefore the good teacher must be given a higher rent.

Note also that, since the left hand side of (1.15) is the good teacher's expected

utility from a contract (f_g, w_g), the scenario where the IRg constraint binds is in fact a special case of the above where the principal sets $w_b = 0$. But since (i) the principal's expected benefit is linear in w_b, and (ii) the good teacher's rent and the bad teacher's expected compensation are both quadratic in w_b, it is always optimal for the principal to offer $w_b > 0$ and thereby induce a strictly positive effort from the bad teacher. In other words, this confirms that the optimal menu of contracts always attracts here both types of teacher.

Furthermore, given the binding IRb and SSg constraints, the SSb constraint implies $w_g \geq w_b$. Thus, in the optimal contract, the compensation scheme designed for the good teacher must be more high-powered than the one for the bad teacher. The principal's maximisation problem now simplifies to

$$\begin{aligned} &\max_{w_b, w_g} p\left[\frac{1}{2}w_b R - \frac{1}{4}\left(1+\gamma^2\right)w_b^2\right] + (1-p)\left[w_g R - \frac{1}{4}\left(1-\gamma^2\right)w_b^2 - \frac{1}{2}w_g^2\right] \text{ s.t.} \\ &\text{(SSb) } w_g \geq w_b. \end{aligned} \tag{1.16}$$

Ignoring for now the SSb constraint and maximising the objective function with respect to w_b and w_g yields $w_b^{\mathcal{D}} = \frac{pR}{1-\gamma^2(1-2p)}$ and $w_g^{\mathcal{D}} = R$. It can be easily checked that $w_g^{\mathcal{D}} > w_b^{\mathcal{D}}$, so the SSb constraint is indeed satisfied.[8] We thus obtain the following proposition:

Proposition 1.1. *When the teacher is privately informed about his type and the principal delegates the choice of exam to him, the optimal contract for the teacher is a menu of*

$$\left(f_b^{\mathcal{D}}, w_b^{\mathcal{D}}\right) = \left(-\frac{\left(1+\gamma^2\right)p^2R^2}{4\left(1-\gamma^2\left(1-2p\right)\right)^2}, \frac{pR}{1-\gamma^2\left(1-2p\right)}\right) \tag{1.17}$$

[8]The fact that such a separating contract is always feasible (as long as $\gamma < 1$) also means that a pooling contract with the IRb constraint binding can never be optimal.

and

$$\left(f_g^{\mathcal{D}}, w_g^{\mathcal{D}}\right) = \left(\frac{\left(1-\gamma^2\right)p^2R^2}{4\left(1-\gamma^2\left(1-2p\right)\right)^2} - \frac{1}{2}R^2, R\right). \tag{1.18}$$

The principal's expected payoff is

$$\Pi_{\mathcal{D}}^* = \frac{1}{2}R\left[\frac{1}{2}pw_b^{\mathcal{D}} + \left(1-p\right)w_g^{\mathcal{D}}\right]. \tag{1.19}$$

Since the good teacher's rent is increasing in w_b, the value of $w_b^{\mathcal{D}}$ is distorted downwards relative to its counterpart in the NHI benchmark whereas $w_g^{\mathcal{D}}$ is not. As a result, we have $w_b^{\mathcal{D}} < w_b^{NHI} < w_b^{FB}$ and $w_g^{\mathcal{D}} = w_g^{NHI} = w_g^{FB}$: the optimal contract provides the "correct" (i.e. first best) incentives to the good type of teacher while those given to the bad teacher are weaker than in the NHI and FB benchmarks. The fact that $w_g^{\mathcal{D}} = R$ means that the principal effectively "sells the project" to the good type for a certain fee. However, unlike in the NHI benchmark, this fee does not fully offset the good teacher's expected benefit—he receives a strictly positive rent.

Furthermore, like in the NHI and FB benchmarks discussed in Section 1.3, the principal's expected payoff is here linear in $\mathbb{E}\left[V \mid w_b^{\mathcal{D}}, w_g^{\mathcal{D}}\right]$, i.e. her expected reward from students obtaining deeper knowledge given $w_b^{\mathcal{D}}$ and $w_g^{\mathcal{D}}$. Therefore, the observation that $w_b^{\mathcal{D}} < w_b^{NHI} < w_b^{FB}$ and $w_g^{\mathcal{D}} = w_g^{NHI} = w_g^{FB}$ implies also that the principal's expected payoff is, unsurprisingly, lower than in the NHI and FB benchmarks.

1.4.2 Optimal Contract under Centralisation

Let us now consider the optimal contract under centralisation. When the principal centralises the choice of exam, she ensures that a difficult exam is set and pays a fixed monitoring cost of $c \geq 0$. There is thus no scope for gaming: the bad teacher has no incentive to exert any effort on teaching to the test. To find the optimal contract under centralisation, we substitute $\gamma = 0$ into the optimal contract under delegation that was

specified in Proposition 1.1:

Proposition 1.2. *Consider* $\left(f_b^{\mathcal{D}}(\gamma), w_b^{\mathcal{D}}(\gamma)\right)$, $\left(f_g^{\mathcal{D}}(\gamma), w_g^{\mathcal{D}}(\gamma)\right)$, *and* $\Pi_{\mathcal{D}}^*(\gamma)$. *When the teacher is privately informed about his type and the principal centralises the choice of exam, the optimal contract for the teacher is a menu of* $\left(f_b^{\mathcal{C}}, w_b^{\mathcal{C}}\right) = \left(f_b^{\mathcal{D}}(0), w_b^{\mathcal{D}}(0)\right)$ *and* $\left(f_g^{\mathcal{C}}, w_g^{\mathcal{C}}\right) = \left(f_g^{\mathcal{D}}(0), w_g^{\mathcal{D}}(0)\right)$. *The principal's expected payoff is* $\Pi_{\mathcal{C}}^* = \Pi_{\mathcal{D}}^*(0) - c$.

We thus have $\Pi_{\mathcal{C}}^* = \frac{1}{2}R\left[\frac{1}{2}pw_b^{\mathcal{C}} + (1-p)w_g^{\mathcal{C}}\right] - c$, where $w_b^{\mathcal{C}} = pR$ and $w_g^{\mathcal{C}} = R$.

Since the optimal contract under centralisation is the same as in a setting with hidden information but gaming not being allowed, we can use the following table to summarise the relationship between the four contracts analysed so far:

	no gaming	gaming
no hidden information	First Best Benchmark	NHI Benchmark
hidden information	Centralisation	Delegation

Table 1.1: The relationship between the contracts analysed in Section 1.4 and the benchmarks of Section 1.3.

1.4.3 Further Intuition

For a better intuition behind the features of the optimal contracts, it is worth having a closer look at how the incentives provided to the teacher and the principal's expected payoff are affected by the probability that the teacher is bad.

Strength of optimal incentives. It can be easily checked that $w_b^{\mathcal{D}}$ is strictly increasing in p, and that $\lim_{p\to 1} w_b^{\mathcal{D}} = \frac{R}{1+\gamma^2} = w_b^{NHI}$. The intuition here is that as it becomes more likely that the teacher is bad, the less likely it is that the principal faces the good type who receives a positive rent under the optimal contract—and this rent is increasing in the incentives provided to the bad type. There is thus more benefit

from providing stronger incentives to the bad teacher and, as a result, the choice of $w_b^{\mathcal{D}}$ is less distorted relative to the NHI benchmark. In the limit, when $p \to 1$, the probability of having to pay the rent to the good type approaches zero, and consequently $w_b^{\mathcal{D}}$ approaches w_b^{NHI}. Note also that $\lim_{p\to 1} w_b^{\mathcal{C}} = R = w_b^{FB}$, which follows from the earlier observation that $\lim_{p\to 1} w_b^{\mathcal{D}} = \frac{R}{1+\gamma^2} = w_b^{NHI}$ and from the fact that centralisation ensures that there is no gaming.

The fact that $w_b^{\mathcal{D}}$ is strictly increasing in p while $w_g^{\mathcal{D}}$ stays constant does not imply that the optimal incentives provided to the teacher are stronger the more likely he is to be bad. Figure 1.1 illustrates how the probability that the teacher is bad affects the "mean power of incentives", which I define as $\overline{w} = pw_b + (1-p)\, w_g$, for different values of the ease of gaming, γ.

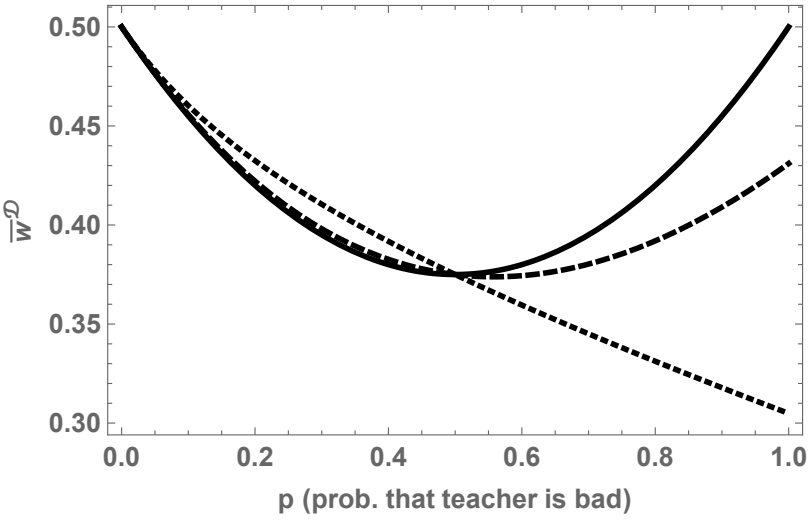


Figure 1.1: The mean power of incentives, $\overline{w}^{\mathcal{D}} = pw_b^{\mathcal{D}} + (1-p)\, w_g^{\mathcal{D}}$, provided by the optimal contract under delegation when $\gamma = 0$ (solid line), when $\gamma = \frac{2}{5}$ (dashed line), and when $\gamma = \frac{4}{5}$ (dotted line). The simulation assumes $R = \frac{1}{2}$.

Intuitively, when gaming is difficult, i.e. γ is low, the effects of hidden information dominate, and therefore the mean power of incentives is highest when there is little uncertainty about the type of teacher and is lowest for some intermediate value of p. In the limit, when $\gamma = 0$ (e.g., under centralisation), the mean power of incentives attains

its minimum when $p = \frac{1}{2}$, i.e. when the problem of hidden information is most acute. On the other hand, when γ is sufficiently high, the effects of gaming dominate over those of hidden information, and an increase in the probability that the teacher is bad always leads to a lower mean power of incentives in the optimal contract.

Principal's expected payoffs. It can be easily verified from (1.19) and $\lim_{p\to 1} w_b^{\mathcal{P}} = w_b^{NHI}$ that when $p \to 0$ or $p \to 1$, the principal's expected payoff under delegation approaches that in the NHI benchmark. In other words, the principal's loss from hidden information approaches zero as the uncertainty about the teacher's type disappears. Furthermore, since centralisation ensures that there is no gaming but entails a cost of c, the principal's expected payoff under centralisation approaches $\Pi_{FB}^* - c$ as $p \to 0$ or $p \to 1$.

The following corollary to Proposition 1.1 shows that the principal's loss from hidden information is highest for some intermediate value of $p \in \left(0, \frac{1}{2}\right]$:

Corollary 1.1. *The principal's loss from hidden information under delegation,* $\Pi_{NHI}^* - \Pi_{\mathcal{D}}^*$*, is strictly increasing in* p *when* $p < \widehat{p}(\gamma)$ *and is strictly decreasing in* p *when* $p > \widehat{p}(\gamma)$*, where* $\widehat{p}'(\gamma) < 0$*,* $\widehat{p}(0) = \frac{1}{2}$*, and* $\lim_{\gamma\to 1} \widehat{p}(\gamma) = 0$*.*

Thus, when the ease of gaming is $\gamma = 0$, the loss is largest when $p = \frac{1}{2}$, i.e. when the problem of hidden information is most acute. This also means that the difference between Π_{FB}^* and $\Pi_{\mathcal{C}}^*$ is largest when $p = \frac{1}{2}$. As gaming becomes easier, the principal's loss from hidden information under delegation is at its maximum for smaller and smaller values of p.

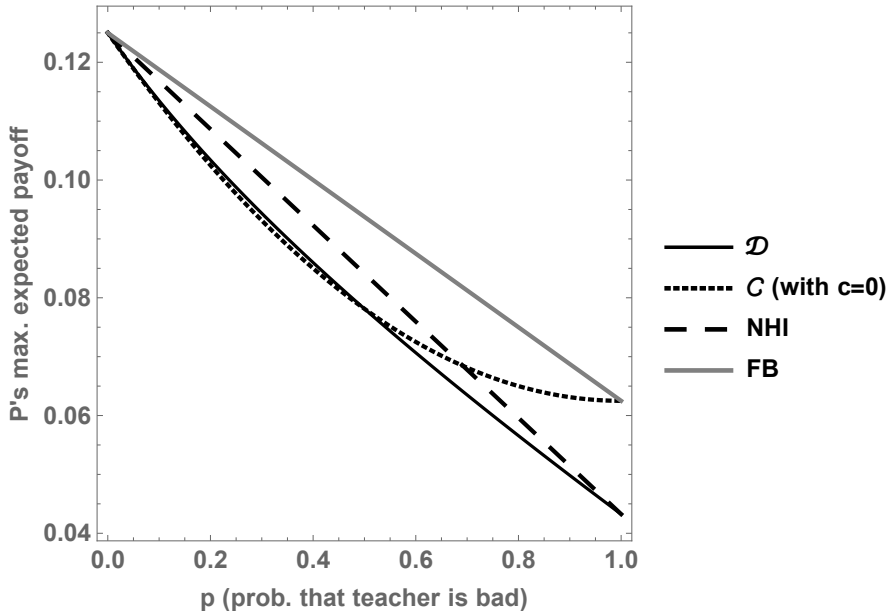


Figure 1.2: The impact of the probability that the teacher is bad, p, on the principal's maximum expected payoff under delegation and under centralisation (with $c = 0$), and in the FB and NHI benchmarks. The simulation assumes $R = \frac{1}{2}$ and $\gamma = \frac{2}{3}$.

Figure 1.2 illustrates how the probability that the teacher is bad affects the principal's maximum expected payoff under delegation and under centralisation (with $c = 0$), and compares them to the equivalents in the FB and NHI benchmarks.

1.5 Are Delegation and Provision of Incentives Complements or Substitutes?

In order to identify the settings where delegation and incentives are complements or substitutes in this model, it proves useful to first look at the following corollary to Proposition 1.1, which shows how the incentives provided to the bad teacher in the optimal contract vary depending on how easy gaming is:

Corollary 1.2. *The variable part of the optimal contract designed for the bad teacher under delegation, $w_b^{\mathcal{P}}$, has the following properties with respect to the ease of gaming,*

γ:

(i) $w_b^{\mathcal{D}}$ *is strictly increasing in* γ *when* $p < \frac{1}{2}$*, invariant to* γ *when* $p = \frac{1}{2}$*, and strictly decreasing in* γ *when* $p > \frac{1}{2}$*;*

(ii) $\lim_{\gamma \to 1} w_b^{\mathcal{D}} = \frac{R}{2} = \lim_{\gamma \to 1} w_b^{NHI}$.

It may seem surprising that easier gaming may prompt the principal to provide stronger incentives to the bad teacher. Given that $w_g^{\mathcal{D}}$ does not vary with γ, part (i) of Corollary 1.2 also implies that the mean power of incentives is strictly increasing in γ whenever $p < \frac{1}{2}$ (which we can also observe in Figure 1.1). As we have seen in Section 1.3.2, this is not possible when there is no hidden information.

The result in part (i) of Corollary 1.2 arises because an increase in γ has two distinct effects on the optimal incentives provided to the bad teacher.

The first effect is that it increases the bad teacher's incentives to exert effort on task A, i.e. on "teaching to the test", for any given w_b. This is costly to the principal because such effort improves the students' chances of passing the exam—and hence the teacher's expected compensation—but at the same time it does not bring any direct benefit to the principal.

The second effect is that, for any given w_b, the good teacher's rent decreases as gaming becomes easier. Intuitively, since the bad teacher's maximum expected utility from a contract (f, w) under delegation is $f + \frac{1}{4}(1 + \gamma^2) w^2$, a higher value of γ means that the bad teacher reaps a larger benefit from being able to privately choose the exam. Hence, the higher γ is, the more the principal can lower the lump-sum compensation for the bad type while still satisfying his individual rationality constraint. This also lowers the good teacher's expected utility from selecting the contract designed for the bad teacher, and so it becomes easier to construct a contract for the good type that satisfies his self-selection constraint. The rent given to the good teacher by the optimal contract is thus reduced.

In the limit, when $\gamma \to 1$, the expected utility obtained by the bad teacher from any given contract (f, w) approaches that of the good teacher, i.e. $f + \frac{1}{2}w^2$. Consequently, the lump-sum compensation for the bad type can be lowered to such a level that the good teacher's utility from selecting the contract designed for the bad teacher is arbitrarily close to zero. In other words, the binding SSg constraint implies then that the IRg constraint is virtually binding, and hence the rent given to the good teacher by the optimal contract approaches zero. The zero rent means that the principal's optimal choice of w_b is then not distorted by the "externality" arising from the need to pay a rent to the good type, and therefore the incentives provided to the bad teacher in the optimal contract under delegation approach those in the NHI benchmark—as stated in part (ii) of Corollary 1.2.

Overall, in the optimal contract under delegation, the bad teacher's expected compensation is $f_b^{\mathcal{D}} + \frac{1}{2}(1+\gamma^2)\left(w_b^{\mathcal{D}}\right)^2 = \frac{1}{4}(1+\gamma^2)\left(w_b^{\mathcal{D}}\right)^2$, while the good teacher's expected rent is $\frac{1}{4}(1-\gamma^2)\left(w_b^{\mathcal{D}}\right)^2$. If the probability that the teacher is bad, p, is more than $\frac{1}{2}$, the impact of γ on raising the former dominates over the impact on lowering the latter, and consequently w_b is decreasing in γ. On the other hand, when $p < \frac{1}{2}$, the reverse holds and the optimal incentives provided to the bad teacher become stronger as gaming becomes easier.

Finally, because the optimal contract under centralisation is the same as the optimal contract under delegation for the case of $\gamma = 0$, part (i) of Corollary 1.2 and the fact that $w_g^{\mathcal{D}}$ is invariant to γ imply that: (i) when $p < \frac{1}{2}$, we have $w_b^{\mathcal{D}} > w_b^{\mathcal{C}}$ and $w_g^{\mathcal{D}} = w_g^{\mathcal{C}}$, and (ii) when $p > \frac{1}{2}$, we have $w_b^{\mathcal{D}} < w_b^{\mathcal{C}}$ and $w_g^{\mathcal{D}} = w_g^{\mathcal{C}}$. We thus obtain the following proposition:

Proposition 1.3. *Relative to centralisation, delegating the choice of exam to the teacher makes the optimal compensation scheme more high-powered when $p < \frac{1}{2}$ and less high-powered when $p > \frac{1}{2}$.*

In other words, provision of incentives is complementary to delegation when $p < \frac{1}{2}$, and is a substitute for it when $p > \frac{1}{2}$.

Figure 1.3 illustrates the impact of γ on the optimal incentives provided to the bad teacher under delegation and under centralisation, and compares them to the NHI and FB benchmarks.

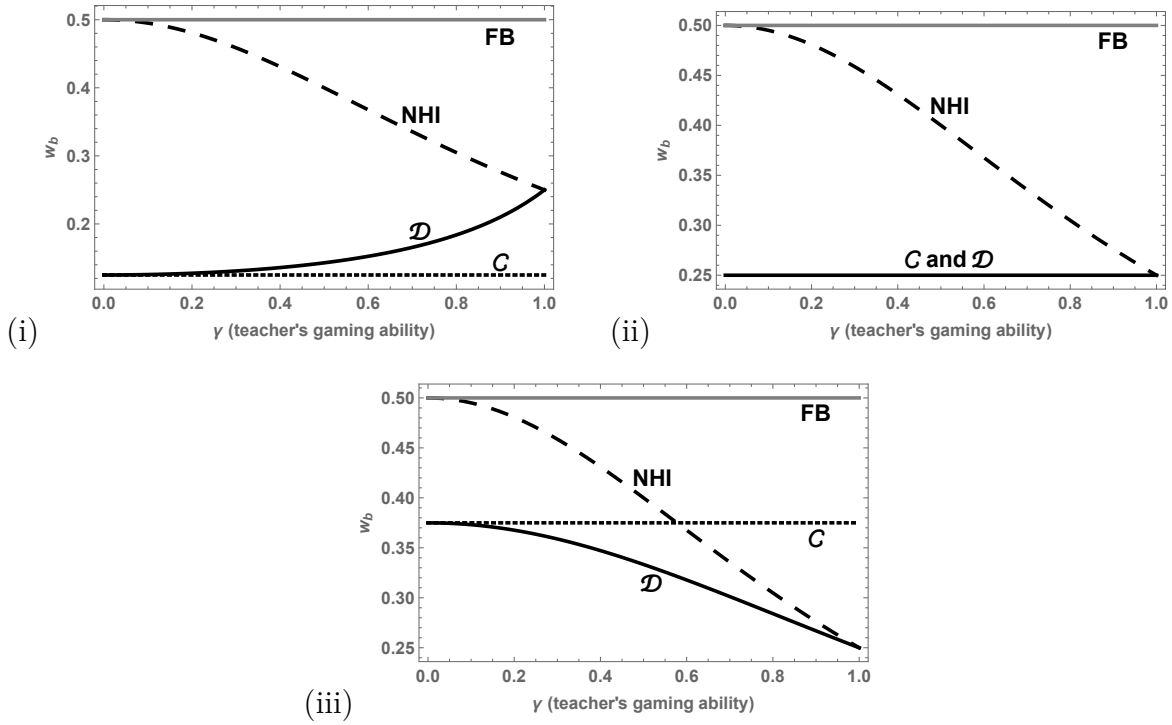


Figure 1.3: The impact of γ on the optimal value of w_b in the optimal contracts under delegation, under centralisation, and in the NHI and FB benchmarks: (i) when $p = \frac{1}{4}$, (ii) when $p = \frac{1}{2}$, and (iii) when $p = \frac{3}{4}$. The simulation assumes $R = \frac{1}{2}$.

Figure 1.3(i) shows that when $p < \frac{1}{2}$, the optimal contract designed for a bad teacher is more high-powered under delegation than it is under centralisation, and that it brings the incentives provided to the bad type closer not only to the NHI but also the FB benchmark. On the other hand, 1.3(iii) demonstrates that when $p > \frac{1}{2}$, the optimal contract designed for a bad teacher is less high-powered under delegation than it is under centralisation. This is because the negative effect of gaming outweighs here the positive effect of being able to better screen good and bad agents; as a result, even

though delegation brings the optimal incentives provided to the bad type closer to the NHI benchmark, it still pushes them down and away from the first best.

1.6 When Is Delegation Optimal?

1.6.1 Principal's Optimal Choice of Delegation vs. Centralisation

Before we specify the settings where it is optimal for the principal to delegate the choice of exam to the teacher, let us have a look at the following corollary to Proposition 1.1, which demonstrates how the principal's expected payoff from the optimal contract under delegation varies depending on how easy gaming is:

Corollary 1.3. *The principal's expected payoff from the optimal contract under delegation, $\Pi^*_{\mathcal{D}}$, has the following properties with respect to the ease of gaming, γ:*

(i) $\Pi^*_{\mathcal{D}}$ is strictly increasing in γ when $p < \frac{1}{2}$, invariant to γ when $p = \frac{1}{2}$, and strictly decreasing in γ when $p > \frac{1}{2}$;

(ii) $\lim_{\gamma \to 1} \Pi^*_{\mathcal{D}} = \lim_{\gamma \to 1} \Pi^*_{NHI}$.

Note that the statements in Corollary 1.3 are parallel to those in Corollary 1.2. In fact, since (i) the principal's maximised expected payoff under delegation, $\Pi^*_{\mathcal{D}}$, is increasing and linear in $w_b^{\mathcal{D}}$ and $w_g^{\mathcal{D}}$, and (ii) $w_g^{\mathcal{D}}$ is invariant to γ, part (i) of Corollary 1.2 implies part (i) of Corollary 1.3. Furthermore, the fact that $w_b^{\mathcal{D}}$ approaches w_b^{NHI} as $\gamma \to 1$ while $w_g^{\mathcal{D}} = w_g^{NHI}$ implies part (ii) of Corollary 1.3.

Figure 1.4 illustrates the impact of γ on the principal's maximised expected payoff under delegation and under centralisation (with $c = 0$), and compares it to the NHI and FB benchmarks. The striking similarity to Figure 1.3 follows from the fact that

(i) $\Pi^*_{\mathcal{D}}$, $\Pi^*_{\mathcal{C}}$, Π^*_{NHI} and Π^*_{FB} are all linear in the respective optimal values of w_b and w_g, and (ii) the optimal value of w_g is invariant to γ in each of the four cases.

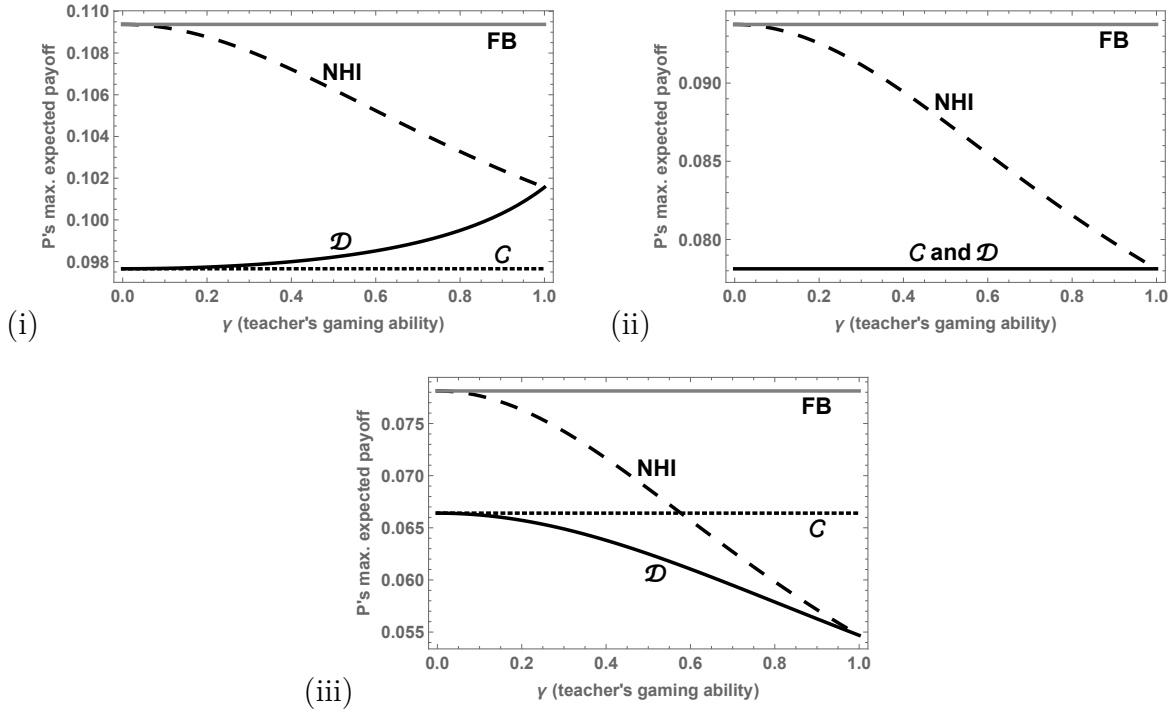


Figure 1.4: The impact of γ on the principal's maximum expected payoff under delegation, under centralisation (with $c = 0$), and in the NHI and FB benchmarks: (i) when $p = \frac{1}{4}$, (ii) when $p = \frac{1}{2}$, and (iii) when $p = \frac{3}{4}$. The simulation assumes $R = \frac{1}{2}$.

It remains to be established when precisely delegation is preferable to centralisation. A comparison of the principal's maximised expected payoff under centralisation, $\Pi^*_{\mathcal{C}}$, with that under delegation, $\Pi^*_{\mathcal{D}}$, yields the following result:

Proposition 1.4. *The principal prefers to delegate the choice of the exam to the teacher rather than to centralise it if and only if the probability that the teacher is bad satisfies* $p \leq p_{\mathcal{D}}(c, \gamma, R)$*, where* $p_{\mathcal{D}} \in \left[\frac{1}{2}, 1\right]$ *is increasing in* c*, and decreasing in* γ *and* R*.*

Proposition 1.4 demonstrates that delegating the choice of exam to the teacher is preferable to centralisation as long as the teacher is sufficiently likely to be good. When the fixed cost of centralisation is zero, the principal's choice between delegation and

centralisation is entirely driven by part (i) of Corollary 1.3 and the relevant threshold is $p_{\mathcal{D}} = \frac{1}{2}$, i.e. the principal strictly prefers to delegate if and only if the teacher is more likely to be good than bad. As centralisation becomes more costly, the principal prefers delegation also for values of p higher than $\frac{1}{2}$. Eventually, when c is sufficiently large, it is optimal for the principal to delegate regardless of the values of the other parameters.

Other things being constant, the threshold p_D is decreasing in the ease of gaming, γ. This follows from the fact that centralisation can be optimal here only if $p \geq \frac{1}{2}$, in which case the principal's maximised expected payoff under delegation, $\Pi^*_{\mathcal{D}}$, is decreasing in γ (by part (i) of Corollary 1.3). This means that the cost of delegation relative to centralisation increases as gaming becomes easier, and therefore the latter becomes more likely to be optimal.

A similar argument applies to the comparative statics of $p_{\mathcal{D}}$ with respect to the principal's reward from the students obtaining deeper knowledge, R. When $p > \frac{1}{2}$ and $\gamma > 0$, $\Pi^*_{\mathcal{D}}(0)$ is strictly higher than $\Pi^*_{\mathcal{D}}(\gamma)$ (by part (i) of Corollary 1.3), and the difference is larger the higher R is. Since $\Pi^*_{\mathcal{C}} = \Pi^*_{\mathcal{D}}(0) - c$, this means that centralisation becomes more likely to be optimal as R increases. Since an increase in R means also that the incentives provided in the optimal contract unambiguously increase, this analysis confirms that delegation and incentives are substitutes when $p > \frac{1}{2}$.

Figure 1.5 illustrates the threshold $p_{\mathcal{D}}$ for different values of the cost of centralisation, c. The threshold is represented by a line that divides the (p, γ) parameter space into two regions: delegation is optimal in the region to the left of the line whereas centralisation is optimal in the region to the right of the line.

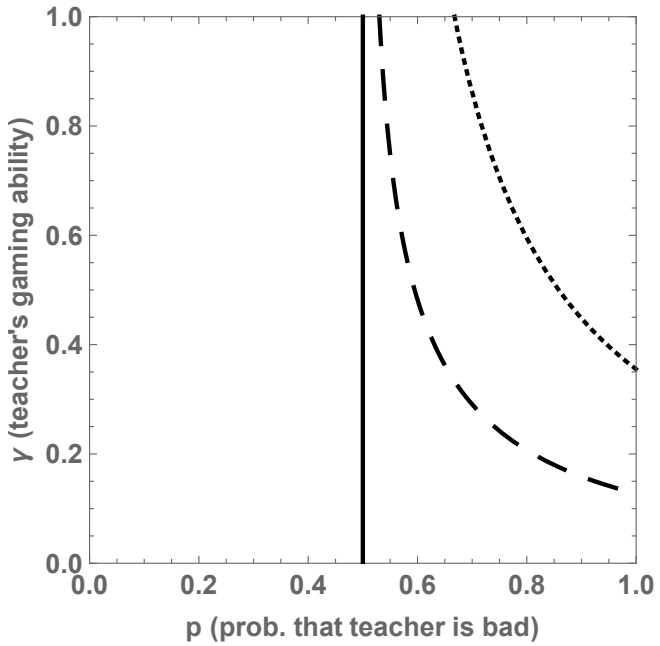


Figure 1.5: An illustration of the threshold $p_{\mathcal{D}}$ which determines the principal's choice between delegation ($p \leq p_{\mathcal{D}}$) and centralisation ($p \geq p_{\mathcal{D}}$): (i) when $c = 0$ (solid line), (ii) when c is strictly positive but low (dashed line), and (iii) when c is high (dotted line). When c is sufficiently high, delegation is preferable for all values of p.

1.6.2 Can Delegation Be Pareto Optimal?

In Proposition 1.4, we established the circumstances under which the principal prefers to delegate the choice of exam to the teacher, and we know from Section 1.4 that the bad teacher receives zero rent from the optimal contracts under delegation and under centralisation. Thus, to determine whether the delegation can be Pareto optimal in our setting, we need to find out whether—and possibly when—it can make the good teacher better off.

I will denote by $\mathcal{R}_{\mathcal{D}}$ the good teacher's expected rent from the optimal contract under delegation. It follows from Section 1.4 that $\mathcal{R}_{\mathcal{D}} = \frac{1}{4}(1-\gamma^2)\left(w_b^{\mathcal{D}}\right)^2$. I start the analysis with the following lemma:

Lemma 1.3. *The good teacher's expected rent from the optimal contract under delega-*

tion, $\mathcal{R}_{\mathcal{D}} = \frac{1}{4}\left(1-\gamma^2\right)\left(w_b^{\mathcal{P}}\right)^2$*, is strictly increasing (decreasing) in the ease of gaming,* γ*, if and only if* $\gamma < (>)\, \widetilde{\gamma}_{\mathcal{R}}$*, where* $\widetilde{\gamma}_{\mathcal{R}} = \sqrt{\frac{1-4p}{1-2p}}$.

It can be easily shown that the threshold $\widetilde{\gamma}_{\mathcal{R}}$ (i) is strictly decreasing in the probability that the teacher is bad, p, (ii) approaches 1 as p gets arbitrarily close to zero, and (iii) equals zero when $p = \frac{1}{4}$.

Interestingly, Lemma 1.3 implies that easier gaming may boost the good teacher's expected rent from the optimal contract under delegation. In particular, this will occur when the ease of gaming, γ, and the probability that the teacher is bad, p, are both sufficiently low.

An increase in γ may lead to an increase in $\mathcal{R}_{\mathcal{D}}$ because easier gaming has two distinct effects on the good teacher's expected rent. First, as was discussed in Section 1.5, the higher γ is, the more the principal can decrease the fixed compensation in the contract designed for the bad teacher, and hence the lower the expected rent of the good teacher is for any given incentives w_b provided in the contract designed for the bad teacher. Second, as was also discussed in Section 1.5, the fact the good teacher's rent is reduced for any given w_b may encourage the principal to offer the bad type a more high-powered contract—even though a higher w_b also exacerbates gaming. As we have seen in Corollary 1.2, the optimal incentives provided to the bad teacher are indeed strictly increasing in γ when $p < \frac{1}{2}$, and they are strictly decreasing in γ when $p > \frac{1}{2}$.

Thus, when $p < \frac{1}{2}$, the first and the second effect operate in opposite directions. Lemma 1.3 tells us that the higher p is, the lower γ must be for the latter effect to dominate over the former. In other words, if the teacher is sufficiently unlikely to be bad and gaming is difficult enough, we obtain the surprising result that the good teacher's expected rent from the optimal contract under delegation is *increasing* in the ease of gaming.

Now that we have identified how the ease of gaming affects the good teacher's expected rent, we can specify the circumstances under which the good teacher is better off under delegation than under centralisation.

Proposition 1.5. *The good teacher's expected rent from the optimal contract under delegation,* $\mathcal{R}_{\mathcal{D}} = \frac{1}{4}(1-\gamma^2)\left(w_b^{\mathcal{D}}\right)^2$*, is strictly higher (lower) than that from the optimal contract under centralisation,* $\mathcal{R}_{\mathcal{C}} = \frac{1}{4}\left(w_b^{\mathcal{C}}\right)^2$*, if and only if* $\gamma < (>)\,\gamma_{\mathcal{R}}$*, where* $\gamma_{\mathcal{R}} = \frac{\sqrt{1-4p}}{1-2p}$.

Similarly to Lemma 1.3, it can be easily shown that the threshold $\gamma_{\mathcal{R}}$ (i) is strictly decreasing in the probability that the teacher is bad, p, (ii) approaches 1 as p gets arbitrarily close to zero, and (iii) equals zero when $p = \frac{1}{4}$. Equivalently, the condition $\gamma < (>)\,\gamma_{\mathcal{R}}$ can be rewritten as $p < (>)\,p_{\mathcal{R}}$, where $p_{\mathcal{R}} = \frac{\sqrt{1-\gamma^2}-\left(1-\gamma^2\right)}{2\gamma^2}$. The threshold $p_{\mathcal{R}}$ has the following properties: (i) it is decreasing in the ease of gaming, γ, (ii) it approaches zero as γ gets arbitrarily close to 1, and (iii) it equals $\frac{1}{4}$ when $\gamma = 0$.

Proposition 1.5 thus states that the good teacher's expected rent from the optimal contract under delegation is higher than the one under centralisation as long as teachers are sufficiently unlikely to be bad and gaming is difficult enough.

The similarity of the statement in Proposition 1.5 to that in Lemma 1.3 follows from the fact that the good teacher's expected rent from the optimal contract under centralisation is $\mathcal{R}_{\mathcal{C}} = \frac{1}{4}\left(w_b^{\mathcal{C}}\right)^2$, which can be obtained by substituting $\gamma = 0$ into $\mathcal{R}_{\mathcal{D}}$. The mechanisms which lead to Proposition 1.5 are thus the same as the ones outlined in the discussion of Lemma 1.3. More generally, Proposition 1.5 implies that the good teacher may indeed benefit from the principal delegating the choice of exam to the teacher, even though such delegation prompts the bad teacher to engage in gaming—which the good teacher does not directly benefit from—and allows the principal to offer a menu of contracts that screen the two types of teacher more effectively—which means that the good teacher receives a lower rent for any given incentives provided to the bad

teacher.

Finally, the results in Propositions 1.4 and 1.5 allow us to identify the circumstances under which delegation or centralisation are Pareto optimal. Note here that, since the bad teacher receives zero rent from the optimal contract under delegation and under centralisation, we only need to consider the principal's expected payoff and the good teacher's expected rent.

Corollary 1.4. *Delegation is Pareto optimal when $p \leq p_{\mathcal{R}}$; centralisation is Pareto optimal when $p \geq p_{\mathcal{D}}$; when $p_{\mathcal{R}} < p < p_{\mathcal{D}}$, neither delegation nor centralisation is Pareto optimal.*

It follows from Propositions 1.4 and 1.5 that the principal and the teacher have aligned preferences regarding delegation and centralisation when $p \leq p_{\mathcal{R}}$ and when $p \geq p_{\mathcal{D}}$. When $p \leq p_{\mathcal{R}}$, they principal and the good teacher both prefer delegation to centralisation: $\Pi^*_{\mathcal{D}} > \Pi^*_C$ and $\mathcal{R}_{\mathcal{D}} \geq \mathcal{R}_{\mathcal{C}}$ (and $\mathcal{R}_{\mathcal{D}} > \mathcal{R}_{\mathcal{C}}$ whenever $p < p_{\mathcal{R}}$). When $p \geq p_{\mathcal{D}}$, they both prefer centralisation: $\Pi^*_{\mathcal{D}} \leq \Pi^*_C$ and $\mathcal{R}_{\mathcal{D}} \leq \mathcal{R}_{\mathcal{C}}$ (and $\Pi^*_{\mathcal{D}} < \Pi^*_C$ whenever $p > p_{\mathcal{D}}$). When $p_{\mathcal{R}} < p < p_{\mathcal{D}}$, the principal's maximised expected payoff is strictly higher when she delegates the choice of exam to the teacher than when she centralises it, whereas the good teacher's expected rent is strictly higher under centralisation than under delegation. Figure 1.6 illustrates these results.

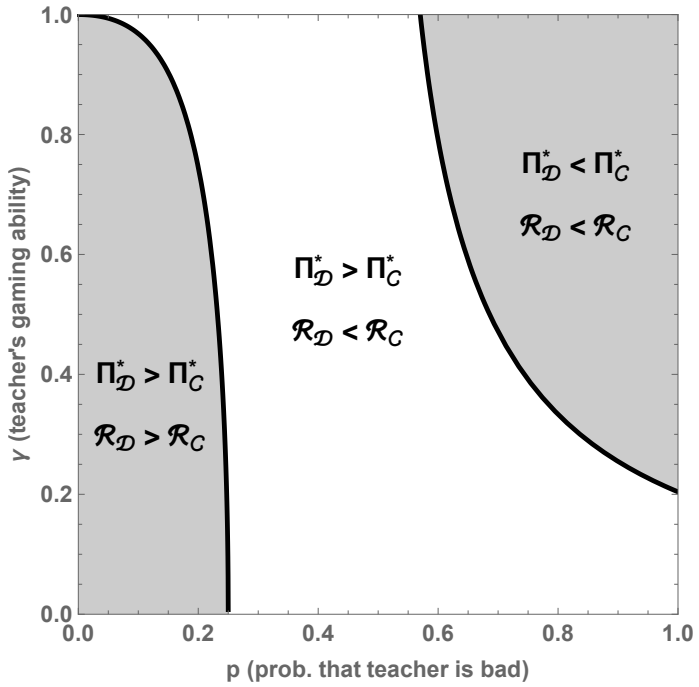


Figure 1.6: An illustration of the regions in the (p, γ) parameter space where the principal's and the teacher's preferences regarding delegation and centralisation are aligned (the grey areas) and where they are not (the white area). The simulation assumes $R = \frac{1}{2}$ and $c = \frac{1}{40}$.

1.7 Further Discussion

1.7.1 The Ease of Gaming Is $\gamma = 1$

The main model assumes that the ease of gaming is $\gamma < 1$, which ensures that, under delegation, the bad teacher's expected utility from any given contract (f, w) is always smaller than that of the good type. This no longer holds when $\gamma = 1$, in which case both types of teacher achieve expected utility of $f + \frac{1}{2}w^2$. Thus, whatever menu of contracts of the form (f, w) the principal would offer, both types of teachers choose the same compensation scheme—the one that maximises $f + \frac{1}{2}w^2$. In other words, the

principal can then no longer use menus of contracts as a screening device.[9]

Hence, the principal must resort to offering a single compensation scheme. Given a contract (f, w), the effort allocation of a bad teacher is $(e_A, e_B) = \left(\frac{1}{2}w, \frac{1}{2}w\right)$, which implies that the probability that students obtain deeper knowledge is $\Pr[\mathcal{K} = 1] = \frac{1}{2}w$, while the probability that they pass the exam is $\Pr[\mathcal{E} = 1] = w$. The effort allocation of a good teacher is $(e_A, e_B) = (0, w)$, which implies $\Pr[\mathcal{K} = 1] = \Pr[\mathcal{E} = 1] = w$.[10] The principal's maximisation problem is then

$$\begin{aligned} &\max_{f,w} p\left[\frac{1}{2}wR - f - w^2\right] + (1-p)\left[wR - f - w^2\right] \text{ s.t.} \\ &\text{(IRb, IRg) } f + \frac{1}{2}w^2 \geq 0. \end{aligned} \tag{1.20}$$

It is optimal here for the principal to make the individual rationality constraint of both types bind, i.e. $f = -\frac{1}{2}w^2$. Given this, the maximisation problem yields $w^{\mathcal{D}}_{\gamma=1} = R\left(1 - \frac{1}{2}p\right)$. We thus obtain the following proposition:

Proposition 1.6. *When $\gamma = 1$, the principal cannot use menus of contracts as a screening device, and the optimal contract under delegation consists of a single compensation scheme*

$$\left(f^{\mathcal{D}}_{\gamma=1}, w^{\mathcal{D}}_{\gamma=1}\right) = \left(-\frac{1}{2}R^2\left(1 - \frac{1}{2}p\right)^2, R\left(1 - \frac{1}{2}p\right)\right). \tag{1.21}$$

The principal's expected payoff is

$$\Pi^*_{\mathcal{D},\gamma=1} = \frac{1}{2}\left(w^{\mathcal{D}}_{\gamma=1}\right)^2. \tag{1.22}$$

[9]When $\gamma < 1$, the good teacher is indifferent between $\left(f^{\mathcal{D}}_b, w^{\mathcal{D}}_b\right)$ and $\left(f^{\mathcal{D}}_g, w^{\mathcal{D}}_g\right)$, but the bad teacher strictly prefers the former, so the principal could slightly "improve" the latter in order to induce a strict preference of the good teacher. However, when $\gamma = 1$, this no longer works because both types of teacher are indifferent between $\left(f^{\mathcal{D}}_b, w^{\mathcal{D}}_b\right)$ and $\left(f^{\mathcal{D}}_g, w^{\mathcal{D}}_g\right)$.

[10]When $\gamma = 1$, a good teacher is in fact indifferent between all effort choices that satisfy $e_A + e_B = w$, but I assume here that he then allocates all his effort to the task that is valued by the principal, i.e. task B.

Proposition 1.6 implies that the optimal incentive scheme becomes less high-powered as the probability that the teacher is bad increases. Furthermore, the principal's expected payoff is strictly decreasing in p, which is in line with common intuition.

It is worth noting that the principal's expected payoff jumps down discontinuously as $\gamma \to 1$: $\Pi^*_{\mathcal{D},\gamma=1} < \lim_{\gamma\to 1} \Pi^*_{\mathcal{D}}$. This is because screening is no longer possible when $\gamma = 1$, and therefore delegation does not bring the benefit of alleviating the problem of hidden information (whereas $\lim_{\gamma\to 1} \Pi^*_{\mathcal{D}} = \lim_{\gamma\to 1} \Pi^*_{NHI}$). The only benefit of delegation relative to centralisation is now that it does not entail the fixed cost, c.

The following proposition states results that are parallel to Propositions 1.3 and 1.4 for the case where $\gamma = 1$:

Proposition 1.7. *Suppose that $\gamma = 1$ and the teacher is privately informed about his type:*

(i) Relative to centralisation, delegating the choice of exam to the teacher always results in the optimal compensation scheme being strictly less high-powered for at least one of the types of teacher;

(ii) The principal prefers to delegate the choice of the exam to the teacher rather than to centralise it if and only if $p \leq p_{\mathcal{D},\gamma=1}(c, R)$, where $p_{\mathcal{D},\gamma=1}$ is increasing in c, decreasing in R, and $p_{\mathcal{D},\gamma=1} < \lim_{\gamma\to 1} p_{\mathcal{D}}$.

Part (i) of Proposition 1.7 is different from the setting with $\gamma < 1$, where—as we have seen in Proposition 1.3—delegation resulted in the optimal contract designed for the bad teacher being more high-powered when $p < \frac{1}{2}$, while the contract designed for the good teacher remained unchanged. Here, delegation always results in strictly lower incentives being offered in the contract designed for the good teacher: $w^{\mathcal{D}}_{\gamma=1} = R\left(1 - \frac{1}{2}p\right) < w^{\mathcal{C}}_g = R$ holds for all p.

However, it is worth noting that as $\gamma \to 1$, the mean power of incentives in the menu of $\left(f^{\mathcal{D}}_b, w^{\mathcal{D}}_b\right)$ and $\left(f^{\mathcal{D}}_g, w^{\mathcal{D}}_g\right)$ approaches the power of incentives in contract $\left(f^{\mathcal{D}}_{\gamma=1}, w^{\mathcal{D}}_{\gamma=1}\right)$:

$\lim_{\gamma\to 1}\overline{w}^{\mathcal{D}} = w^{\mathcal{D}}_{\gamma=1}$.[11] Hence, when $\gamma = 1$, it remains true that as long as $p < \frac{1}{2}$, the mean power of incentives in the optimal contract is higher under delegation than under centralisation.

The statement in part (ii) of Proposition 1.7 follows from the fact that screening is no longer possible when the ease of gaming is $\gamma = 1$ and therefore the principal's expected payoff under delegation jumps down discontinuously from $\Pi^*_{\mathcal{D}}$ to $\Pi^*_{\mathcal{D},\gamma=1}$ as $\gamma \to 1$. As a result, we also observe a discontinuous fall in the threshold value of p below which delegation is preferable to centralisation: $p_{\mathcal{D},\gamma=1} < \lim_{\gamma\to 1} p_{\mathcal{D}}$. In particular, when $c = 0$, we have $p_{\mathcal{D}} = \frac{1}{2}$ (for all values of γ) and $p_{\mathcal{D},\gamma=1} = 0$.

1.7.2 Agent Has Limited Liability

When the teacher has limited liability in the sense that any contract of the form (f, w) must satisfy $f \geq 0$, the results in Corollaries 1.2 and 1.3 no longer hold:

Proposition 1.8. *When the teacher has limited liability with $f \geq 0$, the incentives provided to both types of teacher by the optimal contract under delegation are decreasing in the ease of gaming, γ, for all p, and so is the principal's expected payoff from the optimal contract under delegation.*

Under limited liability with $f \geq 0$, an increase in the ease of gaming, γ, no longer allows the principal to decrease the fixed part of the compensation designed for the bad teacher. As a result, the good teacher's self-selection constraint is not relaxed and therefore his rent is not reduced. The presence of gaming under delegation has, therefore, an unambiguously negative effect on the principal's expected payoff: it only

[11] Thus, since $\lim_{\gamma\to 1}\overline{w}^{\mathcal{D}}$ is equal to the mean power of incentives provided by the optimal contract in the NHI benchmark, $\overline{w}^{NHI}$, we also have $w^{\mathcal{D}}_{\gamma=1} = \lim_{\gamma\to 1}\overline{w}^{NHI}$. But $\lim_{\gamma\to 1} w_b^{NHI} = \frac{R}{2} < w^{\mathcal{D}}_{\gamma=1}$ and $\lim_{\gamma\to 1} w_g^{NHI} = R > w^{\mathcal{D}}_{\gamma=1}$, which means that—relative to the NHI benchmark—contract $\left(f^{\mathcal{D}}_{\gamma=1}, w^{\mathcal{D}}_{\gamma=1}\right)$ provides too strong incentives to the bad teacher and too weak incentives to the good teacher. This is another way of explaining the discontinuity in the principal's maximum expected payoff under delegation as $\gamma \to 1$.

serves to increase the bad teacher's incentives to exert effort on task A, i.e. on "teaching to the test", which boosts his expected compensation but does not bring any benefit to the principal.[12] In other words, since centralisation ensures that there is no gaming, the only advantage of delegation over centralisation is that no monitoring costs are incurred.

More generally, when the teacher has limited liability with $f \geq -k$, the implications for the benefits from delegation depend on the magnitude of k. For sufficiently high values of k, the limited liability constraint is never binding and hence delegation brings the same benefit of alleviating the problem of hidden information as when there is no limited liability. For intermediate values of k, the benefit from delegation is only partial: delegation allows the principal to decrease the fixed part of the compensation designed for the bad teacher but only so long as the limited liability constraint does not bind. Finally, for sufficiently low values of k, the limited liability is always binding as in Proposition 1.8 and there is no benefit from delegation relative to centralisation other than that it does not involve a monitoring cost.

1.8 Conclusion

Most theoretical and empirical research supports the view that delegation and provision of incentives are complements. At the same time, the common perception is that one should not delegate the choice of performance measure to an employee whose contract is performance-based. This paper addresses a potential reason for this skepticism, which is that—when given discretion—employees would select measures that can be improved by performing easier tasks which do not necessary bring any benefit to the company.

In the model, I provide a rationale for giving discretion about performance measures

[12] In the appendix, in the proof of Proposition 1.8, I formally establish the optimal contract for the teacher when he has limited liability.

to agents who receive performance pay. The rationale applies to environments where: (i) the agents are privately informed about the degree of substitutability of their effort on different tasks (which affects how likely they are to engage in gaming), and (ii) the principal can use menus of contracts as a screening device.

Allowing agents to privately choose their own performance measures may be optimal because it leads in fact to two distinct effects. First, delegation prompts some agents to engage in actions that improve their apparent performance (and hence increase their compensation) but do not bring direct benefits to the organisation. This is costly to the principal and, since stronger incentives would only encourage more gaming, makes the compensation incentive scheme lower-powered. But delegation also has a positive side-effect: it allows the principal to reduce the fixed part of the compensation designed for "bad" agents who do not perceive efforts on different tasks as substitutes and hence engage in gaming. This in turn means that, for any given incentives provided to the "bad" agents, the principal can offer a lower rent in the contract designed for the "good" agents who perceive efforts on different tasks as perfect substitutes and therefore are not prone to gaming. Delegation thus alleviates the problem of hidden information and encourages the principal to offer a higher-powered compensation scheme.

If the distribution of agents is such that they are more likely to be "bad" than "good", we obtain the *a priori* more natural results that (i) delegating the choice of performance measure should be accompanied by provision of weaker incentives in performance contracts, and (ii) centralisation yields a higher expected payoff to the principal than delegation as long as the monitoring that it involves is sufficiently cheap. However, when the agent is more likely to be "good", then the positive effect of alleviating hidden information dominates over the negative effect of gaming, resulting in (i) delegation and the provision of incentives being complements, and (ii) delegation being preferable to centralisation even when monitoring is costless.

The setting of the model—albeit stylised—is designed to make the effects under consideration as clear and as prominent as possible. First, the cost of effort functions of the two types of agent have been specifically designed to isolate the role of substitutability of effort on the two tasks. Second, I have assumed that only one of the tasks is valuable to the principal. Nevertheless, the result that delegating the choice of performance measure to agents can lead to more effective screening should hold more generally—especially in environments where agents are privately informed about their inclination for gaming and those agents who receive rents under screening are unlikely to be the ones who choose performance measures for their own self-interest.

Appendix to Chapter 1

A.1 An Alternative to Centralisation: Hiring an External Examiner

In some scenarios, a principal may be unable to use the centralisation option, for example, she may be unable to set the exam herself. But she may still be able to hire an external examiner to do this. Here I consider a setup where the principal may hire a risk-neutral external examiner whose job is to set an exam (and who has no other influence on the students' final results). By doing so, the principal can separate the decision on the type of exam from the allocation of teaching efforts. The external examiner can either (i) ask the teacher to supply her with an exam, which is costless for her and is equivalent to the teacher setting the exam himself, or (ii) set her own difficult exam at cost $c \geq 0$.

I assume that the external examiner's contract—like the one for the teacher—can only be contingent on whether students pass the exam ($\mathcal{E} = 1$ or $\mathcal{E} = 0$) and not on

whether they learn deeper knowledge ($\mathcal{K} = 1$ or $\mathcal{K} = 0$). The external examiner's contract consists of a fixed compensation, f_e, and a variable compensation, w_e. If the students fail the exam ($\mathcal{E} = 0$), the external examiner receives a lump-sum payment, f_e, while if they pass the exam ($\mathcal{E} = 1$), she receives a compensation of $f_e + w_e$. Importantly, I also assume that the contract offers which the principal makes to the teacher and to an external examiner are simultaneous and public.[13]

I analyse two scenarios: (i) the examiner is not informed about the teacher's type and cannot be bribed by him (Section A.1.1), (ii) the examiner is informed about the teacher's type and can be bribed by him (Section A.1.2).[14]

The timing of the game is then as follows:

1. The principal (P) simultaneously and publicly offers a menu of contracts to the teacher (T) and a contract to the external examiner (E).

2. T and E simultaneously and publicly choose whether to accept or reject.

3. T chooses one contract from the menu (if he has accepted it), and this choice is either not observed (Section A.1.1) or observed (Section A.1.2) by E.

4. T chooses whether to offer E a bribe (only in Section A.1.2).

5. E chooses whether to set her own exam or to ask T to supply her with his exam (if she has accepted her contract).

[13]Otherwise, contracting externalities along the lines of Segal (1999) would arise. For example, suppose that the offers were simultaneous but not public. Then, given a contract offered to the external examiner, the principal would have an incentive to deviate by offering a different contract to the teacher that takes into account the payment to the examiner contingent on students passing the exam. The principal would have a similar incentive to deviate if the contract offers were made publicly, but she offered the menu of contracts to the teacher after the external examiner has signed the contract with the principal. The inability to commit to a contracting scheme would make the principal worse off.

[14]The examiner knows that only a bad teacher would ever bribe him. When discussing the second scenario, I also briefly discuss the setting where the examiner is informed about the teacher's type but bribing is not possible.

6. T observes E's choice (and chooses the type of exam if E has not done so) and then chooses his efforts on teaching.

7. The teaching outcome ($\mathcal{K} = 0$ or $\mathcal{K} = 1$), the exam outcome ($\mathcal{E} = 0$ or $\mathcal{E} = 1$), and the payoffs of P, T, and E are realised.

Since the expected cost of inducing the external examiner to set a difficult exam is c, it is optimal for the principal to hire her if and only if centralisation is optimal, with the appropriate condition having been identified in Proposition 1.4. Thus, the analysis here focuses only on the setting where centralisation is optimal.

To start with, assume that the external examiner has agreed to a contract that induces her to set a difficult exam. The optimal contract for the teacher is then a menu of $\left(f_b^{\mathcal{C}}, w_b^{\mathcal{C}}\right)$ and $\left(f_g^{\mathcal{C}}, w_g^{\mathcal{C}}\right)$. Given this optimal contract and the fact that a difficult exam is set, the optimal effort choices of a bad teacher are $(e_A, e_B) = \left(0, \frac{1}{2}pR\right)$, while those of a good teacher are $(e_A, e_B) = (0, R)$.

A.1.1 Bribing Is Not Allowed

Suppose that (i) the examiner does not know the teacher's type, and (ii) the teacher cannot bribe the external examiner so as to induce her to delegate the choice of the exam to the teacher. From the external examiner's perspective, the probability that students pass the exam is

$$\Pr\left[\mathcal{E} = 1 \mid diff.exam\right] = \frac{1}{2}p^2R + (1-p)\,R \tag{1.23}$$

if she sets a difficult exam. On the other hand, if she delegates the choice of the exam to the teacher, then a bad teacher sets an easy exam and his effort allocation becomes $(e_A, e_B) = \left(\frac{1}{2}\gamma pR, \frac{1}{2}pR\right)$. As a result, the probability that students pass the exam

increases to

$$\Pr\left[\mathcal{E}=1 \mid easy\, exam\right]=p\left[\frac{1}{2}\gamma pR+\frac{1}{2}pR\right]+(1-p)\,R. \tag{1.24}$$

Thus, given a contract (f_e, w_e), the external examiner's expected utility from setting a difficult exam is

$$U_{diff}^{ext,nb}\left(f_e, w_e\right)=f_e+\left[\frac{1}{2}p^2R+(1-p)\,R\right]w_e-c, \tag{1.25}$$

while if she sets an easy exam, her expected utility is

$$U_{easy}^{ext,nb}\left(f_e, w_e\right)=f_e+\left[\frac{1}{2}\left(1+\gamma\right)p^2R+(1-p)\,R\right]w_e. \tag{1.26}$$

We thus obtain the following proposition:

Proposition 1.9. *When bribing is not possible and the principal wants to induce the external examiner to set a difficult exam, the optimal contract for the external examiner,* $(f_{e,nb}, w_{e,nb})$ *satisfies* $w_{e,nb}\leq-\frac{2c}{\gamma p^2R}$ *and* $f_{e,nb}=c-\left[\frac{1}{2}p^2R+(1-p)\,R\right]w_e$.

This result follows from the fact that the external examiner prefers to set a difficult exam if and only if $U_{diff}^{ext,nb}\left(f_{e,nb}, w_{e,nb}\right)\geq U_{easy}^{ext,nb}\left(f_{e,nb}, w_{e,nb}\right)$, which is satisfied as long as $w_e\leq-\frac{2c}{\gamma p^2R}$. The lump-sum compensation is constructed in such a way that the external examiner's individual rationality constraint binds. The main implication of Proposition 1.9 is that if the external examiner is to be induced to set a difficult exam, she needs to be punished whenever the students pass the exam.

Unsurprisingly, the required punishment is increasing in the external examiner's private cost of setting a difficult exam, c, which is in line with common intuition. Moreover, the punishment is decreasing in (i) the ease of gaming, γ, (ii) the probability that the teacher is bad, p, and (iii) the principal's reward for the students obtaining

deeper knowledge, R. This follows from the fact that the higher γ, p, and R are, the larger the difference is in the probability of student passing the exam between an easy and a difficult exam. Consequently, as γ, p, and R increase, we see a decrease in the required "per-unit" punishment that induces the external examiner to set a difficult exam.

A.1.2 Bribing Is Allowed

Suppose that (i) the examiner knows the teacher's type, and (ii) the teacher can bribe the external examiner so as to induce her to delegate the choice of the exam to the teacher. In order to induce the external examiner to set a difficult exam, the principal now needs to compensate her not only for the fact that students are more likely to pass the exam when it is easy, but also for the bribe she may receive from the teacher. The setting is also different from the one in Section A.1.1 in that the external examiner learns the teacher's type (she knows that only a bad teacher would have incentives to bribe her).

Given the optimal contract for the teacher under centralisation, i.e. the menu of $\left(f_b^{\mathcal{C}}, w_b^{\mathcal{C}}\right)$ and $\left(f_g^{\mathcal{C}}, w_g^{\mathcal{C}}\right)$, and the fact that the teacher is known to be bad, the probability that students pass the exam is

$$\Pr\left[\mathcal{E} = 1 \mid diff.exam,\, bad\right] = \frac{1}{2}pR \tag{1.27}$$

if the external examiner sets a difficult exam. On the other hand, if she delegates the choice of the exam to the teacher, then a bad teacher will set an easy exam and the probability that students pass the exam will be

$$\Pr\left[\mathcal{E} = 1 \mid easy\,exam,\, bad\right] = \frac{1}{2}\gamma pR + \frac{1}{2}pR. \tag{1.28}$$

Given a contract (f_b, w_b), a bad teacher's expected utility is $f_b + \frac{1}{4}w_b^2$ when a difficult exam is set, and $f_b + \frac{1}{4}(1+\gamma^2)w_b^2$ if an easy exam is set, and therefore the difference in the expected utility is $\frac{1}{4}\gamma^2 w_b^2$. Since the optimal contract for the bad teacher (by Proposition 1.2) specifies that $(f_b, w_b) = \left(-\frac{1}{4}p^2R^2, pR\right)$, the maximum bribe which a bad teacher is willing to pay the external examiner to induce her not to set a difficult exam is $\mathcal{B} = \frac{1}{4}\gamma^2 w_b^2 = \frac{1}{4}(\gamma pR)^2$.

Thus, given a contract (f_e, w_e) and the maximum bribe $\mathcal{B}$, the external examiner's expected utility from setting a difficult exam is

$$U_{diff}^{ext,b}(f_e, w_e) = f_e + \left[\frac{1}{2}pR\right] w_e - c, \tag{1.29}$$

while if she sets an easy exam, her expected utility is

$$U_{easy}^{ext,b}(f_e, w_e) = f_e + \left[\frac{1}{2}\gamma pR + \frac{1}{2}pR\right] w_e + \frac{1}{4}(\gamma pR)^2. \tag{1.30}$$

These observations lead to the following proposition:

Proposition 1.10. *When bribing is possible and the principal wants to induce the external examiner to set a difficult exam, the optimal contract for the external examiner* $(f_{e,b}, w_{e,b})$ *satisfies* $w_{e,b} \leq -\frac{2c}{\gamma pR} - \frac{1}{2}\gamma pR$ *and* $f_{e,b} = c - \frac{1}{2}pRw_e$.

The result follows from the fact that the external examiner prefers to set a difficult exam if and only if $U_{diff}^{ext,b}(f_{e,b}, w_{e,b}) \geq U_{easy}^{ext,b}(f_{e,b}, w_{e,b})$, which is satisfied as long as $w_{e,b} \leq -\frac{2c}{\gamma pR} - \frac{1}{2}\gamma pR$.[15] As in Section A.1.1, if the external examiner is to be induced to set a difficult exam, she needs to be punished sufficiently severely for the fact that students pass the exam. Moreover, the required punishment remains increasing in the examiner's private cost of setting an exam, c.

[15] If the external examiner were able to observe the teacher's type but could not bribe him, the optimal contract for the external examiner would satisfy $w_e \leq -\frac{2c}{\gamma pR}$ and $f_e = c - \frac{1}{2}pRw_e$.

However, an important difference from Section A.1.1 is that—because of the possibility of bribing—the external examiner now needs to be punished, i.e. $w_{e,b} < 0$, even if the external examiner's private cost of setting a difficult exam is $c = 0$. Moreover, the relationship between parameters γ, p, R and the required punishment is now more complex as there are two distinct effects. Both effects rely on the fact that the higher γ, p, and R are, the larger the difference is in the probability of student passing the exam between an easy and a difficult exam, $\Pr[\mathcal{E} = 1 \mid easy\, exam,\, bad] - \Pr[\mathcal{E} = 1 \mid diff.\, exam,\, bad]$.

The first effect is qualitatively equivalent to the one in the setting where bribing is not possible. Since students are more likely to pass if they are given an easy exam, the external examiner needs to be punished whenever students pass the exam. The higher the values of γ, p, and R, the lower the minimum "per-unit" punishment that induces her to set a difficult exam. The second effect is new and operates through the bad teacher's incentives to bribe the external examiner. Higher values of γ, p, and R mean that the bad teacher's expected benefit from an easy exam rather than a difficult one being set is higher. As a result, a bad teacher is willing to pay a larger bribe in order to convince the external examiner not to set a difficult exam. Overall, if c is sufficiently high, the first effect dominates and $w_{e,b}$ is decreasing in γ, p, and R. However, when c is sufficiently low, $w_{e,b}$ is U-shaped with respect to γ, p, and R.

The more general implication of the analysis in Section A.1 is twofold. First, if agents game performance measures whenever given discretion about them and it hurts the principal, it may be optimal for the principal to separate the choice of performance measure from the allocation of efforts to tasks by delegating the former to an external entity. Second, when the principal wants a high-quality performance measure to be set but cannot observe this quality, the external monitor's compensation should not be positively correlated with the agent's observed performance.

A.2 Omitted Proofs

Proof of Lemma 1.1

Substituting the binding IRg constraint into the objective function in (1.10), and maximising it with respect to w_g yields the optimal contract for the good teacher, which is $\left(f_g^{NHI}, w_g^{NHI}\right) = \left(-\frac{1}{2}R^2, R\right)$. Analogous steps for (1.11) yield the optimal contract for the bad teacher: $\left(f_b^{NHI}, w_b^{NHI}\right) = \left(-\frac{R^2}{4(1+\gamma^2)}, \frac{R}{1+\gamma^2}\right)$. Substituting these into the principal's expected payoff, we obtain

$$\Pi_{NHI}^{*} = \frac{1}{2}R\left[\frac{pR}{2\left(1+\gamma^2\right)} + (1-p)\,R\right] = \frac{1}{2}R\left[\frac{1}{2}pw_b^{NHI} + (1-p)\,w_g^{NHI}\right]. \tag{1.31}$$

Proof of Lemma 1.2

In order to find the optimal contracts in the FB benchmark, we need to substitute $\gamma = 0$ into $\left(f_b^{NHI}, w_b^{NHI}\right)$ and $\left(f_g^{NHI}, w_g^{NHI}\right)$. This yields $\left(f_b^{FB}, w_b^{FB}\right) = \left(-\frac{1}{4}R^2, R\right)$ and $\left(f_g^{FB}, w_g^{FB}\right) = \left(-\frac{1}{2}R^2, R\right)$. Substituting these into the principal's expected payoff, we obtain

$$\Pi_{FB}^{*} = \frac{1}{2}R\left[\frac{1}{2}pR + (1-p)\,R\right] = \frac{1}{2}R\left[\frac{1}{2}pw_b^{FB} + (1-p)\,w_g^{FB}\right]. \tag{1.32}$$

Proof of Proposition 1.1

The proof follows from the main text. To confirm that the IRg constraint cannot bind in the optimal contract, note that a binding IRg constraint implies $f_g = -\frac{1}{2}w_g^2$. Substituting this into the SSg constraint, we obtain $f_b + \frac{1}{2}w_b^2 \leq 0$, which implies $f_b + \frac{1}{4}\left(1+\gamma^2\right)w_b^2 < 0$, i.e. the IRb constraint cannot be satisfied. Thus, under the contract discussed here, only the good type of teacher is attracted. The maximisation problem

becomes

$$\max_{w_g} (1-p)\left[w_g R - \frac{1}{2}w_g^2\right], \tag{1.33}$$

which yields $w_g^* = R = w_g^{NHI}$. It follows from the binding IRg constraint that $f_g^* = -\frac{1}{2}R^2 = f_g^{NHI}$. The good teacher is thus offered exactly the same compensation scheme as in the NHI benchmark. The principal's expected payoff is then $\Pi\left(f_g^*, w_g^*\right) = \frac{1}{2}R^2\left[1-p\right] = \frac{1}{2}R\left[\left(1-p\right)w_g^{\mathcal{D}}\right] < \Pi_{\mathcal{D}}^*$.

Proof of Proposition 1.2

The proof follows from the main text.

Proof of Corollary 1.1

Using the expressions for Π_{NHI}^* and $\Pi_{\mathcal{D}}^*$ from, respectively, Lemma 1.1 and Proposition 1.1, we obtain

$$\Pi_{NHI}^* - \Pi_{\mathcal{D}}^* = \frac{1}{4}R^2\left(\frac{1-\gamma^2}{1+\gamma^2}\right)\left(\frac{p\left(1-p\right)}{1-\gamma^2\left(1-2p\right)}\right). \tag{1.34}$$

We thus have

$$\begin{aligned}
\frac{\partial}{\partial p}\left[\Pi_{NHI}^* - \Pi_{\mathcal{D}}^*\right] &\overset{sgn}{=} \frac{\partial}{\partial p}\left[\frac{p\left(1-p\right)}{1-\gamma^2\left(1-2p\right)}\right] && (1.35)\\
&= \frac{-2\gamma^2p^2 - 2\left(1-\gamma^2\right)p + 1 - \gamma^2}{\left(1-\gamma^2\left(1-2p\right)\right)^2} && (1.36)\\
&\overset{sgn}{=} -2\gamma^2p^2 - 2\left(1-\gamma^2\right)p + 1 - \gamma^2. && (1.37)
\end{aligned}$$

As long as $\gamma > 0$, $G\left(p\right) := -2\gamma^2p^2 - 2\left(1-\gamma^2\right)p + 1 - \gamma^2$ is a concave quadratic function with respect to p. The roots of $G\left(p\right)$ are $p_1 = \frac{-\sqrt{1-\gamma^4}-\left(1-\gamma^2\right)}{2\gamma^2} < 0$ and $p_2 = \frac{\sqrt{1-\gamma^4}-\left(1-\gamma^2\right)}{2\gamma^2} > 0$. Given that $p \in (0,1)$, we thus have that $\frac{\partial}{\partial p}\left[\Pi_{NHI}^* - \Pi_{\mathcal{D}}^*\right] > (<)\,0$ if and only if $p < (>)\,\widehat{p}\left(\gamma\right)$, where $\widehat{p}\left(\gamma\right) = \frac{\sqrt{1-\gamma^4}-\left(1-\gamma^2\right)}{2\gamma^2}$. Furthermore, we have $\widehat{p}\,'\left(\gamma\right) =$

$\frac{\sqrt{1-\gamma^4}-1}{\gamma^3\sqrt{1-\gamma^4}} < 0$ and it can be checked trivially that $\lim_{\gamma\to 1}\widehat{p}\,(\gamma) = 0$. Finally, when $\gamma = 0$, we have $\Pi^*_{NHI} - \Pi^*_{\mathcal{D}} = \frac{1}{4}R^2 p\,(1-p)$, which is minimised by $p = \frac{1}{2}$. Also, $\widehat{p}\,(\gamma)$ is continuous as $\gamma \to 0$ since

$$\lim_{\gamma\to 0}\widehat{p}\,(\gamma) = \left[\frac{0}{0}\right] \overset{H}{=} \lim_{\gamma\to 0}\frac{\frac{-2\gamma^2}{\sqrt{1-\gamma^4}}+1}{2} = \frac{1}{2}, \tag{1.38}$$

where "H" denotes the use of L'Hôpital's rule.

Proof of Corollary 1.2

The statements in parts (i) and (ii) follow trivially by inspection of $w_b^{\mathcal{D}}$.

Proof of Proposition 1.3

The proof follows from the main text.

Proof of Corollary 1.3

The results in both parts follow from Corollary 1.2 and from the fact that (i) Π^*_{NHI} is linear in w_b^{NHI} and w_g^{NHI}, (ii) $\Pi^*_{\mathcal{D}}$ is linear in $w_b^{\mathcal{D}}$ and $w_g^{\mathcal{D}}$, and (iii) w_g^{NHI} and $w_g^{\mathcal{D}}$ are invariant to γ.

Proof of Proposition 1.4

First, note that since $\Pi^*_{\mathcal{D}}$ is increasing in γ when $p < \frac{1}{2}$, we can have $\Pi^*_{\mathcal{D}} - \Pi^*_{\mathcal{C}} \leq 0$ only if $p \geq \frac{1}{2}$. Using the expressions for $\Pi^*_{\mathcal{D}}$ and $\Pi^*_{\mathcal{C}}$ from Propositions 1.1 and 1.2, we obtain

$$\Pi^*_{\mathcal{D}} - \Pi^*_{\mathcal{C}} = \frac{1}{4}p^2R^2\left[\frac{\gamma^2\,(1-2p)}{1-\gamma^2\,(1-2p)}\right] + c. \tag{1.39}$$

We have

$$\frac{\partial}{\partial p}\left[\Pi_{\mathcal{D}}^{*}-\Pi_{\mathcal{C}}^{*}\right] = \frac{\gamma^{2}pR^{2}\left(1-\gamma^{2}\left(1-2p\right)^{2}-3p\right)}{2\left(1-\gamma^{2}\left(1-2p\right)\right)^{2}} \tag{1.40}$$

$$\overset{sgn}{=} \underbrace{1-3p}_{<0 \text{ since } p\geq\frac{1}{2}} \quad -\gamma^{2}\left(1-2p\right)^{2}<0, \tag{1.41}$$

which implies that there exists a critical value $p_{\mathcal{D}}\left(c,\gamma,R\right)$ such that the principal prefers to delegate if and only if $p\leq p_{\mathcal{D}}\left(c,\gamma,R\right)$, and otherwise prefers to centralise. The value of p_D is given by $F\left(p_{\mathcal{D}}\right)=\Pi_{\mathcal{D}}^{*}\left(p_{\mathcal{D}}\right)-\Pi_{\mathcal{C}}^{*}\left(p_{\mathcal{D}}\right)=0$.

We can use implicit differentiation to determine the comparative statics of p_D with respect to c, γ, and R. We know from (1.41) that $\frac{\partial F}{\partial p_{\mathcal{D}}}<0$. Clearly, $\frac{\partial F}{\partial c}>0$, and hence $\frac{\partial p_{\mathcal{D}}}{\partial c}=-\frac{\partial F/\partial c}{\partial F/\partial p_{\mathcal{D}}}>0$. Since $p\geq\frac{1}{2}$, we can also easily verify that $\frac{\partial F}{\partial R}\leq 0$, and therefore $\frac{\partial p_{\mathcal{D}}}{\partial R}=-\frac{\partial F/\partial R}{\partial F/\partial p_{\mathcal{D}}}\leq 0$. Finally, it can be easily checked that $\frac{\partial F}{\partial\gamma}=\frac{\frac{1}{2}\gamma p^{2}(1-2p)R^{2}}{\left(1-\gamma^{2}(1-2p)\right)^{2}}\leq 0$, where the inequality again follows from the fact, as argued above, it must be that $p\geq\frac{1}{2}$. Thus, we have $\frac{\partial p_{\mathcal{D}}}{\partial\gamma}=-\frac{\partial F/\partial\gamma}{\partial F/\partial p_{\mathcal{D}}}\leq 0$.

Proof of Lemma 1.3

Differentating $\mathcal{R}_{\mathcal{D}}=\frac{1}{4}\left(1-\gamma^{2}\right)\left(w_{b}^{\mathcal{D}}\right)^{2}$ with respect to γ, we obtain:

$$\frac{\partial\mathcal{R}_{\mathcal{D}}}{\partial\gamma}=\frac{\gamma p^{2}R^{2}\left(1-4p-\gamma^{2}\left(1-2p\right)\right)}{2\left(1-\gamma^{2}\left(1-2p\right)\right)^{3}}. \tag{1.42}$$

Since $\gamma\in[0,1)$ and $p\in(0,1)$ imply that the denominator is strictly positive, we have that $\partial\mathcal{R}_{\mathcal{D}}/\partial\gamma>0$ if and only if $1-4p-\gamma^{2}\left(1-2p\right)>0$, which simplifies to

$$\gamma^{2}\left(1-2p\right)<1-4p. \tag{1.43}$$

This can be rearranged to yield the condition $\gamma<\widetilde{\gamma}_{\mathcal{R}}$, where $\widetilde{\gamma}_{\mathcal{R}}=\sqrt{\frac{1-4p}{1-2p}}$. Note also that (1.43) cannot be satisfied when $p>\frac{1}{4}$. When $p\in\left(\frac{1}{2},\frac{1}{4}\right]$, the left-hand side of

(1.43) is positive while the right-hand side is strictly negative. When $p \in \left(\frac{1}{2}, 1\right)$, both sides of (1.43) are strictly negative; however, since $1 - 2p > 1 - 4p$ and $\gamma^2 < 1$, the left-hand side of (1.43) is then always strictly larger than the right-hand side.

The threshold $\widetilde{\gamma}_{\mathcal{R}}$ is strictly decreasing in p:

$$\frac{\partial \widetilde{\gamma}_{\mathcal{R}}}{\partial p} = -\frac{1}{\sqrt{1-4p}\,(1-2p)^{\frac{3}{2}}} < 0. \tag{1.44}$$

The remaining statements of Lemma 1.3 can be trivially checked by inspection.

Proof of Proposition 1.5

The good teacher's expected rent from the optimal contract under delegation is $\mathcal{R}_{\mathcal{D}} = \frac{1}{4}(1-\gamma^2)\left(w_b^{\mathcal{D}}\right)^2 = \frac{1}{4}(1-\gamma^2)\frac{p^2R^2}{(1-\gamma^2(1-2p))^2}$. The good teacher's expected rent from the optimal contract under centralisation is $\mathcal{R}_{\mathcal{C}} = \frac{1}{4}\left(w_b^{\mathcal{C}}\right)^2 = \frac{1}{4}p^2R^2$. The condition $\mathcal{R}_{\mathcal{D}} > \mathcal{R}_{\mathcal{C}}$ can be rearranged to $\gamma^2(1-2p)^2 < 1-4p$. Since the left-hand side of this inequality is positive, the condition can be satisfied only if $p \leq \frac{1}{4}$. We thus obtain that $\mathcal{R}_{\mathcal{D}} > \mathcal{R}_{\mathcal{C}}$ if and only if $\gamma < \gamma_{\mathcal{R}}$, where $\gamma_{\mathcal{R}} = \frac{\sqrt{1-4p}}{1-2p}$.

The threshold $\gamma_{\mathcal{R}}$ is strictly decreasing in p:

$$\frac{\partial \gamma_{\mathcal{R}}}{\partial p} = -\frac{4p}{\sqrt{1-4p}\,(1-2p)^2} < 0. \tag{1.45}$$

The remaining statements of Proposition 1.5 can be trivially checked by inspection. Note that $\gamma_{\mathcal{R}} \geq \widetilde{\gamma}_{\mathcal{R}}$, both $\gamma_{\mathcal{R}}$ and $\widetilde{\gamma}_{\mathcal{R}}$ approach 1 as $p \to 0$, and both are equal to zero when $p = \frac{1}{4}$.

Proof of Corollary 1.4

The result follows directly from Propositions 1.4 and 1.5.

Proof of Proposition 1.6

The proof follows from the main text.

Proof of Proposition 1.7

Part (i). First, note that $w^{\mathcal{D}}_{\gamma=1} = R\left(1 - \frac{1}{2}p\right) > (<)\, w^{\mathcal{C}}_{b} = pR$ if and only if $p < (>)\, \frac{2}{3}$. Second, observe that $w^{\mathcal{D}}_{\gamma=1} = R\left(1 - \frac{1}{2}p\right) < w^{\mathcal{C}}_{g} = R$ for all $p \in (0,1)$.

Part (ii). Using the expressions for $\Pi^{*}_{\mathcal{C}}$ and $\Pi^{*}_{\gamma=1}$ from, respectively, Propositions 1.2 and 1.6, we obtain

$$\Pi^{*}_{\gamma=1} - \Pi^{*}_{\mathcal{C}} = -\frac{1}{8}R^2p^2 + c. \tag{1.46}$$

Since $p_{\mathcal{D},\gamma=1}$ satisfies $\Pi^{*}_{\gamma=1} - \Pi^{*}_{\mathcal{C}} = 0$, we obtain $p_{\mathcal{D},\gamma=1} = \frac{2\sqrt{2c}}{R}$, which is clearly increasing in c and decreasing in R. Note also that $p_{\mathcal{D},\gamma=1} = 0$ when $c = 0$.

Furthermore, $\lim_{\gamma\to 1} p_{\mathcal{D}} > p_{\mathcal{D},\gamma=1}$ follows from the fact that

$$\lim_{\gamma\to 1} \Pi^{*}_{\mathcal{D}} = \frac{1}{2}R^2\left(1 - p + \frac{1}{4}p\right) > \Pi^{*}_{\gamma=1} = \frac{1}{2}R^2\left(1 - p + \frac{1}{4}p^2\right) \tag{1.47}$$

holds for all $p \in (0,1)$.

Proof of Proposition 1.8

When the teacher has limited liability, the principal's maximisation problem is:

$$\begin{aligned}
&\max_{f_b,w_b,f_g,w_g} p\left[\frac{1}{2}w_bR - f_b - \frac{1}{2}\left(1+\gamma^2\right)w_b^2\right] + (1-p)\left[w_gR - f_g - w_g^2\right] \text{ s.t.} \\
&\text{(IRb) } f_b + \frac{1}{4}\left(1+\gamma^2\right)w_b^2 \geq 0 \\
&\text{(IRg) } f_g + \frac{1}{2}w_g^2 \geq 0 \\
&\text{(SSb) } f_b + \frac{1}{4}\left(1+\gamma^2\right)w_b^2 \geq f_g + \frac{1}{4}\left(1+\gamma^2\right)w_g^2 \\
&\text{(SSg) } f_g + \frac{1}{2}w_g^2 \geq f_b + \frac{1}{2}w_b^2 \\
&\text{(LLb \& LLg) } f_b \geq 0,\ f_g \geq 0, \qquad (1.48)
\end{aligned}$$

where LLb and LLg denote the limited liability constraints for the two types of teacher.

First, note that we cannot have $f_b > 0$ and $f_g > 0$ in the optimal contract: the principal could then decrease both f_b and f_g by $\varepsilon > 0$ while still satisfying the IRb and IRg constraints and without affecting the SSb and SSg constraints. Thus, it must be that either $f_b = 0$ and $f_g \geq 0$ or $f_b \geq 0$ and $f_g = 0$.

1. Consider $f_b = 0$ and $f_g \geq 0$. Then the LLg and the SSb constraints imply that $0 \leq f_g \leq \frac{1}{4}\left(1+\gamma^2\right)w_b^2 - \frac{1}{4}\left(1+\gamma^2\right)w_g^2$, and hence that $w_b \geq w_g$. But the SSg and the SSb constraints imply that $\frac{1}{2}w_b^2 - \frac{1}{2}w_g^2 \leq f_g \leq \frac{1}{4}\left(1+\gamma^2\right)w_b^2 - \frac{1}{4}\left(1+\gamma^2\right)w_g^2$, and hence that $w_g \geq w_b$. Therefore, it must be that $w_b = w_g$. The SSb and LLg constraints then imply that $f_g = 0$.

2. Consider $f_b \geq 0$ and $f_g = 0$. Then the LLb and the SSg constraints imply that $0 \leq f_b \leq \frac{1}{2}w_g^2 - \frac{1}{2}w_b^2$, and hence that $w_g \geq w_b$. Furthermore, the SSb constraint implies $f_b \geq \frac{1}{4}\left(1+\gamma^2\right)w_g^2 - \frac{1}{4}\left(1+\gamma^2\right)w_b^2$. Whenever $f_b > \frac{1}{4}\left(1+\gamma^2\right)w_g^2 - \frac{1}{4}\left(1+\gamma^2\right)w_b^2$, the principal can decrease f_b by $\epsilon > 0$ while still satisfying the LLb, the SSb and the SSg constraints; therefore, the SSb constraint must bind in the optimal

contract. The principal's maximisation problem can be then rewritten as

$$\max_{w_b,w_g} p\left[\frac{1}{2}w_bR - \frac{1}{4}\left(1+\gamma^2\right)\left(w_g^2+w_b^2\right)\right] + (1-p)\left[w_gR - f_g - w_g^2\right] \text{ s.t.}$$
$$\text{(SSb \& LLb) } w_g \geq w_b. \tag{1.49}$$

First order conditions with respect to w_b and w_g yield, respectively, $w_b^* = \frac{R}{1+\gamma^2}$ and $w_g^* = \frac{(1-p)R}{2(1-p)+\frac{1}{2}p(1+\gamma^2)}$. But it is clear that $w_b^* > \frac{1}{2}R$ while $w_g^* < \frac{1}{2}R$, so the constraint that $w_g \geq w_b$ is violated and therefore the optimal contract must satisfy $w_b = w_g$. The SSg and LLb constraints then imply that $f_b = 0$.

With $f_b = f_g = 0$ and $w_b = w_g = w$, the principal's maximisation problem can be rewritten as

$$\max_w p\left[\frac{1}{2}wR - \frac{1}{2}\left(1+\gamma^2\right)w^2\right] + (1-p)\left[wR - w^2\right]. \tag{1.50}$$

The first order condition with respect to w yields $w^* = \frac{R\left(1-\frac{1}{2}p\right)}{2-p(1-\gamma^2)}$, while the principal's expected payoff from the optimal contract is $\Pi^* = \frac{1}{2}R^2\frac{\left(1-\frac{1}{2}p\right)^2}{2-p(1-\gamma^2)}$. Both w^* and Π^* are strictly decreasing in γ.

Proof of Proposition 1.9

The proof follows from the main text.

Proof of Proposition 1.10

The proof follows from the main text.

Chapter 2

What Drives Regimes to Manipulate Information: Criticism, Collective Action, and Coordination

2.1 Introduction

Authoritarian regimes often manipulate information in order to improve the citizens' perception of the ruling government and to prevent protests that could ultimately result in an overthrow. They have several instruments at their disposal, with the most common being censorship, direct media control, employment of social media commentators, as well as the use of political pressure and repressions to influence journalist reporting.[1] Out of sixty-five countries assessed in the "Freedom on the Net 2015" report by the Freedom House, thirty-one engage in censorship of political, social and religious content, twenty-four employ pro-government commentators who manipulate online discussions,

[1]The Freedom House publishes annual reports on media independence and internet freedom, "Freedom of the Press" and "Freedom on the Net", which document the ways in which governments influence the information available to citizens across the world.

and forty have arrested, imprisoned or put journalists into detention for political or social content.[2]

The scale of information manipulation is extensive in some countries, for example, King, Pan and Roberts (2017) estimate that there are as many as two million pro-regime online commentators employed in China. This large group of people came to be called the Fifty Cent Party as they have allegedly been paid 50 cents (5 jiao) for every post that advances the line of the Communist Party.[3] The Chinese regime also uses a sophisticated system of censorship, which involves both automated mechanisms and human monitors. In recent years, this apparatus was extensively used in the period surrounding the 25th anniversary of the Tiananmen Square Massacre and during the Umbrella Revolution in Hong Kong.[4]

Regimes also distort the information available to citizens through firm control of media. For example, the Russian state controls—either directly or through proxies—all the major national television networks, as well as national radio networks, important national newspapers, and national news agencies. In addition to this, the regime controls more than 60 percent of the country's estimated 45,000 regional and local newspapers and other periodicals.[5]

On a more general level, information manipulation may take different forms. One approach is to persuade the citizens who criticise the government and its policies that the situation is better than it really is. Alternatively, a regime may distract citizens so as to redirect their attention from discussions on unfavourable topics or events. King,

[2]Freedom House, "Freedom on the Net 2015: A Global Assessment of Internet and Digital Media," edited by S. Kelly, M. Earp, L. Reed, A. Shahbaz, and M. Truong; https://freedomhouse.org/report/freedom-net/freedom-net-2015

[3]Han (2015) provides a comprehensive list of tasks performed by online commentators. For similar recent developments in Russia, see, e.g.: http://www.nytimes.com/2015/06/07/magazine/the-agency.html; http://www.theguardian.com/world/2015/apr/02/putin-kremlin-inside-russian-troll-house

[4]See "Freedom on the Net" (2015).

[5]Freedom House, "Freedom of the Press 2016"; https://freedomhouse.org/report/freedom-press/2016/russia

Pan and Roberts (2017) provide empirical evidence on these two forms of information manipulation in the context of government-employed social media commentators in China. They show that, contrary to the common views on the subject, it is not part of the Chinese regime's strategy to engage in arguments with skeptics and to defend the party and the government. They instead observe that these paid internet trolls do not even discuss controversial issues and mostly write posts that involve cheerleading for Chinese state and the Communist Party. King et al. (2017) thus infer that the regime employs online commentators to distract the public and to change the subject of discussion when needed.

The choice of specific goals of information manipulation depends on what the authoritarian regime believes to be critical to staying in power. When analysing the Chinese censorship programme, King, Pan and Roberts (2013; 2014) have proposed two distinct theories of what constitutes the goals of the regime: the *state critique theory* and the *theory of collective action potential*. According to the former, censorship is aimed at restricting any criticism of the government and its policies, while the latter assumes that the objective is to stop the spread of information that could lead to any kind of collective action, regardless of whether or not the expression is in opposition to the state and whether or not it is related to the government's policies. In principle, it could be that either or both of the above theories are correct or incorrect. Remarkably, King et al. (2013; 2014) demonstrate empirical evidence suggesting that, in the context of China, the theory of collective action potential is correct while the state critique theory is not.

The aim of this paper is to formally analyse the incentives of authoritarian regimes to manipulate information and to highlight the role of the citizens' ability to coordinate a collective action. In the paper, I also attempt to provide *a possible* theoretical rationale for the empirical findings of King et al. (2013; 2014; 2017). The model uses the standard

global games framework—introduced by Carlsson and van Damme (1993)—in which a continuum of citizens observe private and heterogeneous signals about the incumbent regime's strength, and then simultaneously choose whether to attack it. If the aggregate attack is sufficiently large relative to the incumbent's strength, the regime change is successful. Citizens receive a positive payoff if they participate in a successful attack; however, this participation is costly. I assume that, before the citizens decide on their actions, the regime is able to manipulate the information available to them by adding noise to their private signals.

An important feature of the model—discussed in Section 2.3 of the paper—is that the regime's incentives to increase noise in the citizens' private information are strongest for moderate values of (i) the citizens' individual cost of attacking the regime, and (ii) the mean strength of the regime. Intuitively, when the regime adds noise to the citizens' private signals, they attach more weight to the prior information when deciding on whether to attack. Consequently, the regime has incentives to add noise *only if* the cost of attacking is sufficiently high or the regime is expected to be strong enough. On the other hand, as the cost of attacking and the mean strength become arbitrarily high, the regime is almost bound to survive and the marginal benefit from adding noise eventually approaches zero. Thus, in the model, if it is ever optimal for the regime to add noise, it is for some moderate values of the two parameters.

In Section 2.4, I consider two possible drivers of the regime's incentives to prevent regime change: (i) the desire to minimise the measure of citizens whose posteriors are below a critical threshold, and (ii) the desire to minimise the size of collective attacks. These two cases can be thought of as capturing the ideas behind, respectively, the state critique theory and the theory of collective action potential. The model predicts that whether the regime's optimal strategy of adding noise focuses on the former or the latter depends on how well informed the citizens are about the regime's strength.

In particular, the regime's incentives are driven predominantly by the former when the citizens are intrinsically well informed about the regime's strength, and by the latter when they are poorly informed about it. Thus, the Chinese regime's focus on preventing collective action rather than suppressing state criticism may be due to the fact that the Chinese citizens' knowledge of the regime is poor—even before the regime's manipulation efforts.

The model also suggests a possible linkage to the main empirical finding of King et al. (2017), which says that the Chinese regime employs social media commentators to distract citizens rather than to persuade them. If a regime wants to prevent an overthrow, information manipulation in the form of adding noise (which could capture distraction[6] as opposed to persuasion) turns out to be effective even if (i) the citizens can observe the regime's manipulative action (i.e. they observe how much noise is being added but not the realisation of the noise term), and (ii) the regime is not better informed about its own strength than the citizens are. On the other hand, as is discussed in more detail below, Edmond (2013) shows that information manipulation in the form of adding a bias (which could capture persuasion) is effective only if the two informational conditions above do not hold. Thus, information manipulation in the form of adding noise is effective under weaker informational requirements than adding a bias.

In Section 2.5, I investigate how the citizens' ability to coordinate their actions affects the regime's incentives to add noise to their private information. I show that a crucial role is played by how well the citizens can aggregate their collective information. In particular, if the citizens can better coordinate their actions only at the cost of impeded information aggregation, the regime's incentives to increase noise may be

[6]For example, a number of papers in the attention allocation literature assume a linear noise reduction technology, i.e. that the attention devoted to an information source is linearly devoted to the precision of the signal. For example, see Myatt and Wallace (2012), Mondria and Quintana-Domeque (2013), and Chen and Suen (2017).

even stronger than if the citizens were imperfectly coordinated but could still partially aggregate their information. As a result, better coordination may actually hurt the citizens and render regime change less likely.

Related Literature

The model draws from the extensive literature on global games, which was introduced by Carlsson and van Damme (1993) and further developed by Morris and Shin (1998; 2001; 2002).[7]

The paper contributes to the literature which investigates how the citizens' information affects the probability of regime change. A number of papers have studied the role of signal precision, e.g., Metz (2002), Bannier and Heinemann (2005), Iachan and Nenov (2015), and Szkup and Trevino (2015). My model also focuses on the role of signal precision, and adds to the literature by looking deeper into the impact of collective action and coordination on the regime's incentives to increase noise, which provides insights into the empirical findings of King, Pan and Roberts (2013; 2014; 2017).

My paper is also related to Edmond (2013) who analyses a coordination game in which a regime can add a bias, instead of noise, to the citizens' information. In his model, the regime is informed about its strength, and its manipulative action—in the form of a linear bias added to private signals—is unobserved by the citizens. This informational setting allows weaker regimes to mimic the strong ones; however, since information manipulation is costly, only the intermediate regimes do so. Crucially, the manipulation is not backed out by the citizens as they need to infer the regime's hidden action jointly with their inferences about the incumbent's strength. As a result, they cannot perfectly disentangle the regime's strength from the endogenous bias. As I show later, contrasting the implications of my model with those of Edmond (2013) creates

[7]Morris and Shin (2003) provide an excellent survey of the global games literature.

interesting linkages to King et al. (2013; 2014; 2017).

Little (2016) and Ananyev, Petrova and Zudenkova (2017) present theoretical models whose results are also related to King et al. (2013; 2014). Little (2016) distinguishes between "political coordination" (i.e. a sufficient number of citizens must protest to successfully overthrow) and "tactical coordination" (i.e. they must choose a similar tactic), and shows that reducing the precision of citizens' information about tactics is always an effective way to reduce protest while reducing the precision of the signal about the regime's popularity is only under certain conditions. Ananyev, Petrova and Zudenkova (2017) make a related distinction between "coordination censorship", i.e. one which makes it less likely that the citizens use the same tactic, and "content censorship", i.e. one which makes a signal of the regime's unpopularity more noisy. However, neither of these papers distinguishes between distraction and persuasion as methods of information manipulation, and thus neither provides insights into their relative merits or the circumstances under which they are likely to be used.

The paper is also related to the growing economic literature on censorship. Shadmehr and Bernhardt (2015) analyse the strategic choices of authoritarian regimes on when to censor the information available to citizens, and they show that bad rulers may prefer media to be strong. This is because when revolution payoffs are high, a ruler values strong media that might uncover good news about the status quo which could prevent a revolution. Although the mechanisms are quite different, one can see parallels to the result that the mean strength of the regime must be sufficiently high for the regime to have incentives to add noise. Lorentzen (2014) analyses the trade-off of an authoritarian regime in setting the level of censorship, where free media provide the information necessary for the regime to discipline lower-level bureaucrats, but—at the same time—they may facilitate a coordinated revolt if discontent is revealed to be widespread. His model provides insights into why an increase in uncontrollable infor-

mation may lead to a reduction in media freedom, for example, in China. The same trade-off is analysed by Egorov, Guriev and Sonin (2009) who show that providing bureaucrats with incentives is less important for authoritarian regimes in resource-rich economies and, therefore, media are more likely to be free in these countries. Guriev and Treisman (2015) investigate how information manipulation, e.g., censorship and state propaganda, may interact with other methods of preventing regime change such as co-opting the elite and equipping the police to repress attempted uprisings.

The idea of information manipulation appears also in the literature on media capture, for example, Besley and Prat (2006), Petrova (2008), Enikopolov, Petrova and Zhuravskaya (2011), and Gehlbach and Sonin (2014). Recent developments in social media have been studied by Enikopolov, Petrova and Sonin (2017), who provide empirical evidence that social media can discipline corruption even in countries with limited political competition. Enikopolov, Makarin and Petrova (2016) find that increased social media penetration in Russia has boosted both the probability of protest and the number of protesters. They also identify the main channel to be the reduction in the costs of coordination, rather than the spread of information critical of the government.

Finally, other political economy applications of global games can be found, among others, in Angeletos, Hellwig and Pavan (2006), Bueno de Mesquita (2010), Chassang and Padró-i-Miquel (2010), Shadmehr and Bernhardt (2011), Boix and Svolik (2013), and Casper and Tyson (2014).

2.2 The Model

2.2.1 Setup

Game of Regime Change. The model uses the global games framework of modelling regime change. There is a continuum of citizens of measure 1, indexed by i and

uniformly distributed over $[0, 1]$. Citizens simultaneously choose between two actions: they can either attack the regime (e.g., protest, join a revolt) or refrain from attacking. The citizens' payoffs from the two possible actions depend on whether the attack turns out to be successful:

	Attack is successful	Status quo remains
Attack	$1 - c$	$-c$
Do Not Attack	0	0

If a citizen chooses not to attack the regime, she receives a payoff of zero. Parameter $c \in (0, 1)$ denotes a citizen's individual cost of attacking the regime. If a citizen attacks and the regime is overthrown, the citizen receives a payoff of $1 - c > 0$. Otherwise, she only incurs the cost with no benefit from regime change, and thus her payoff is $-c < 0$.[8] The assumption that $c \in (0, 1)$ ensures that neither action is a dominant strategy for the citizen.

The regime is overthrown if $A \geq \theta$, where A is the measure of citizens who choose to attack the regime, and $\theta \in \mathbb{R}$ is the regime's strength. A citizen's incentive to attack thus increases with the aggregate size of the attack, implying that the citizens' action choices are strategic complements.

Information Structure. Citizens have heterogeneous information about the regime's strength, θ. They share the common prior $\theta \sim \mathcal{N}\left(\bar{\theta}, \sigma_\theta^2\right)$, and each citizen i receives a private signal $x_i = \theta + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is independently and identically distributed across citizens and independent of θ. Importantly, the regime can (publicly) control σ_ε^2, i.e. the variance of noise in citizens' private signals. Given this, a citizen i's

[8] It is straightforward to introduce additional features to the citizens' payoffs in this framework, e.g., a citizen could receive a spillover payoff, $s \in (0, 1 - c)$, if she does not attack and the regime is still overthrown because a sufficient measure of the other citizens have attacked, or a citizen could incur an additional cost of being repressed, $r > 0$, if she unsuccessfully attacks. It can be shown that an increase in either s or r would be equivalent to an appropriately scaled increase in c.

posterior is:

$$\mathbb{E}\left[\theta \mid x_i\right] = \frac{\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\overline{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}x_i. \tag{2.1}$$

Thus, the higher the noise, σ_ε^2, the more weight citizens attach to the prior mean relative to their private signals.

Regime's Objective. The regime's objective function is assumed to be

$$U_{\mathcal{R}}\left(\sigma_\varepsilon^2\right) := \left[1 - \Phi_{RC}\left(\theta^*\left(\sigma_\varepsilon^2\right)\right)\right] - \lambda C\left(\left|\sigma_\varepsilon^2 - \underline{\sigma_\varepsilon^2}\right|\right), \tag{2.2}$$

where $\Phi_{RC}\left(\theta^*\right)$ is the probability of regime change and function $C\left(\cdot\right)$ parametrises the regime's cost of information manipulation, with $C'\left(\cdot\right) > 0$ and $C''\left(\cdot\right) > 0$.

The intrinsic level of noise in the citizens' private signals is given by $\underline{\sigma_\varepsilon^2}$, and so the cost is higher the further from that level the value of σ_ε^2 set by the regime.[9] Parameter λ measures the importance of the cost of manipulating information relative to the benefits of staying in power.

Timeline of the Game. The game proceeds in the following steps:

1. The nature draws θ from a normal distribution $\mathcal{N}\left(\bar{\theta}, \sigma_\theta^2\right)$, which defines the initial common prior about the incumbent regime's strength, θ.

2. The regime (publicly) chooses the variance of noise in the citizens' private signals, σ_ε^2.

3. The regime's strength, θ, is realised.

4. Each citizen i receives a private signal $x_i = \theta + \varepsilon_i$, then forms a posterior about the regime's strength, θ, and decides whether or not to attack the regime.

[9] This implies that making the citizens' private signals more precise is also costly. In this paper, I focus on the regime's incentives to add noise; the incentives to make the signals more precise will be to a large extent a mirror image.

5. The regime is overthrown, or it survives. All players' payoffs are realised.

The possible interpretations of this setting are not limited to political change, but can also include self-fulfilling bank runs, currency attacks, and debt crises.[10] In the appendix, I present an application of the model to the context of paid online commentators employed by authoritarian regimes.

2.2.2 Equilibrium

In what follows, I describe the equilibrium in the subgame following the regime's choice of the level of noise in the citizens' private information, σ_ε^2. I consider perfect Bayesian Nash equilibria in monotone strategies, that is, equilibria in which the citizens' strategies are non-decreasing in each citizen's private signal, x.[11]

An equilibrium consists of a critical strength of the regime and a critical private signal, (θ^*, x^*), such that a citizen attacks whenever her private signal is below the critical signal, $x \leq x^*$, and the regime is overthrown whenever its true strength is less than the critical strength, $\theta \leq \theta^*$. An alternative statement is that there is a critical strength of the regime, θ^*, which generates a distribution of private signals such that the proportion of citizens who observe signals smaller than or equal to the critical value, x^*, is exactly θ^*, resulting in a regime change. In other words, in this equilibrium, whenever $\theta \leq \theta^*$, at least θ players attack and the regime is overthrown. Conversely, whenever $\theta > \theta^*$, no more than θ players attack and the regime remains in place.

Two conditions determine an equilibrium. The first is the *payoff indifference condition* (e.g., Hellwig, 2002), which states that—in equilibrium—a citizen must be indifferent between attacking and not attacking when she observes a private signal x^*.

[10] For example, see Angeletos, Hellwig and Pavan (2007).

[11] Restricting attention to monotone Bayesian equilibria is common in the global games literature, e.g., see Angeletos, Hellwig and Pavan (2007) and Loeper, Steiner and Stewart (2013).

With the structure of the equilibrium defined as above, this yields the condition

$$\Pr\left(\theta \leq \theta^* \mid x^*\right) = c. \tag{2.3}$$

Given the assumptions on the distribution of θ, the left-hand side can be conveniently rewritten as $\Phi\left(\left(\theta^* - \mathbb{E}(\theta \mid x^*)\right) / \sqrt{Var(\theta \mid x^*)}\right)$. The payoff indifference condition is then:

$$c = \Phi\left(\sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2 \sigma_\varepsilon^2}}\left(\theta^* - \bar{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\left(x^* - \bar{\theta}\right)\right)\right). \tag{2.4}$$

The citizen attacks if and only if $x \leq x^*$, where x^* is determined in (2.4).

The second equilibrium condition is the *critical mass condition* (e.g., Hellwig, 2002). In equilibrium, whenever the true strength of the regime is exactly θ^*, the proportion of citizens who receive a signal $x \leq x^*$ is equal to precisely θ^*:

$$\theta^* = \Pr\left(x \leq x^* \mid \theta^*\right). \tag{2.5}$$

With the distributional assumptions stated above, this can be rewritten as:

$$\theta^* = \Phi\left(\frac{x^* - \theta^*}{\sqrt{\sigma_\varepsilon^2}}\right). \tag{2.6}$$

The two equilibrium conditions stated in (2.4) and (2.6) determine an equilibrium of the game:

Proposition 2.1. *(Morris-Shin) In a Perfect Bayesian Equilibrium of the game, a critical strength of the regime, θ^*, solves:*

$$\theta^* = \Phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - \bar{\theta}\right) - \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}(c)\right). \tag{2.7}$$

There is a unique equilibrium if $\sigma_\varepsilon^2 \leq 2\pi \left(\sigma_\theta^2\right)^2$.

As usual in global games, I will assume that the citizens' private signals are sufficiently precise so that there is a unique equilibrium. Since σ_ε^2 is an instrument that is chosen by the regime before the subgame described here, the equilibrium uniqueness condition from Proposition 2.1 thus requires that the cost of manipulating the level of noise and the weighting parameter λ in the regime's objective function are not too low.

Finally, it is worth noting that in equilibrium the regime is overthrown whenever $\theta \leq \theta^*$, so the probability of regime change is

$$\Phi_{RC}\left(\theta^*\right) = \Phi\left(\frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}}\right), \tag{2.8}$$

with θ^* given by (2.7).

2.3 Regime's Incentives to Manipulate Noise

In this section, I analyse the regime's incentives to manipulate noise in the citizens' private information about the regime's strength in the context of the objective function in (2.2).[12] I start the analysis with two useful lemmas.

Lemma 2.1. *The probability of regime change,* $\Phi_{RC}\left(\theta^*\right)$*, is decreasing in:*

(i) the mean strength of the regime, $\bar{\theta}$*, i.e.* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \bar{\theta}} < 0$*;*

(ii) the citizens' individual cost of attacking, c*, i.e.* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial c} < 0$.

These results are in line with common intuition. It follows from (2.4) that an increase in either the mean strength of the regime, $\bar{\theta}$, or the citizens' individual cost of attacking, c, results in a decrease in x^*, i.e. the critical value of the private signal

[12]The impact of the precision of the citizens' private signals on the chances of regime change has been previously studied—in other contexts and settings—by Metz (2002), Bannier and Heinemann (2005), Iachan and Nenov (2015), and Szkup and Trevino (2015).

that makes a citizen indifferent between attacking the regime and not doing so. As a result, as long as the equilibrium uniqueness condition in Proposition 2.1 is satisfied, the regime's critical strength, θ^*, is decreasing in both $\bar{\theta}$ and c, which ensures that fewer citizens choose to attack and the regime becomes less likely to be successfully overthrown.

Lemma 2.2. *The probability of regime change is decreasing in the noise in citizens' private signals, i.e.* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} < 0$*, if and only if* $\theta^* < T_{RC}$*, where*

$$T_{RC} = \bar{\theta} + \sqrt{\frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}} \Phi^{-1}(c). \tag{2.9}$$

In other words, the regime's critical strength, θ^*, must be low enough for the probability of regime change to be decreasing in the noise in the citizens' private signals, σ_ε^2. Note here that, by Lemma 2.1, we have that $\partial\theta^*/\partial\bar{\theta} < 0$ and $\partial\theta^*/\partial c < 0$, while threshold T_{RC} is increasing in both $\bar{\theta}$ and c. Therefore, the higher the mean strength of the regime, $\bar{\theta}$, or the individual cost of attacking, c, the more likely it is that an increase in σ_ε^2 lowers the chances of a regime change.[13]

The intuition behind the result in Lemma 2.2 is as follows. When the mean strength of the incumbent (or the cost of attacking the regime) is high, the citizens' prior incentives to attack are rather weak. However, since the fundamentals of the model are normally distributed, there is always a positive probability that a citizen receives a very low private signal inducing her to attack the regime. By lowering the precision of private signals, the regime makes the citizens attach less weight to them and more to the prior. This may lead the citizens to refrain from attacking even for very low private signals, and consequently a regime change becomes less likely.

[13]This result is parallel to one of the main results of Metz (2002) and Bannier and Heinemann (2005). The two papers consider the central bank's optimal rules for minimising the probability of currency crises and show that full transparency is not always optimal: when prior beliefs about economic performance are good, the central bank should disclose imprecise information.

Information manipulation is nonetheless costly for the regime. If we incorporate these costs into the analysis of the regime's incentives, we arrive at the following:

Proposition 2.2. *The regime's incentives to increase the noise in the citizens' private signals, σ_ε^2, are non-monotonic with respect to the mean strength of the regime, $\bar{\theta}$, and the citizens' individual cost of attacking the regime, c:*

(i) for sufficiently low and sufficiently high values of $\bar{\theta}$ (or c), the regime chooses not to increase σ_ε^2;

(ii) for moderate values of $\bar{\theta}$ (or c), the regime chooses to increase σ_ε^2 if and only if λ is small enough.

The fact that the regime would never choose to add noise to the citizens' private signals for sufficiently low values of $\bar{\theta}$ and c follows from Lemma 2.2. On the other hand, we know from Lemma 2.1 that as $\bar{\theta}$ or c increase to very high levels, the probability of regime change, Φ_{RC}, becomes very small. As a result, when $\bar{\theta}$ or c reaches an arbitrarily high level, the marginal decrease in the probability of regime change caused by raising σ_ε^2 approaches zero. Thus, given that increasing σ_ε^2 is costly, the regime would also not choose to raise σ_ε^2 for sufficiently high values of $\bar{\theta}$ and c.

If the weighting parameter λ is small enough, then the benefits from a decrease in Φ_{RC} are large relative to the marginal costs, and there exists an interval of moderate values of $\bar{\theta}$ and c where it is optimal for the regime to increase σ_ε^2. Figure 2.1 presents a simulation of the result stated in Proposition 2.2.[14]

[14] A simulation with respect to the citizens' individual cost of attacking, c, produces a very similar figure.

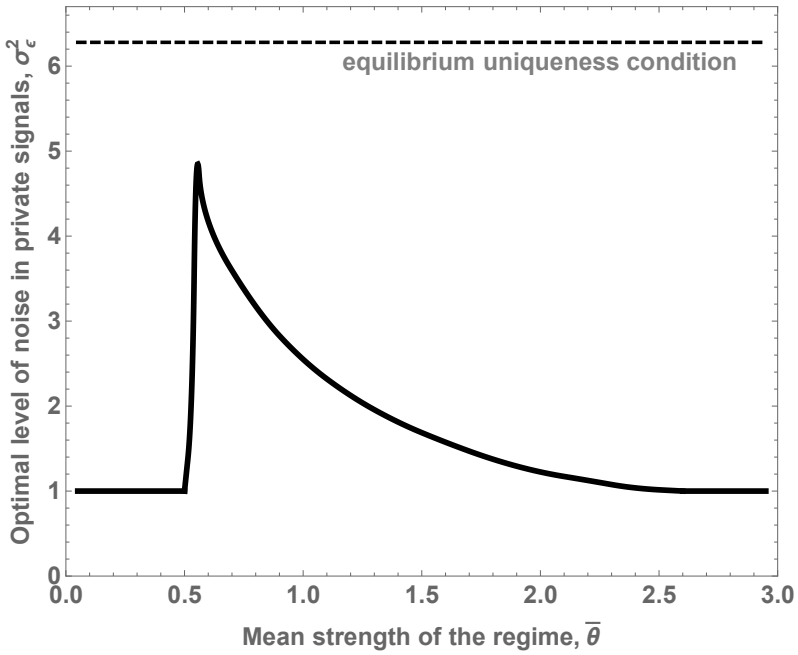


Figure 2.1: A simulation of the regime's optimal choice of the variance of noise in the citizens' private signals, σ_ε^2, as a function of the mean strength of the regime, $\bar{\theta}$. Calibration assumptions: $\sigma_\theta^2 = 1$, $\underline{\sigma_\varepsilon^2} = 1$, $c = \frac{1}{2}$, $C(x) = x^2$, and $\lambda = 1/250$.

Discussion

It is useful to contrast the results in this section with Edmond (2013), where the regime's ability to increase the likelihood of survival by adding a bias to the citizens' private signals stems from the fact that the regime is privately informed about its own strength, and so the citizens need to infer the regime's hidden action jointly with its strength. On the other hand, as Proposition 2.2 shows, information manipulation in the form of adding noise—which could illustrate distraction of citizens—is effective under weaker informational requirements than adding a bias to their signals—which could capture persuasion.

The model thus provides one possible explanation for the main finding of King, Pan and Roberts (2017), which states that the reason why the Chinese regime fabricates social media posts is not because it wants to persuade the skeptics but because it wants

to distract them. It may be that the Chinese regime is aware of the fact that preventing a collective attack through persuasion (or by adding a bias to the citizens' information) would be rather difficult in that it would require (i) the citizens to be uninformed about the government's manipulative action and (ii) the regime to be better informed about its strength than the citizens are.[15] Adding noise to the citizens' information provides an easier alternative that may be effective even if these conditions are not satisfied.

2.4 Targeting State Criticism and Collective Action

In this section, I investigate what features of the environment determine whether the regime's incentives to prevent regime change (as outlined in Section 2.3) are driven (i) by the need to minimise the measure of citizens whose posteriors are below a critical threshold, and (ii) by the need to minimise the size of collective attacks.

First, let us consider how adding noise affects the size of collective attacks, and to what extent this impact is different from the one on the probability of regime change. Given that in the equilibrium a citizen attacks if and only if $x \leq x^*$, the expected size of the collective attack is $\Pr(x \leq x^*) = \Phi\left(\left(x^* - \bar{\theta}\right)/\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}\right)$, i.e.

$$\mathbb{E}\left[A\left(\theta^*\right)\right] = \Phi\left(\left(\theta^* - \bar{\theta}\right)\frac{\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}}{\sigma_\theta^2} - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(c\right)\right), \tag{2.10}$$

where θ^* is given by (2.7).

Like for the probability of regime change, it can be shown that the expected size of the attack, $\mathbb{E}\left[A\left(\theta^*\right)\right]$, is decreasing in both the mean strength of the regime, $\bar{\theta}$, and the individual cost of attacking the regime, c. Furthermore, as $\bar{\theta} \to \infty$ or $c \to 1$,

[15] Edmond (2013) provides numerical results which show that manipulating noise would also be effective in his informational setting (i.e. where the regime is informed about θ and takes a hidden action).

the expected size of the attack approaches zero, and it approaches 1 when $\bar{\theta} \to -\infty$ or $c \to 0$,. This observation suggests that the regime's incentives to minimise the expected size of the attack follow a similar pattern to the one in Proposition 2.2, i.e. that they are non-monotonic with respect to $\bar{\theta}$ and c, being highest for moderate values.

Comparing (2.10) with (2.8), we can see however that, for a sufficiently small value of $\bar{\theta}$ (or c), the expected size of the attack is larger than the probability of regime change, but any increase in $\bar{\theta}$ (or c) always has a more prominent impact on $\mathbb{E}\left[A\left(\theta^*\right)\right]$ than it has on Φ_{RC}. Consequently, for a sufficiently high value of $\bar{\theta}$ (or c), the converse holds: $\mathbb{E}\left[A\left(\theta^*\right)\right] < \Phi_{RC}$.[16] This follows from the fact that the attack size approaches zero as as $\bar{\theta} \to \infty$ or $c \to 1$, while the probability of regime change remains strictly positive: there is still a possibility that $\theta \leq 0$, in which case the regime is overthrown even if it is not attacked at all.

The contrast in how Φ_{RC} and $\mathbb{E}\left[A\left(\theta^*\right)\right]$ behave as functions of $\bar{\theta}$ and c suggests that the regime's incentives to minimise collective attacks are different from its incentives to prevent regime change. The following proposition confirms this:

Proposition 2.3. *Consider the marginal impact of noise in the citizens' private signals on the probability of regime change,* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2}$*, and on the expected size of the collective attack,* $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2}$*:*
(i) when $\sigma_\varepsilon^2 \to 0$, $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} < 0$ *if and only if* $\bar{\theta} > \tau_{RC,0} = 1 - c$, *and* $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2} < 0$ *if and only if* $\bar{\theta} > \tau_{A,0}$*;*
(ii) when $\sigma_\varepsilon^2 \to \infty$, *both* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} < 0$ *and* $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2} < 0$ *if and only if* $\bar{\theta} > \tau_{A,\infty} = \tau_{RC,\infty}$*, where* $\tau_{A,0} > \tau_{RC,\infty} > \tau_{RC,0}$ *when* $c < \frac{1}{2}$*, and* $\tau_{A,0} < \tau_{RC,\infty} < \tau_{RC,0}$ *when* $c > \frac{1}{2}$*.*
Furthermore, $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} = \frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2} = 0$ *for both limiting cases when* $\bar{\theta} = c = \frac{1}{2}$*.*

It is best to explain the statements of Proposition 2.3 with a graphical illustration.

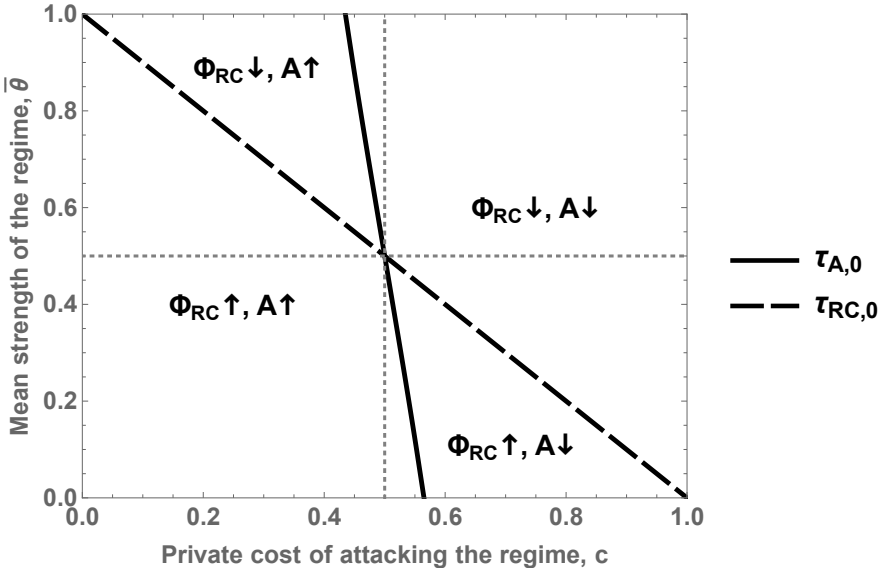
[16] As σ_ε^2 becomes arbitrarily small, we have $\mathbb{E}\left[A\left(\theta^*\right)\right] = \Phi_{RC}$. To see this, observe that the citizens then observe the regime's strength almost perfectly and thus attack whenever $\theta < \lim_{\sigma_\varepsilon^2 \to 0} \theta^* = 1 - c$, which is also precisely when the regime is overthrown.

With c on the x-axis and $\bar{\theta}$ on the y-axis, Figure 2.2 illustrates the parameter spaces where $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2}$ and $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2}$ are positive or negative in two limiting cases: when the citizens' private signals are intrinsically very precise, $\sigma_\varepsilon^2 \to 0$, and when they are intrinsically very noisy, $\sigma_\varepsilon^2 \to \infty$. Threshold τ_{RC} depicts the critical value of $\bar{\theta}$—as a function of c—such that whenever $\bar{\theta}$ is larger than this critical value, the probability of regime change in the global game is decreasing in σ_ε^2, i.e. $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} < 0$. Threshold τ_A illustrates the equivalent for the size of the collective attack.

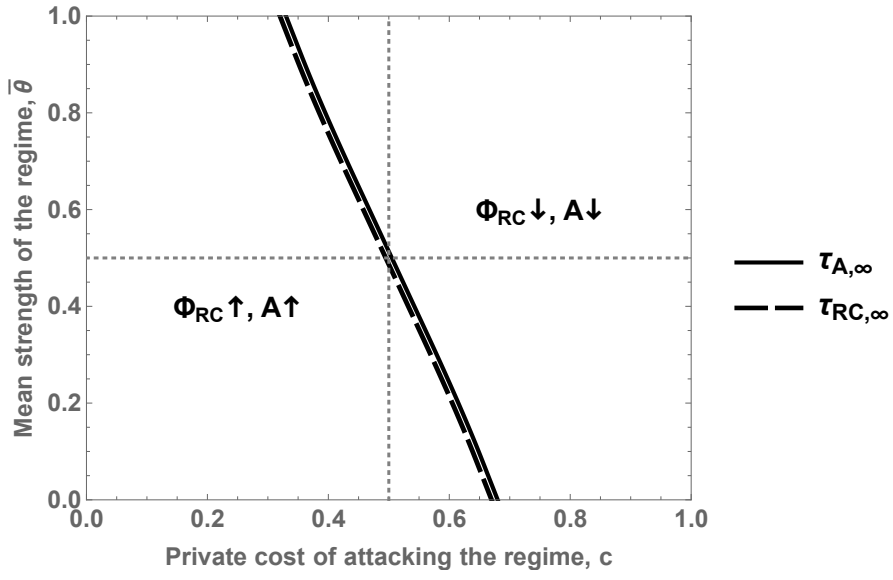
Proposition 2.3 implies that when $\sigma_\varepsilon^2 \to 0$, the threshold τ_A is above (below) the threshold τ_{RC} as long as $c < (>) \frac{1}{2}$. On the other hand, when $\sigma_\varepsilon^2 \to \infty$, the two thresholds coincide. Furthermore, the very last statement in the proposition implies that the two thresholds cross when $\bar{\theta} = c = \frac{1}{2}$. Proposition 2.3 thus tells us that the regime's incentives to prevent regime change and to prevent collective attacks are substantially different when the citizens' private signals are intrinsically very precise but are qualitatively the same when the citizens' signals are intrinsically noisy.

Consider first the case $\sigma_\varepsilon^2 \to 0$ in Figure 2.2(i). As part (i) of Proposition 2.3 shows, in order to prevent regime change, the regime should increase noise whenever the mean strength of the regime, $\bar{\theta}$, is greater than $\lim_{\sigma_\varepsilon^2 \to 0} \theta^* = 1 - c$. This is because—with private signals that are intrinsically very precise—the citizens observe the regime's strength, θ, almost perfectly and attack when they see that θ is less than the threshold θ^*. The regime's best strategy of survival is then to increase noise whenever $\bar{\theta}$ is greater than θ^*, or in other words to minimise the expected proportion of citizens whose posteriors are below the critical threshold. On the other hand, if the regime aimed to minimise all collective attacks—as opposed to preventing regime change—then it would be less concerned about whether any given attack is likely to be successful. As we can see from Figure 2.2(i), its optimal strategy of whether to increase noise would then mostly depend on the individual cost of attacking the regime rather than the mean

strength.



(i) $\sigma_\varepsilon^2 \to 0$;



(ii) $\sigma_\varepsilon^2 \to \infty$.

Figure 2.2: The sign of the marginal impact of an increase in σ_ε^2 on the change in the probability of regime change, Φ_{RC}, and on the size of the collective attack, $\mathbb{E}\left[A\left(\theta^*\right)\right]$, in two limiting cases: (i) $\sigma_\varepsilon^2 \to 0$, and (ii) $\sigma_\varepsilon^2 \to \infty$.

Second, consider the case $\sigma_\varepsilon^2 \to \infty$ in Figure 2.2(ii). As part (ii) of Proposition 2.3 demonstrates, it is then optimal for the regime to add noise in order to prevent regime change whenever it also leads to a smaller size of the attack. The intuition for

this result is as follows. Since the citizens' private signals then provide little additional information about the regime's strength, the citizens decide whether to attack relying almost solely on their prior information. Consequently, all the attacks are targeted very poorly, and the regime optimally behaves as if its aim were to minimise the expected size of attacks.

Taken together, the cases $\sigma_\varepsilon^2 \to 0$ and $\sigma_\varepsilon^2 \to \infty$ imply that the regime's incentives to add noise to the citizens' information can be decomposed into (i) minimising the proportion of citizens whose posteriors are below the critical threshold, and (ii) minimising the expected size of collective attacks.

Finally, Proposition 2.3 and Figure 2.2 also demonstrate that when both $\bar{\theta}$ and c are high or both are low, the regime's incentives to prevent regime change do not differ significantly from its incentives to minimise the size of collective attacks. However, for high values of c and sufficiently low values of $\bar{\theta}$, adding noise decreases the size of collective attacks but makes regime change more likely—especially when the citizens' private signals are intrinsically precise. Thus, too much focus on minimising the size of all collective attacks—rather than on those that might overthrow the regime—may result here in the regime adding too much noise to the citizens' private signals. Conversely, when c is low and $\bar{\theta}$ is sufficiently high, adding noise boosts the size of collective attacks but makes regime change less likely, and therefore too much focus on minimising the size of all collective attacks may lead to less noise being added than is optimal.

Discussion

The results of this section provide a certain structure for thinking about the empirical findings of King, Pan and Roberts (2013; 2014) and their two theories of what constitutes the goals of the Chinese censorship programme: the state critique theory and the theory of collective action potential. King et al. (2013; 2014) show empirical

evidence suggesting that, in the context of censorship in China, the latter theory is correct, while the former is not; in other words, the censorship system in China aims to stop the spread of information that could lead to any kind of collective action, rather than to restrict criticism of the state *per se*. If we extrapolate the results of King et al. (2013; 2014) to manipulation that aims to add noise to the citizens' information, the model can provide some theoretical rationale for their empirical findings.

The analysis in this section has demonstrated that the regime's incentives to prevent regime change can be decomposed into the need to minimise the measure of citizens whose posteriors are below a critical threshold and the need to minimise the size of all collective attacks. Whether the regime's strategy of increasing the noise should be driven predominantly by the former or the latter depends on how well informed the citizens are about the regime's strength. The following corollary recapitulates a crucial observation from Proposition 2.3:

Corollary 2.1. *Consider the marginal impact of noise in the citizen's private signals on the probability of regime change,* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2}$*, and on the expected size of the collective attack,* $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2}$*:*

(i) if $\sigma_\varepsilon^2 \to 0$*, then* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} < (>)\, 0$ *if and only if* $\bar{\theta} > (<)\, \theta^*$*;*

(ii) if $\sigma_\varepsilon^2 \to \infty$*, then* $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} < (>)\, 0$ *if and only if* $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2} < (>)\, 0$*.*

This result suggests that when the citizens are intrinsically well informed about the regime's strength, it is optimal for the regime to focus on minimising the proportion of citizens whose posteriors are below the critical threshold. In other words, the model indicates that the regime should be then more concerned about how critical the citizens' opinions are about the regime, and therefore *the state critique theory* should apply. The opposite is true, however, when the citizens are intrinsically poorly informed about the regime's strength: the regime's incentives to prevent regime change are then driven predominantly by the need to minimise the size of collective attacks. Consequently,

under these circumstances, we should expect *the theory of collective action potential* to prove correct.

In the context of King et al. (2013; 2014), these results imply that the Chinese regime's focus on preventing collective action rather than suppressing state criticism may be due to the Chinese citizens being intrinsically (i.e. even before the regime's manipulation efforts) poorly informed about the regime's strength.

2.5 Role of Coordination and Information Aggregation

An important implication of the empirical findings in King, Pan and Roberts (2013; 2014) is that the citizens' ability to coordinate a collective action should have a significant impact on the incentives of authoritarian regimes to manipulate information. In this section, I investigate how the regime's incentives to add noise to the citizens' information are affected by the degree to which citizens can coordinate their actions. In particular, while the model has so far assumed that citizens are imperfectly coordinated, I now also allow for the possibility that they are perfectly coordinated.

2.5.1 Perfect Coordination Benchmark

Suppose that the mass of citizens acts like one "large citizen" who is able to overthrow the regime with certainty as long as the incumbent's strength, θ, is less than the large citizen's mass, 1.[17] It turns out that the implications of perfect coordination for the regime's incentives depend on the extent to which the citizens can aggregate their

[17]Similar perfect coordination benchmarks—but with a different focus of the analysis—appear in Corsetti, Dasgupta, Morris and Shin (2004) and Edmond (2013).

collective information.[18]

At one extreme is a scenario where the citizens can perfectly aggregate all the information they collectively possess in their private signals. If they were able to do so, they would be able to back out the regime's strength, θ, regardless of the level of noise in their private signals, σ_ε^2. Consequently, the regime would have no incentives to add noise to the citizens' private signals, which is in stark contrast to the regime's incentives when citizens are imperfectly coordinated, as discussed in Section 2.3. The citizens would attack (i.e. $A^{PCA} = 1$) and overthrow the regime whenever $\theta \leq 1$, and would not attack it otherwise (i.e. $A^{PCA} = 0$). The *ex ante* probability of regime change would thus be $\Phi_{RC}^{PCA} = \Phi\left(\left(1 - \bar{\theta}\right)/\sqrt{\sigma_\theta^2}\right)$.

At the other extreme is a scenario in which the citizens are perfectly coordinated but are unable to aggregate the information they collectively possess. In other words, suppose there is a "large citizen" who receives a single signal, $x = \theta + \varepsilon$, with the variance of the error term being σ_ε^2 as before. We could interpret this environment as one with a "political leader" who is not better informed than any other citizen but who is able to make all the citizens attack the regime if she so desires. In this case, the regime still has incentives to add noise, but—as I show below—these turn out to be different from those in Section 2.3.

2.5.2 Impact of Coordination and Information Aggregation on the Regime's Incentives to Manipulate Noise

I focus my analysis on a comparison of the regime's incentives to add noise in the perfect coordination benchmark with no information aggregation and in the setting where citizens are imperfectly coordinated (i.e. in the global game). This allows me to

[18]Henceforth, in the mathematical notation, I use "PCA" as an abbreviation for "perfect coordination with (information) aggregation", and "PCN" as an abbreviation for "perfect coordination with no (information) aggregation".

highlight the extent to which the citizens' inability to aggregate their information can affect the role of coordination in the regime's manipulation incentives.

In the perfect coordination benchmark with no information aggregation, the "large citizen" chooses to attack the regime if $\Pr(\theta \leq 1 \mid x) \leq c$ and does not attack it otherwise. This means that the critical mass condition from the global game is no longer necessary to establish the equilibrium and it is only the payoff indifference condition that matters. The "large citizen" is indifferent between attacking and not attacking if

$$c = \Phi\left(\sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2\sigma_\varepsilon^2}}\left(1-\bar{\theta}-\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2}\left(x_{PCN}^*-\bar{\theta}\right)\right)\right), \tag{2.11}$$

which is notably different from (2.4). Thus, the citizen attacks if and only if her private signal is less than the critical value, $x \leq x_{PCN}^*$, where:

$$x_{PCN}^* = \frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2} - \frac{\sigma_\varepsilon^2}{\sigma_\theta^2}\bar{\theta} - \sqrt{\frac{\sigma_\varepsilon^2(\sigma_\theta^2+\sigma_\varepsilon^2)}{\sigma_\theta^2}}\Phi^{-1}(c). \tag{2.12}$$

Inspection of (2.4) shows that $x_{PCN}^* > x^*$ for all $\theta^* < 1$, i.e. the critical threshold of the private signal below which the citizens choose to attack the regime is always higher when they are perfectly coordinated and cannot aggregate their information than when they are imperfectly coordinated.[19] Furthermore, the expected size of the attack is

$$\mathbb{E}\left[A^{PCN}\right] = \Phi\left(\left(1-\bar{\theta}\right)\frac{\sqrt{\sigma_\theta^2+\sigma_\varepsilon^2}}{\sigma_\theta^2} - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}(c)\right), \tag{2.13}$$

which is always higher than (2.10), i.e. the expected size of the attack when the citizens are imperfectly coordinated.

Thus, perfect coordination with no information aggregation, on average, always

[19]Note that, with the citizens being perfectly coordinated, the regime faces aggregate uncertainty since it does not know which value of x will be realised. The size of the attack is 1 if $x \leq x_{PCN}^*$ and 0 otherwise.

yields more collective attacks than we observe when the citizens are imperfectly coordinated. This statement does not imply, however, that also regime change is more likely. For the regime to be overthrown, not only do we need the citizen to attack (which happens whenever $x \leq x^*_{PCN}$), but we also need the regime's strength to satisfy $\theta < 1$. The ex ante probability of regime change is then $\Phi^{PCN}_{RC} = \Pr\left(x \leq x^*_{PCN}, \theta \leq 1\right)$.

In the subsequent analysis, I focus on the case where the negative impact of increasing σ^2_ε on Φ^{PCN}_{RC} is largest in absolute terms:

Lemma 2.3. *$\frac{\partial}{\partial \sigma^2_\varepsilon}\Phi^{PCN}_{RC}$ is negative and is minimised for all σ^2_ε when $\bar{\theta} = 1$ and $c = \frac{1}{2}$.*

While being restrictive, the analysis of the special case still allows for interesting observations about how the regime's incentives to increase noise are affected by the citizens' ability to coordinate actions and aggregate their information.[20]

First, note from (2.13) that when $\bar{\theta} = 1$ and $c = \frac{1}{2}$, the expected size of an attack in the perfect coordination benchmark with no information aggregation is $\mathbb{E}\left[A^{PCN}\right] = \Phi\left(0\right) = \frac{1}{2}$, i.e. it is the same regardless of the level of σ^2_ε set by the regime. On the other hand, when $\bar{\theta} = 1$ and $c = \frac{1}{2}$, the expected size of an attack with imperfectly coordinated citizens is

$$\mathbb{E}\left[A\left(\theta^*\right)\right] = \Phi\left(\left(\theta^* - 1\right)\frac{\sqrt{\sigma^2_\theta + \sigma^2_\varepsilon}}{\sigma^2_\theta}\right), \tag{2.14}$$

where $\theta^* = \Phi\left(\sqrt{\sigma^2_\varepsilon}\left(\theta^* - 1\right)/\sigma^2_\theta\right)$, which we obtain from (2.7) and (2.10). Since $\theta^* \in (0,1)$, $\mathbb{E}\left[A\left(\theta^*\right)\right]$ is decreasing in σ^2_ε. Thus, adding noise to the citizens' private information lowers the expected size of an attack when they are imperfectly coordinated, but does not affect it when they are perfectly coordinated and rely only on a single signal.

[20] Lemma 2.3 is a useful result since, in general, calculating Φ^{PCN}_{RC} is non-trivial. Generically, there are no closed form solutions for $F\left(\boldsymbol{x}\right) = \Pr\left(\mathbf{X} \leq \mathbf{x}\right)$, where $\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. There is, however, an analytical solution for the standardised bivariate normal distribution in the special case where $\mathbf{x} = (0,0)$. Setting $\bar{\theta} = 1$ and $c = \frac{1}{2}$ allows me to use it.

This means that if the regime aimed to minimise the expected size of the collective attack, it would choose to increase the noise in the former case only.

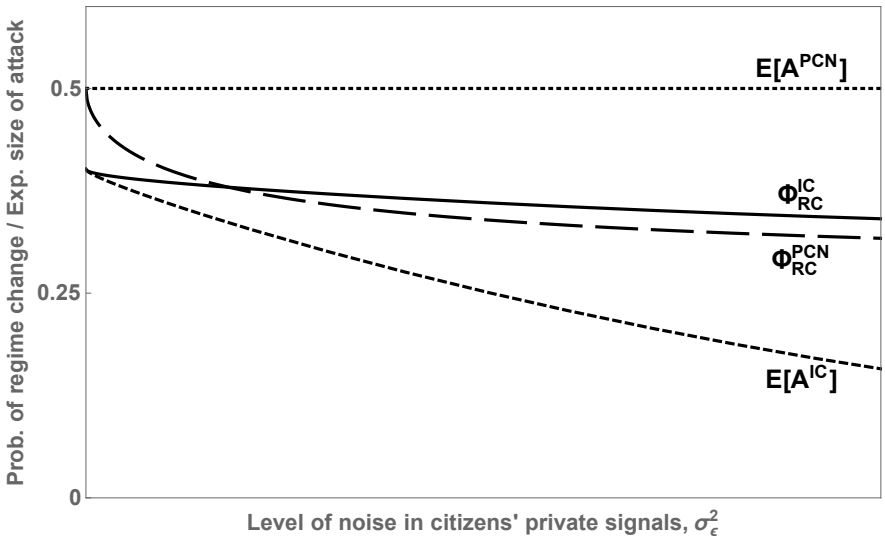
However, the implications for the regime's incentives to prevent regime change may be dramatically different:

Proposition 2.4. *Suppose* $\bar{\theta} = 1$ *and* $c = \frac{1}{2}$. *If* σ_θ^2 *is sufficiently low, then* $\Phi_{RC}(\theta^*) < \Phi_{RC}^{PCN}$ *for all values of* σ_ε^2. *Otherwise,* $\Phi_{RC}(\theta^*) < \Phi_{RC}^{PCN}$ *for sufficiently low values of* σ_ε^2 *and* $\Phi_{RC}(\theta^*) > \Phi_{RC}^{PCN}$ *for sufficiently high values of* σ_ε^2.

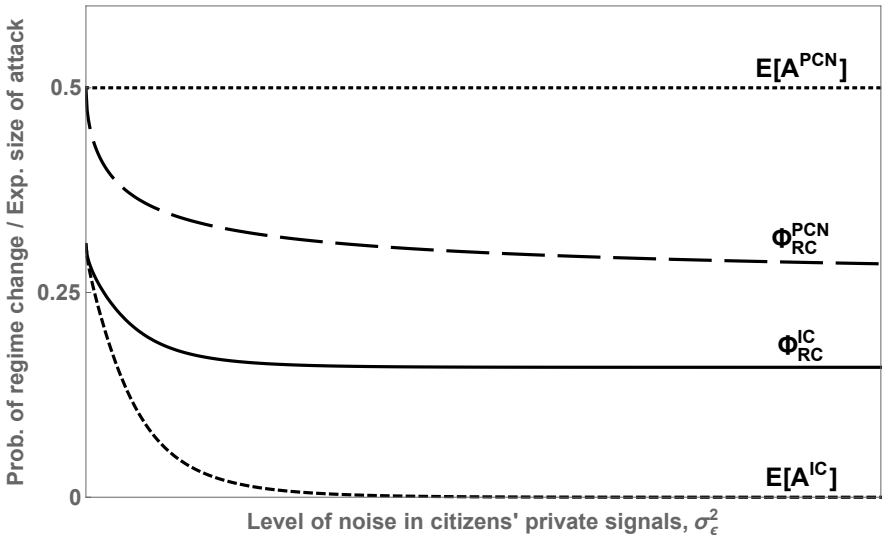
Proposition 2.4 states that when the strength of the regime, θ, is distributed tightly, then for a given σ_ε^2 the regime is always more likely to survive when citizens are imperfectly coordinated rather than when they are perfectly coordinated but cannot aggregate their collective information. The situation is however different when θ is dispersed, i.e. when σ_θ^2 is high. Then the regime is more likely to survive with imperfectly coordinated citizens when σ_ε^2 is low, but when σ_ε^2 becomes very high, the regime is actually better off with citizens who are perfectly coordinated but their ability to aggregate their collective information is impeded.

2.5.3 A Graphical Illustration

The implications of Proposition 2.4, and in particular the difference between the cases with a high and a low value of σ_θ^2, are illustrated in Figure 2.3. In this figure, Φ_{RC}^{IC} and $\mathbb{E}\left[A^{IC}\right]$ denote, respectively, the probability of regime change and the expected size of the attack when citizens are imperfectly coordinated (i.e. in the global game setting); Φ_{RC}^{PCN} and $\mathbb{E}\left[A^{PCN}\right]$ are the equivalents for the scenario where the citizens are perfectly coordinated but cannot aggregate their collective information.



(i) strength of the regime, θ, is dispersed, i.e. σ_θ^2 is high;



(ii) strength of the regime, θ, is distributed tightly, i.e. σ_θ^2 is low.

Figure 2.3: The probability of regime change in the global game, Φ_{RC}^{IC}, and in the perfect coordination benchmark with no information aggregation, Φ_{RC}^{PCN}, as a function of the noise in the citizens' private signals, σ_ε^2 (in the special case where $\bar{\theta} = 1$ and $c = \frac{1}{2}$).

In Figure 2.3(i), σ_θ^2 is assumed to be high. While the expected size of the "large citizen's" attack stays constant as σ_ε^2 increases, this attack is more and more poorly targeted, i.e. it often occurs when the regime is strong enough to suppress it. Consequently, the probability of regime change falls quite abruptly. On the other hand,

with imperfectly coordinated citizens, we observe that the expected size of the attack falls as σ_ε^2 increases, yet the probability of regime change falls less prominently than in the perfect coordination benchmark with no information aggregation. This is because these imperfectly coordinated citizens collectively possess more information about the regime's strength than the "large citizen", and this allows them to target their attacks more precisely. In fact, for a sufficiently high σ_ε^2, they overthrow the regime with higher probability than the "large citizen" who receives a single signal. This implies that when σ_θ^2 is high, the regime may have stronger incentives to increase noise when citizens are perfectly coordinated but cannot aggregate their collective information than when they are imperfectly coordinated.

Figure 2.3(ii) illustrates a scenario where σ_θ^2 is low. The situation is now markedly different from the one in Figure 2.3(i): the fact that the imperfectly coordinated citizens collectively possess more information never compensates for their poor coordination, and the regime is always more likely to be overthrown by the "large citizen" who receives a single signal.

In sum, the analysis in Section 2.5 has two important implications. First, it demonstrates that adding noise to the citizens' information may be an effective method of preventing regime change even if it does not reduce the expected size of collective attacks. This is because it still makes the attacks less precisely targeted, that is, they occur more often when the regime is strong enough to overcome them. Second, the analysis shows that coordination and information aggregation are both important—yet distinct—factors affecting collective action. In particular, if better coordination of citizens comes at the cost of impeded aggregation of information, it may actually boost the regime's incentives to manipulate information. As a result, the regime may have better chances of survival with perfectly rather than imperfectly coordinated citizens. As we have seen above, this is especially likely when there is significant prior uncertainty

about the regime's strength.

2.6 Conclusion

In this paper, I have used the global games framework to investigate the incentives of authoritarian regimes to manipulate the information available to the citizens by adding noise to their private signals, with a particular emphasis on the role of the citizens' ability to coordinate a collective action. The analysis is also a step towards providing a theoretical rationale for the empirical findings of King, Pan and Roberts (2013; 2014; 2017). The first two of these papers have shown empirically that, when censoring content, the Chinese regime aims to silence comments that may lead to a collective action and is not concerned about mere criticism of the state. The last one, on the other hand, demonstrates that the Chinese regime fabricates social media posts to distract the public rather than to engage in discussions with skeptics and try to improve their opinion of the government.

The empirical observations of King et al. (2017) are consistent with a formal setting in which the regime is aware that manipulating information by adding a bias to the public opinion is informationally difficult. In particular, as Edmond (2013) has shown, for adding a bias to be effective, the citizens would need to be uninformed about the regime's manipulative actions and the incumbent would need to be better informed about its strength than the citizens are. On the other hand, adding noise into citizens' information provides an effective alternative even when these two conditions are not satisfied. Furthermore, when they are not satisfied, this paper shows that the regime's incentives to increase noise are non-monotonic in (i) the citizens' individual cost of attacking the regime and (ii) the mean strength of the incumbent, being strongest for moderate values of these parameters.

In addition to this, I have investigated the settings in which the regime's incentives to add noise are predominantly driven by the desire (i) to minimise the measure of citizens whose posteriors are below a critical threshold, or (ii) to minimise the size of all collective attacks. One can think of these two characteristics as capturing the ideas behind, respectively, the state critique theory and the theory of collective action potential, as defined by King et al. (2013; 2014). I have found that the better the citizens are informed about the regime's strength *a priori*, the more the regime's optimal strategy of whether to add noise focuses on the former rather than on the latter. The model thus suggests that the Chinese regime's focus on preventing collective action rather than suppressing state criticism may be due to the fact that the Chinese citizens' knowledge about the regime is poor. While the analysis presented in this paper is a step toward a better understanding of what drives authoritarian regimes to manipulate information—as well as the empirical findings of King et al. (2013; 2014; 2017)—it is not without limitations. In particular, it would be fruitful to investigate the role of the citizens' knowledge about the regime in more general environments with weaker assumptions on the information structure.

Finally, I have demonstrated that the impact of citizens' better coordination on the regime's incentives to add noise depends on how well the citizens can aggregate the information that they collectively possess. If they can perfectly coordinate an attack but rely only on the signal of their leader, the regime's incentives to increase noise may in fact be stronger than if the citizens were imperfectly coordinated but could still partially aggregate their collective information. Better coordination may thus be counterproductive as it may lower the chances that the regime is overthrown. This paper shows that this is especially likely to be true when there is significant prior uncertainty about the regime's strength. Although the analysis of the interactions between the citizens' coordination and information aggregation abilities is stylised, I conjecture that the

mechanisms identified here should still play an important role in more general settings. Nevertheless, the analysis of how the citizens' ability to aggregate their information interacts with their ability to coordinate an attack, and how this interplay affects the regime's incentives to manipulate information, remains an interesting avenue for future research.

Appendix to Chapter 2

B.1 Further Discussion

B.1.1 Spillovers and Repressions

Suppose that the citizens' payoffs are as follows:

	Attack is successful	Status quo remains
Attack	$1-c$	$-c-r$
Do Not Attack	s	0

As before, the parameter $c \in (0,1)$ denotes the citizens' individual cost of attacking the regime. I introduce here an additional cost, r, incurred by the citizen when he attacks the regime and the attack turns out to be unsuccessful. This reflects the fact that a citizen may then be repressed by the incumbent regime. Furthermore, I also allow for a positive payoff of s for a citizen who chooses not to attack the regime but the collective attack turns out to be successful (because sufficiently many other citizens have attacked the regime). I assume that $s \in [0, 1-c)$, where s illustrates a spillover effect. The difference $1-s$ measures the extent to which, once the regime is overthrown, those who were among the attackers gain privileged status relative to those who held back.[21]

By assuming that $s < 1-c$, we ensure that the game does not collapse to a prisoner's dilemma; otherwise, it would be a strictly dominant strategy not to attack. It is also worth noting that, as long as $s > 0$, there is a positive externality to attacking: if an attack is successful, the non-attackers are also better off. In principle, one could also imagine a situation where $s < 0$, i.e. that the citizens who do not participate in the

[21] The idea of a privileged status of attackers appears in political science literature, e.g., Bueno de Mesquita (2010).

attack are worse off after a regime change, for example, because of prosecution by the new regime, ostracism leading to diminished career prospects, etc.

With the payoffs specified as above, the citizen's payoff indifference condition is:

$$(1-c)\Pr\left(\text{attack successful} \mid x^*\right) + (-c-r)\left(1-\Pr\left(\text{attack successful} \mid x^*\right)\right)$$

$$= s\Pr\left(\text{attack successful} \mid x^*\right). \tag{2.15}$$

This simplifies to:

$$\Pr\left(\text{attack successful} \mid x^*\right) = \frac{c+r}{1-s}. \tag{2.16}$$

With the structure of the equilibrium defined as in Section 2.2, this yields

$$\Pr\left(\theta \leq \theta^* \mid x^*\right) = \frac{c+r}{1-s}, \tag{2.17}$$

and the payoff indifference condition is then:

$$\frac{c+r}{1-s} = \Phi\left(\sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2\sigma_\varepsilon^2}}\left(\theta^* - \bar{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2}\left(x^* - \bar{\theta}\right)\right)\right). \tag{2.18}$$

Given that the new payoff structure does not affect the critical mass condition, an equilibrium value of θ^* is given by:

$$\theta^* = \Phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - \bar{\theta}\right) - \sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(\frac{c+r}{1-s}\right)\right). \tag{2.19}$$

Thus, the only way in which the introduction of parameters r and s affects (2.19) is through the ratio $\frac{c+r}{1-s}$. In other words, in this model, the impact of $r > 0$ or $s > 0$ is equivalent to an appropriately scaled increase in c.

B.1.2 Example: Online Commentators

I present here how the model, with a few additions, could be applied to the context of online commentators (or "internet trolls") who are employed by authoritarian regimes to post comments that are favourable to the incumbent. This exercise is helpful in that it highlights the kind of assumptions that have to be made to apply the main model to a particular context.

Suppose that, in addition to the regime and the citizens, there is a continuum of online commentators of measure 1, who are indexed by i like the citizens and uniformly distributed over $[0, 1]$. This means that there is a one-to-one correspondence (a bijective function) between commentators and citizens, i.e. each citizen follows only one commentator and each commentator is listened to by only one citizen.

Each commentator i has a certain intrinsic valuation for money, which is denoted by v. Their utility from praising the government at a level $b_i \in R$ is given by:

$$U_o := (\theta + v_i y)b_i - c_o(b_i), \tag{2.20}$$

where θ is the state of the world and y is a payment that the online commentator receives for an additional unit of b_i. The praise level, b_i, can be interpreted as the number of pro-government posts written by an online commentator.[22] Intuitively, the stronger the regime is (i.e. the higher θ is), the more eager a commentator is to write positive comments about the government (*intrinsic motivation*). This may also reflect the fact that it is easier for the commentator to praise the government when the regime is strong. But the degree to which he praises the government depends also on the payment, y, which he receives from the government (*extrinsic motivation*). The extent to which a commentator is responsive to this payment depends on his intrinsic valuation

[22] In fact, an online commentator's optimal action, b, would be a measure of both the quantity of pro-government posts and the extent to which they praise the regime (their "intensity").

for money, v.

However, online commentators incur also a cost of writing comments, $c_o(b_i) = \frac{1}{2}kb_i^2$, which is increasing in the praise level, b_i. Parameter k measures here the relative importance of the cost of effort in the online commentator's utility function. The optimal choice of the number of pro-government comments written is then $b_i^* = (\theta + v_i y)/k$. For simplicity, I assume that $k = 1$.[23]

In order to make the global games model applicable to this context, I assume that online commentators are the only players to observe the exact values of θ and v_i; the former is realised after the contract is signed, and the latter is each commentator's private information. However, the distribution of θ and v is common knowledge among all players, with $v \sim \mathcal{N}(\bar{v}, \sigma_v^2)$ and no correlation between θ and v.[24]

Each citizen i observes the online commentator's action, b_i, and uses it to update her information about the prior. Given this and the publicly known value of the payment, y, the posterior of a citizen i is then

$$\mathbb{E}[\theta \mid b_i] = \frac{y^2\sigma_v^2}{\sigma_\theta^2 + y^2\sigma_v^2}\overline{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2}(b_i - \bar{v}y), \tag{2.21}$$

which is equivalent to (2.1) since $b_i - \bar{v}y$ is normally distributed with mean zero and variance $y^2\sigma_v^2$. In other words, an increase in y acts in the same way as an appropriately scaled increase in σ_ε^2. Furthermore, given the bijective relationship between commentators and citizens, the noise is identically and independently distributed across citizens. Finally, no correlation between θ and v ensures, given their normal distributions, that

[23] These preferences of online commentators are inspired by Bénabou and Tirole (2003; 2006), who study the behaviour of citizens choosing the extent of their participation in some prosocial activity, e.g., contribution to a public good.

[24] It could be argued, however, that allowing for $Cov(\theta, v) > 0$ would be more realistic. For example, when the regime is indeed strong and the state of the world is high, online commentators might feel less guilty about being responsive to the payment they receive, and thus their intrinsic valuation for money may also be higher.

noise is independent of θ. The setting with online commentators thus becomes equivalent to the global games model of Section 2.2, with the following timeline:

1. The nature draws θ from a normal distribution $\mathcal{N}\left(\bar{\theta}, \sigma_\theta^2\right)$, which defines the initial common prior about the regime's strength, θ.

2. The government (publicly) offers a common contract to each commentator i: a payment of y per unit of praise b_i.

3. Each commentator i observes the regime's strength, θ, and based on that and on his (privately-known) intrinsic valuation for money, v_i, he chooses his level of praise for the government, b_i.

4. Each citizen i observes the commentator's level of praise, b_i, forms a posterior about the regime's strength, θ, and decides whether to attack the regime or not.

5. The regime is overthrown, or it survives. All players' payoffs are realised.

B.2 Omitted Proofs

Proof of Proposition 2.1

Proposition 2.1 is a standard result in global games, e.g., see Morris and Shin (2003). According to the payoff indifference condition, a citizen is indifferent between attacking the regime and not attacking it when she observes a private signal x^*. Hence, we must have $\Pr\left(attack\, successful \mid x^*\right) = c$, which—given the structure of the equilibrium—yields $\Pr\left(\theta \leq \theta^* \mid x^*\right) = c$. Given the assumed distribution of θ, the left-hand side is equal to $\Phi\left(\left(\theta^* - \mathbb{E}\left[\theta \mid x^*\right]\right) / \sqrt{Var\left(\theta \mid x^*\right)}\right)$. Applying the standard results of the normal learning model, a citizen's posterior expectation of the regime's strength, θ,

conditional on a given private signal, x, is

$$\mathbb{E}\left[\theta \mid x\right] = \overline{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\left(x - \bar{\theta}\right). \tag{2.22}$$

This is another way of stating (2.1). The higher σ_ε^2 is, the less informative the citizen's private signal is about θ, which makes the citizen put more weight on the mean strength of the regime, $\bar{\theta}$, and less on the private signal, x, when forming her posterior.

The conditional variance of θ given x is given by:

$$\begin{aligned} Var\left(\theta \mid x\right) &= \sigma_\theta^2\left(1 - \left(Corr\left(\theta, x\right)\right)^2\right) && (2.23)\\ &= \sigma_\theta^2\left(1 - \left(\frac{\sigma_\theta^2}{\sqrt{\sigma_\theta^2}\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}}\right)^2\right) && (2.24)\\ &= \frac{\sigma_\theta^2\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}. && (2.25) \end{aligned}$$

Adding in the results on conditional variance, the payoff indifference condition is:

$$c = \Phi\left(\sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2\sigma_\varepsilon^2}}\left(\theta^* - \bar{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\left(x^* - \bar{\theta}\right)\right)\right). \tag{2.26}$$

We can rearrange (2.26) to obtain the following expression for the "critical" value of the private signal, x^*:

$$x_{PI}^* = \frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}\theta^* - \frac{\sigma_\varepsilon^2}{\sigma_\theta^2}\bar{\theta} - \sqrt{\frac{\sigma_\varepsilon^2\left(\sigma_\theta^2 + \sigma_\varepsilon^2\right)}{\sigma_\theta^2}}\Phi^{-1}\left(c\right). \tag{2.27}$$

The critical mass condition states that, in equilibrium, whenever the true strength of strength is exactly θ^*, the proportion of citizens who receive a signal $x \leq x^*$ should

equal precisely θ^*:

$$\theta^* = \Pr\left(x \leq x^* \mid \theta^*\right), \tag{2.28}$$

where, given the distributional assumptions on θ, the right-hand side is:

$$\Pr\left(x \leq x^* \mid \theta^*\right) = \Phi\left(\frac{x^* - \mathbb{E}\left[x \mid \theta^*\right]}{\sqrt{Var\left(x \mid \theta^*\right)}}\right) \tag{2.29}$$

$$= \Phi\left(\frac{x^* - \theta^*}{\sqrt{\sigma_\varepsilon^2}}\right). \tag{2.30}$$

This leads us to the critical mass condition:

$$x^*_{CM} = \theta^* + \sqrt{\sigma_\varepsilon^2}\Phi^{-1}\left(\theta^*\right). \tag{2.31}$$

The two equilibrium conditions stated in (2.27) and (2.31) yield an equilibrium value of the critical state of the world, θ^*, as stated in Proposition 2.1.

As far as the statement on equilibrium uniqueness is concerned, by substituting (2.31) into (2.27), we obtain the following condition:[25]

$$U^{st}\left(\theta;\cdot\right) \equiv 1 - \Phi\left(\sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2\sigma_\varepsilon^2}}\left(\frac{\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\left(\bar{\theta} - \theta\right) + \frac{\sigma_\theta^2\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2 + \sigma_\varepsilon^2}\Phi^{-1}\left(\theta\right)\right)\right) - c$$

$$= 1 - \Phi\left(\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\bar{\theta} - \theta\right) + \Phi^{-1}\left(\theta\right)\right)\right) - c = 0. \tag{2.32}$$

By inspection, we observe that (i) $U^{st}\left(\theta;\cdot\right)$ is continuous and differentiable in $\theta \in (0,1)$, (ii) $\lim_{\theta\to 0} U^{st}\left(\theta;\cdot\right) = 1 - c > 0$, and (iii) $\lim_{\theta\to 1} U^{st}\left(\theta;\cdot\right) = -c < 0$. This implies that

[25] I follow here the steps of the proof for the equilibrium uniqueness condition in Angeletos, Hellwig and Pavan (2007).

a solution to $U^{st}(\theta;\cdot)=0$ always exists. Furthermore, note that

$$\frac{\partial U^{st}(\theta;\cdot)}{\partial\theta}=-\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2}}\phi(\cdot)\left(\frac{1}{\phi\left(\Phi^{-1}(\theta)\right)}-\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\right). \tag{2.33}$$

By the properties of the normal distribution, $\min_{\theta\in(0,1)}\frac{1}{\phi(\Phi^{-1}(\theta))}=\sqrt{2\pi}$. This means that a sufficient condition for $U^{st}(\theta;\cdot)$ to be monotonic in θ is that $\sqrt{2\pi}\geq\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}$, which can be rearranged to yield the statement on equilibrium uniqueness in Proposition 2.1.

Proof of Lemma 2.1

For part (i), note that by partially differentiating θ^* from (2.7) with respect to $\bar{\theta}$, we obtain:

$$\frac{\partial\theta^*}{\partial\bar{\theta}} \;=\; \frac{-\phi(\cdot)\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}}{1-\phi(\cdot)\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}}. \tag{2.34}$$

where $\phi(\cdot)=\phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^*-\bar{\theta}\right)-\sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}(c)\right)$. The numerator is negative. By the properties of the normal distribution, the maximum value of $\phi(\cdot)$ is $1/\sqrt{2\pi}$. The equilibrium uniqueness condition stated in Proposition 2.1 ensures that the denominator is positive. The probability of regime survival is $\Phi_{RC}=\Phi\left(\frac{\theta^*-\bar{\theta}}{\sqrt{\sigma_\theta^2}}\right)$. Hence, $\frac{\partial\theta^*}{\partial\theta}<0$ implies that $\frac{\partial\Phi_{RC}}{\partial\theta}=\phi\left(\frac{\theta^*-\bar{\theta}}{\sqrt{\sigma_\theta^2}}\right)\left(\frac{1}{\sqrt{\sigma_\theta^2}}\left(\frac{\partial\theta^*}{\partial\theta}-1\right)\right)<0$.

The steps for part (ii) are analogous. First, partially differentiate the expression for θ^* in (2.7) with respect to c:

$$\frac{\partial\theta^*}{\partial c}=\frac{-\phi(\cdot)\sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\frac{\partial\Phi^{-1}(c)}{\partial(c)}}{1-\phi(\cdot)\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}}, \tag{2.35}$$

where $\phi(\cdot)=\phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^*-\bar{\theta}\right)-\sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}(c)\right)$. The numerator is negative, whereas

the denominator is again positive by the properties of normal distribution and the condition for equilibrium uniqueness stated in Proposition 2.1. Furthermore, $\frac{\partial \theta^*}{\partial c} < 0$ implies that $\frac{\partial \Phi_{RC}}{\partial c} = \phi\left(\frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}}\right)\left(\frac{1}{\sqrt{\sigma_\theta^2}}\frac{\partial \theta^*}{\partial c}\right) < 0$.

Proof of Lemma 2.2

By partially differentiating the expression for θ^* in (2.7) with respect to σ_ε^2, we obtain:

$$\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2} = \frac{\phi(\cdot)\left(\frac{1}{2\sigma_\theta^2\sqrt{\sigma_\varepsilon^2}}\left(\theta^* - \bar{\theta}\right) - \frac{1}{2\sqrt{\sigma_\theta^2\left(\sigma_\theta^2+\sigma_\varepsilon^2\right)}}\Phi^{-1}(c)\right)}{1 - \phi(\cdot)\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}}, \tag{2.36}$$

where $\phi(\cdot) = \phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - \bar{\theta}\right) - \sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}(c)\right)$.

Note that by the properties of the normal distribution, the maximum value of $\phi(\cdot)$ is $1/\sqrt{2\pi}$. As long as the equilibrium uniqueness condition stated in Proposition 2.1 is satisfied, this ensures that the denominator is positive. However, the sign of the numerator is ambiguous and is strictly positive if and only if

$$\frac{1}{\sigma_\theta^2\sqrt{\sigma_\varepsilon^2}}\left(\theta^* - \bar{\theta}\right) - \frac{1}{\sqrt{\sigma_\theta^2\left(\sigma_\theta^2+\sigma_\varepsilon^2\right)}}\Phi^{-1}(c) > 0. \tag{2.37}$$

Given that $\Phi_{RC} = \Phi\left(\frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}}\right)$, and so $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < (>)\, 0$ if and only if $\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2} > (<)\, 0$, this can be rearranged to yield the condition in Lemma 2.2.

Proof of Proposition 2.2

Recall from Lemma 2.2 that the probability of regime change is decreasing in the variance of noise in the citizens' private signals, $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < 0$, if and only if:

$$\theta^* < T_{RC}, \tag{2.38}$$

where $T_{RC} = \bar{\theta} + \sqrt{\frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}} \Phi^{-1}(c)$.

Given that the regime finds it costly to increase σ_ε^2, it is optimal for the regime to do so only if $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < 0$ holds. To see why $\bar{\theta}$ must be high enough—for any given c—for the regime to have incentives to increase σ_ε^2, note that the left-hand side of (2.38) is decreasing in $\bar{\theta}$ (by Lemma 2.1), while the right-hand side is increasing in $\bar{\theta}$. Hence, a sufficiently high value of $\bar{\theta}$ ensures that the condition in (2.38) is satisfied. We can use an analogous argument to show that, for any given $\bar{\theta}$, the cost c must be high enough for $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < 0$ to hold.

It remains to be shown that the regime would not choose to increase σ_ε^2, i.e. $\frac{\partial U_R}{\partial \sigma_\varepsilon^2} \leq 0$, if $\bar{\theta}$ (or c) is exceedingly high. The key here is to use two properties of the normal distribution: (i) $\phi'(x) = x\phi(x)$ and (ii) $\lim_{x \to \pm\infty} \phi'(x) = 0$. The claim here is that:

$$\lim_{\bar{\theta} \to \infty} \frac{\partial \theta^*}{\partial \sigma_\varepsilon^2} = \lim_{\bar{\theta} \to \infty} \frac{\phi(\cdot) \left(\frac{1}{2\sigma_\theta^2 \sqrt{\sigma_\varepsilon^2}} \left(\theta^* - \bar{\theta} \right) - \frac{1}{2\sqrt{\sigma_\theta^2 \left(\sigma_\theta^2 + \sigma_\varepsilon^2 \right)}} \Phi^{-1}(c) \right)}{1 - \phi(\cdot) \frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}} = 0, \tag{2.39}$$

where $\phi(\cdot) = \phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2} \left(\theta^* - \bar{\theta} \right) - \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} \Phi^{-1}(c) \right)$, and an analogous statement holds for $\lim_{c \to 1} \frac{\partial \theta^*}{\partial \sigma_\varepsilon^2}$.

To see this, note that θ^* approaches 1 as $\bar{\theta} \to \infty$ (or as $c \to 1$), and the result in (2.39) then follows by applying the two properties of the normal distribution, $\phi'(x) = x\phi(x)$ and $\lim_{x \to \pm\infty} \phi'(x) = 0$. Since $\Phi_{RC} = \Phi\left(\frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}} \right)$, this implies that $\lim_{\bar{\theta} \to \infty} \frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} = 0$ and $\lim_{c \to 1} \frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} = 0$. Thus, given that $C'(\cdot) > 0$, the regime will never choose to increase σ_ε^2 if $\bar{\theta}$ (or c) takes a sufficiently high value.

Hence, for a given value of c, if λ is small enough (so that the role of costs in the regime's utility function is sufficiently small), there exists an interval of moderate values of $\bar{\theta}$ where it is optimal for the regime to increase σ_ε^2. Similarly, for a given value of $\bar{\theta}$, if λ is small enough, there exists an interval of moderate values of c where the regime

chooses to increase σ_ε^2.

Proof of Proposition 2.3

Note that

$$\frac{\partial \mathbb{E}\left[A\left(\theta^*\right)\right]}{\partial \sigma_\varepsilon^2} = \phi\left(\cdot\right)\left(\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2}\frac{\sqrt{\sigma_\theta^2+\sigma_\varepsilon^2}}{\sigma_\theta^2} + \left(\theta^* - \bar{\theta}\right)\frac{1}{2\sigma_\theta^2\sqrt{\sigma_\theta^2+\sigma_\varepsilon^2}} - \frac{1}{2\sqrt{\sigma_\theta^2\sigma_\varepsilon^2}}\Phi^{-1}\left(c\right)\right), \tag{2.40}$$

where $\phi\left(\cdot\right) = \phi\left(\left(\theta^* - \bar{\theta}\right)\frac{\sqrt{\sigma_\theta^2+\sigma_\varepsilon^2}}{\sigma_\theta^2} - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(c\right)\right)$. Hence, $\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2} > 0$ if and only if $\theta^* < T_A$, where T_A is given by:

$$T_A = \bar{\theta} + \sqrt{\frac{\sigma_\theta^2\left(\sigma_\theta^2+\sigma_\varepsilon^2\right)}{\sigma_\varepsilon^2}}\Phi^{-1}\left(c\right) - 2\left(\sigma_\theta^2+\sigma_\varepsilon^2\right)\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2}. \tag{2.41}$$

For part (i), note that:

$$\begin{aligned}
\lim_{\sigma_\varepsilon^2 \to 0}\theta^* &= \lim_{\sigma_\varepsilon^2 \to 0}\Phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - \bar{\theta}\right) - \sqrt{\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(c\right)\right) && (2.42)\\
&= \Phi\left(-\Phi^{-1}\left(c\right)\right) && (2.43)\\
&= 1-c. && (2.44)
\end{aligned}$$

First, consider the case where $c > \frac{1}{2}$ and $\bar{\theta} > 1-c = \lim_{\sigma_\varepsilon^2 \to 0}\theta^*$. Then $\Phi^{-1}\left(c\right) > 0$ and, by Lemma 2.2, $\partial\theta^*/\partial\sigma_\varepsilon^2 < 0$. This implies that both the second and the third term in the expression for T_A in (2.41) are positive, and so given $\theta^* < \bar{\theta}$, then it must also be the case that $\theta^* < T_A$, and hence $\partial \mathbb{E}\left[A\left(\theta^*\right)\right]/\partial\sigma_\varepsilon^2 < 0$. Thus, when $c > \frac{1}{2}$ and $\bar{\theta} > 1-c$, we observe that $\partial \mathbb{E}\left[A\left(\theta^*\right)\right]/\partial\sigma_\varepsilon^2 < 0$. An analogous argument holds for the case where $c < \frac{1}{2}$ and $\bar{\theta} < 1-c$, in which case we have $\partial \mathbb{E}\left[A\left(\theta^*\right)\right]/\partial\sigma_\varepsilon^2 > 0$.

Second, note that when $c > \frac{1}{2}$ and $\bar{\theta} < 1-c = \lim_{\sigma_\varepsilon^2 \to 0}\theta^*$, the second and the third term in the expression for T_A in (2.41) have opposing signs, and they both become

arbitrarily large in absolute terms as $\sigma_\varepsilon^2 \to 0$. However, we can see that the third term is second-order in the limit $\sigma_\varepsilon^2 \to 0$: the second term approaches $+\infty$ at rate $1/\sqrt{\sigma_\varepsilon^2}$, while the third term approaches $-\infty$ at rate $g\left(\sigma_\varepsilon^2\right)/\sqrt{\sigma_\varepsilon^2}$, where the numerator is decreasing as σ_ε^2 gets closer to 0. Hence, as $\sigma_\varepsilon^2 \to 0$, the second term dominates the third term, and we have that $\partial \mathbb{E}\left[A\left(\theta^*\right)\right]/\partial \sigma_\varepsilon^2 < 0$ as long as $\theta^* > \lim_{\sigma_\varepsilon^2 \to 0} T_A$, where $\lim_{\sigma_\varepsilon^2 \to 0} T_A > \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(c) > \lim_{\sigma_\varepsilon^2 \to 0} T_{RC}$. An analogous argument holds for the case where $c < \frac{1}{2}$ and $\bar{\theta} > 1 - c = \lim_{\sigma_\varepsilon^2 \to 0} \theta^*$.

For part (ii), note that as $\sigma_\varepsilon^2 \to \infty$, the second term in the expression for T_A goes to $\sqrt{\sigma_\theta^2}\Phi^{-1}(c)$, while the third term vanishes. To see the latter, use the properties of the normal distribution, $\phi'(x) = x\phi(x)$ and $\lim_{x \to \pm\infty} \phi'(x) = 0$, and note that when $\sigma_\varepsilon^2 \to \infty$, the third term can be rewritten as $k\phi\left(\sqrt{\sigma_\varepsilon^2}h\left(\sigma_\varepsilon^2\right)\right)\sqrt{\sigma_\varepsilon^2}h\left(\sigma_\varepsilon^2\right)$, where k is some constant and $\left|h\left(\sigma_\varepsilon^2\right)\right|$ is increasing in σ_ε^2. Consequently, $\lim_{\sigma_\varepsilon^2 \to \infty} T_A = \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(c)$. At the same time, we clearly also have $\lim_{\sigma_\varepsilon^2 \to \infty} T_{RC} = \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(c)$, since the second term in the expression for T_{RC} in (2.9) also approaches $\sqrt{\sigma_\theta^2}\Phi^{-1}(c)$ as $\sigma_\varepsilon^2 \to \infty$. Thus, we have $\lim_{\sigma_\varepsilon^2 \to \infty} T_A = \lim_{\sigma_\varepsilon^2 \to \infty} T_{RC} = \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(c)$.

For the very last statement in the proposition, first note that when $\bar{\theta} = \frac{1}{2}$ and $c = \frac{1}{2}$, then $\theta^* = \frac{1}{2}$. Furthermore, note that then $\Phi^{-1}(c) = 0$ and so $\partial\theta^*/\partial\sigma_\varepsilon^2 = 0$ (since $\theta^* = \frac{1}{2}$ and $T_{RC} = \bar{\theta} = \frac{1}{2}$). This implies in turn that $T_A = \frac{1}{2}$ as well, and hence we have $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_\varepsilon^2} = \frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \sigma_\varepsilon^2} = 0$.

Proof of Lemma 2.3

The probability of regime change in the perfect coordination benchmark with no information aggregation is $\Phi_{RC}^{PCN} = \Pr\left(x \leq x_{PCN}^*, \theta \leq 1\right)$. Since θ and $x = \theta + \varepsilon$ follow a bivariate normal distribution with correlation coefficient $\rho = \rho(\theta, x) = \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}$, we can use the properties of multivariate normal distributions.

Note from (2.12) that setting $\bar{\theta} = 1$ and $c = \frac{1}{2}$ implies that $x_{PCN}^* = \bar{\theta} = 1$.

Consider then variables $x_1 = (x - x^*_{PCN}) / \sqrt{\sigma^2_\theta + \sigma^2_\varepsilon} = (x-1) / \sqrt{\sigma^2_\theta + \sigma^2_\varepsilon}$ and $x_2 = \left(\theta - \bar{\theta}\right) / \sqrt{\sigma^2_\theta} = (\theta - 1) / \sqrt{\sigma^2_\theta}$. Variables x_1 and x_2 follow a standardised bivariate normal distribution, i.e. one with $(\mu_1, \mu_2) = (0,0)$ and $(\sigma_1, \sigma_2) = (1,1)$, and have a correlation coefficient $\rho = \sqrt{\frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_\varepsilon}}$.

The value of the CDF for a bivariate standard normal distribution is given by:

$$\Phi(a_1, a_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \exp\left(-\frac{x_1^2 - \rho x_1 x_2 + x_2^2}{2(1-\rho^2)} \right) dx_1 dx_2. \tag{2.45}$$

Sibuya (1960) and Sungur (1990) show that:

$$\frac{\partial \Phi(a_1, a_2; \rho)}{\partial \rho} = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{a_1^2 - 2\rho a_1 a_2 + a_2^2}{2(1-\rho^2)} \right). \tag{2.46}$$

Thus, we have $\frac{\partial}{\partial \rho}\Phi(a_1, a_2; \rho) = \phi(a_1, a_2; \rho)$. Since $\phi(a_1, a_2; \rho) > 0$, this implies that $\Phi(a_1, a_2; \rho)$ is strictly increasing in ρ for all a_1 and a_2. Because $\rho = \sqrt{\frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_\varepsilon}}$ is strictly decreasing in σ^2_ε, it follows that that $\Pr(x < x^*_{PCN}, \theta < 1)$ is strictly decreasing in σ^2_ε for all x^*_{PCN}.

We are interested here in the values of a_1 and a_2 for which $\frac{\partial}{\partial \rho}\Phi(a_1, a_2; \rho) = \phi(a_1, a_2; \rho)$ is maximised. To obtain these, we simply need to find

$$\arg\min_{a_1, a_2} g(a_1, a_2; \rho) = \frac{a_1^2}{2} - \rho a_1 a_2 + \frac{a_2^2}{2}. \tag{2.47}$$

Taking the first order condition with respect to a_1 and a_2 shows that the only critical point is $(a_1^*, a_2^*) = (0,0)$. Furthermore, $\frac{\partial^2}{\partial a_1^2} g(a_1^*, a_2^*) = \frac{\partial^2}{\partial a_2^2} g(a_1^*, a_2^*) = 1$ and $\frac{\partial^2}{\partial a_1 \partial a_2} g(a_1^*, a_2^*) = -\rho$. Since

$$\begin{aligned} D(a_1^*, a_2^*) &= \left(\frac{\partial^2}{\partial a_1^2} g(a_1^*, a_2^*) \right) \left(\frac{\partial^2}{\partial a_2^2} g(a_1^*, a_2^*) \right) - \left(\frac{\partial^2}{\partial a_1 \partial a_2} g(a_1^*, a_2^*) \right)^2 & (2.48) \\ &= 1 - \rho^2 > 0, & (2.49) \end{aligned}$$

and $\frac{\partial^2}{\partial a_1^2} g\left(a_1^*, a_2^*\right) > 0$ as well as $\frac{\partial^2}{\partial a_2^2} g\left(a_1^*, a_2^*\right) > 0$, it follows that $\left(a_1^*, a_2^*\right) = (0,0)$ is indeed the minimum of $g\left(a_1, a_2\right)$. Hence, $\frac{\partial}{\partial \rho}\Phi\left(a_1, a_2; \rho\right)$ is maximised by $\left(a_1^*, a_2^*\right) = (0,0)$ for all ρ.

The fact that with $\bar{\theta} = 1$ and $c = \frac{1}{2}$ we have $x_{PCN}^* = \bar{\theta} = 1$ for all σ_ε^2 implies that we can apply the above results to our model. Note that we can write

$$\begin{aligned} \Pr\left(x < x_{PCN}^*, \theta < 1\right) &= \Pr\left(\frac{x - \bar{\theta}}{\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}} < \frac{x_{PCN}^* - \bar{\theta}}{\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}}, \frac{\theta - \bar{\theta}}{\sqrt{\sigma_\theta^2}} < \frac{1 - \bar{\theta}}{\sqrt{\sigma_\theta^2}}\right) \quad (2.50) \\ &= \Pr\left(x_1 < a_1, x_2 < a_2\right), \quad (2.51) \end{aligned}$$

and $x_{PCN}^* - \bar{\theta} = 0$ and $\bar{\theta} - 1 = 0$ (so that $a_1 = 0$ and $a_2 = 0$) holds if and only if $\bar{\theta} = 1$ and $c = \frac{1}{2}$. Therefore, $\frac{\partial}{\partial \rho}\Pr\left(x < x_{PCN}^*, \theta < 1\right)$ is maximised for $\bar{\theta} = 1$ and $c = \frac{1}{2}$. Finally, since $\rho = \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}$ is strictly decreasing in σ_ε^2, we have that $\frac{\partial}{\partial \sigma_\varepsilon^2}\Pr\left(x < x_{PCN}^*, \theta < 1\right)$ is minimised by $\bar{\theta} = 1$ and $c = \frac{1}{2}$ for all levels of σ_ε^2 (recall that we have shown above that $\frac{\partial}{\partial \sigma_\varepsilon^2}\Pr\left(x < x_{PCN}^*, \theta < 1\right) < 0$).

Proof of Proposition 2.4

We start with the following lemma:

Lemma 2.4. *Suppose that $\bar{\theta} = 1$ and $c = \frac{1}{2}$. The probability of regime change in the perfect coordination benchmark with no information aggregation is:*

$$\Phi_{RC}^{PCN} = \frac{1}{4} + \frac{\arcsin\left(\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}\right)}{2\pi}. \quad (2.52)$$

Moreover, (i) $\frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} < 0$, (ii) $\lim_{\sigma_\varepsilon^2 \to 0} \frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} = -\infty$, and (iii) $\lim_{\sigma_\varepsilon^2 \to \infty} \frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} = 0$.

Proof. In general, there are no closed form solutions for the cumulative distribution function $F\left(\boldsymbol{x}\right) = \Pr\left(\mathbf{X} \leq \mathbf{x}\right)$, where $\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. There is, however, an analytical

solution for the standardised bivariate normal distribution in the special case where $\mathbf{x} = (0, 0)$. Setting $\bar{\theta} = 1$ and $c = \frac{1}{2}$ means that this property can be used in the model.

The quadrant probability for the bivariate standard normal distribution is given by:[26]

$$\begin{aligned} \Pr(x_1 \leq 0, x_2 \leq 0) &= \Pr(x_1 \geq 0, x_2 \geq 0) && (2.53) \\ &= \int_{-\infty}^{0} \int_{\infty}^{0} P(x_1, x_2)\, dx_1 dx_2 && (2.54) \\ &= \frac{1}{4} + \frac{\arcsin(\rho)}{2\pi}. && (2.55) \end{aligned}$$

From the proof of Lemma 2.3, we know that variables $x_1 = (x - 1)/\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}$ and $x_2 = (\theta - 1)/\sqrt{\sigma_\theta^2}$ follow a bivariate normal distribution with $(\mu_1, \mu_2) = (0, 0)$ and $(\sigma_1, \sigma_2) = (1, 1)$, and have a correlation coefficient $\rho = \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}$. Applying this to (2.55), we obtain the following:

$$\begin{aligned} \Phi_{RC}^{PCN} &= \Pr(x \leq x_{PCN}^*, \theta \leq 1) && (2.56) \\ &= \Pr(x_1 \leq 0, x_2 \leq 0) && (2.57) \\ &= \frac{1}{4} + \frac{\arcsin\left(\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}\right)}{2\pi}. && (2.58) \end{aligned}$$

For parts (i)-(iii), we can apply the property that $\frac{d}{dz}\arcsin(z) = 1/\sqrt{1 - z^2}$ to find the derivative of Φ_{RC}^{PCN} with respect to σ_ε^2 at $\bar{\theta} = 1$ and $c = \frac{1}{2}$:

$$\frac{\partial}{\partial \sigma_\varepsilon^2} \Phi_{RC}^{PCN} \mid_{\bar{\theta}=1, c=\frac{1}{2}} = -\frac{\sqrt{\sigma_\theta^2}}{4\pi(\sigma_\theta^2 + \sigma_\varepsilon^2)\sqrt{\sigma_\varepsilon^2}} < 0. \quad (2.59)$$

Furthermore, $\lim_{\sigma_\varepsilon^2 \to 0} \frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} = -\infty$ and $\lim_{\sigma_\varepsilon^2 \to \infty} \frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} = 0$. □

[26] See, e.g., Rose and Smith (1996, 2002) and Stuart and Ord (1998), p. 231.

Now, consider the statement in Proposition 2.4. Note from the proof of Lemma 2.4 that $\frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} < 0$ for all $\sigma_\varepsilon^2 > 0$. On the other hand, with $\bar{\theta} = 1$ and $c = \frac{1}{2}$, the probability of regime change in the global game (i.e. with imperfectly coordinated citizens) is $\Phi_{RC} = \Phi\left(\left(\theta^* - 1\right)/\sqrt{\sigma_\theta^2}\right)$, where $\theta^* = \Phi\left(\left(\sqrt{\sigma_\varepsilon^2}/\sigma_\theta^2\right)\left(\theta^* - 1\right)\right)$. Consequently, we have

$$\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} = \phi\left(\frac{\theta^* - 1}{\sqrt{\sigma_\theta^2}}\right)\frac{1}{\sqrt{\sigma_\theta^2}}\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2}, \tag{2.60}$$

where

$$\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2} = \phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - 1\right)\right)\frac{\theta^* - 1}{2\sigma_\theta^2\sqrt{\sigma_\varepsilon^2}\left(1 - \phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - 1\right)\right)\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\right)}. \tag{2.61}$$

This is also negative since $1 - \phi\left(\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^* - 1\right)\right)\frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2} > 0$, which is implied by the equilibrium uniqueness condition stated in Proposition 2.1: $\sigma_\varepsilon^2 \leq 2\pi\left(\sigma_\theta^2\right)^2$. Hence, when $\bar{\theta} = 1$ and $c = \frac{1}{2}$, we have $\frac{\partial \Phi_{RC}^{PCN}}{\partial \sigma_\varepsilon^2} < 0$ and $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < 0$ for all $\sigma_\varepsilon^2 > 0$.

To arrive at the result in Proposition 2.4, note that

$$\lim_{\sigma_\varepsilon^2 \to 0} \Phi_{RC}\left(\theta^*\right)|_{\bar{\theta}=1, c=\frac{1}{2}} = \Phi\left(\frac{-1}{2\sqrt{\sigma_\theta^2}}\right) \tag{2.62}$$

$$\lim_{\sigma_\varepsilon^2 \to \infty} \Phi_{RC}\left(\theta^*\right)|_{\bar{\theta}=1, c=\frac{1}{2}} = \Phi\left(\frac{-1}{\sqrt{\sigma_\theta^2}}\right) \tag{2.63}$$

$$\lim_{\sigma_\varepsilon^2 \to 0} \Phi_{RC}^{PCN} = \frac{1}{2} \tag{2.64}$$

$$\lim_{\sigma_\varepsilon^2 \to \infty} \Phi_{RC}^{PCN} = \frac{1}{4}. \tag{2.65}$$

where (2.64) and (2.65) follow from, respectively, the fact that $\arcsin\left(1\right) = \frac{\pi}{2}$ and $\arcsin\left(0\right) = 0$. Because $\Phi\left(-1/2\sqrt{\sigma_\theta^2}\right) < \frac{1}{2}$, for sufficiently small values of σ_ε^2 we observe that $\Phi_{RC}\left(\theta^*\right) < \Phi_{RC}^{PCN}$. Furthermore, if $\Phi\left(-1/\sqrt{\sigma_\theta^2}\right) > \frac{1}{4}$, which can be rearranged to $\sigma_\theta^2 > 1/\left(\Phi^{-1}\left(\frac{3}{4}\right)\right)^2$, then—since both $\Phi_{RC}\left(\theta^*\right)$ and Φ_{RC}^{PCN} are continuous and decreas-

ing in σ_{ε}^2—there exists a value of σ_{ε}^2, denoted by $\sigma_{\varepsilon,h}^2$, such that $\Phi_{RC}(\theta^*) > \Phi_{RC}^{PCN}$ when $\sigma_{\varepsilon}^2 > \sigma_{\varepsilon,h}^2$.

Chapter 3

Persuasion by Falsification of Evidence, and Measures Against It

3.1 Introduction

In recent decades, many interest groups have developed strategies to falsify scientific findings in order to defend their own agenda. Moreover, they have often done so while being aware that the scientific truth is against them. One well-known example is that of tobacco producers, who knew about the risks which their products pose already in the 1950s, but they consistently denied adverse effects of active smoking in the 1950s and 1960s and of second-hand smoke exposure from the 1970s until the 1990s. Their objective was to prevent any new regulation that would harm them or at least to delay it for as long as possible. It was only in 2009 that the U.S. Congress gave the authority to the Food and Drug Administration to regulate tobacco as an addictive drug.[1] There is also evidence of scientific findings being falsified to counter changes in regulations on issues such as climate change, lead in petrol, asbestos, insecticides, acid rain, ozone-layer

[1] http://www.wsj.com/articles/SB124474789599707175

depletion, eating meat, and many others.[2,3]

There are two noteworthy features of this phenomenon. First, it is costly for interest groups to falsify scientific findings. Falsification often requires developing an own research programme and publishing only favourable results or simply fabricating research studies with the desired results. One example is the Tobacco Institute created by the tobacco industry with the aim of fostering research but the primary objective was in fact to develop "a cadre of experts who could be called upon in time of need".[4] Philip Morris, one of the leading tobacco companies, was reported to have spent large amounts of money to fund scientists who would help the firm counter the emerging consensus on the health risks of second-hand tobacco smoke.[5,6]

Second, falsification of scientific findings may go undetected. People are often not able to identify whether a study has been done by objective academic researchers or by an interest group. This may be because they have insufficient knowledge or not enough time to analyse. In addition, interest groups make a great effort to render their research hardly distinguishable from that done by objective scientists. For example, it has been reported that the tobacco industry financed conferences and workshops, as well as created newsletters, magazines, and scientific journals where the results of industry-sponsored research could be reported, published, and cited, and an impression was made that these were independent (Proctor, 2012). There is even evidence of an interest group formatting an article to look like a reprint from the Proceedings of the

[2]The many parallels between these controversies are the subject of Oreskes and Conway (2010).

[3]See, for example:
http://www.theguardian.com/environment/planet-oz/2015/mar/05/doubt-over-climate-science-is-a-product-with-an-industry-behind-it;
http://www.theguardian.com/environment/2015/feb/27/what-happened-to-lobbyists-who-tried-reshape-us-view-climate-change

[4]Oreskes and Conway (2010), p. 244.

[5]https://www.washingtonpost.com/archive/politics/1997/05/09/philip-morris-sought-experts-to-cloud-issue-memo-details/17f42ecd-ec47-4046-b372-80ea4fb1157f/

[6]http://www.nytimes.com/2015/03/06/movies/review-merchants-of-doubt-separating-science-from-spin.html

National Academy of Sciences, which later featured in the Wall Street Journal.[7,8] The problem is aggravated further by the fact that media often do not inform their audiences that the "experts" being quoted have links to interest groups. This may be because they feel that doing so would be "editorialising" or they simply are not aware of the connections themselves (Oreskes and Conway, 2010).

In this paper, we draw upon the above observations to develop a theoretical model of communication where falsification is costly for the sender and may be not detected by the decision maker. There are two players, a sender and a decision maker, and the state of the world is binary, either high or low. The decision maker chooses between two actions—accepting or rejecting—and would like to accept when the state is high, and to reject when the state is low. The sender, on the other hand, would like the decision maker always to accept, regardless of the state of the world. This illustrates a setting in which, for example, a policy maker would like to adjust its policy on tobacco products to the true scientific evidence on whether or not smoking has adverse effects on human health. At the same time, an interest group such as a tobacco company would always like this policy to be lenient.

The sender first observes the state of the world, interpreted as the scientific evidence on an issue. He then decides whether to falsify it at a cost. If he does falsify, then—with some predetermined probability—the falsification is successful: the decision maker sees the falsified evidence rather than the true scientific evidence. If it is unsuccessful or the sender does not falsify, the decision maker sees the true scientific evidence. The decision maker then takes her action. Thus, the setup captures the fact that falsification is costly and may be not detected by people—who are, however, still sophisticated enough to be aware of the possibility of falsification. The fact that only the sender observes the true state of the world captures the observation that an interest group is often better

[7]http://science.sciencemag.org/content/280/5361/195.1.full
[8]http://www.wsj.com/articles/SB881189526293285000

informed about the scientific evidence on the issue.[9]

For limiting cases of the cost and the undetectability of falsification, the model reconciles the predictions of the three canonical models of communication—in particular, the decision maker's welfare and the probability that the decision maker accepts conditional on the state of the world. As undetectability approaches zero, these predictions are the same as in a model of verifiable disclosure (Milgrom, 1981; Grossman, 1981). As cost approaches zero and undetectability approaches zero, they are the same as in a model of cheap talk (Crawford and Sobel, 1982). As cost and undetectability both approach one, they are the same as in a model of Bayesian persuasion (Kamenica and Gentzkow, 2011).

We investigate two possible measures the decision maker could take to improve her chances of taking the action that matches the state of the world. The first one is to acquire additional independent information about the issue, for example, a policy maker could conduct an independent study on the impact of smoking on health. The second one is to control how much information to acquire from the sender, e.g., a policy maker could choose how much time to spend on meetings with lobbyists.

Acquiring additional independent information has two important implications. First, it naturally provides the decision maker with extra information about the state of the world. Second, more interestingly, it also affects the sender's incentives to falsify the scientific evidence. This is because, when deciding whether to falsify, the sender compares the cost of doing so with the expected benefit, which depends on how likely the falsified evidence is to change the decision maker's action. For example, it may be that the sender's information on its own cannot possibly persuade the decision maker to accept, but it can persuade her if the decision maker's independent information also turns out to be favourable to the sender. In that case, if the cost of falsifying is low and the sender

[9]We later relax the assumption that the sender's signal perfectly reveals the true state of the world and we instead assume the signal is noisy—but still more informative than that of the decision maker.

expects the decision maker's independent information to be indeed favourable, then his incentives to falsify can be boosted by the presence of the independent evidence.

We demonstrate that the impact of acquiring independent evidence on the sender's incentives to falsify is in general ambiguous and depends on a number of factors. The sender's incentives to falsify are strengthened when (i) falsification is cheap, (ii) detecting it is difficult, (iii) the quality of the decision maker's independent evidence is low, and (iv) the prior belief is in favour of the low state. The increased falsification diminishes the value of the independent evidence, and—in the extreme—can even fully offset the informational benefit it brings. On the other hand, the sender's incentives to falsify are dampened when (i) falsification is expensive, (ii) detecting it is easy, (iii) the quality of the decision maker's independent evidence is high, and (iv) the prior belief is not too much in favour of the high state. In this case, since there is a side effect of less falsification, the value of the independent evidence exceeds the mere informational benefit it brings. Despite its ambiguous effect on falsification by the sender, acquiring private independent evidence always makes the decision maker weakly better off in our model.

Our analysis proceeds by comparing the decision maker's benefits from acquiring private independent evidence (i.e. observed only by her) and public independent evidence (i.e. observed by both players). The aim here is to shed light on policy makers' benefits from conducting public independent research, e.g., by establishing independent research institutes. We show that the decision maker is weakly better off with private independent evidence. The preference for private independent evidence is strict when the quality of independent research is sufficiently high and the prior belief is not too much in favour of the high state. Otherwise, acquiring private and public is equally good for the decision maker—despite the fact that, under some conditions, acquiring public evidence actually leads to less falsification than acquiring private evidence.

We make further observations by relaxing the assumption that the sender observes a perfect signal of the true state of the world. If the sender instead observes a noisy signal of the state, then it becomes possible that acquiring private independent evidence completely deters falsification by the sender. Moreover, we demonstrate that an increase in the quality of the decision maker's independent evidence can actually decrease her welfare.

The second measure which the decision maker could take is to control how much attention she devotes to the sender. By committing to ignore the sender with some positive probability, the decision maker can lower the chances that the sender's message will be pivotal to her decision, and this may in turn dampen his incentives to falsify scientific evidence. At the same time, the negative consequence of ignoring the sender is naturally that she has no information other than the prior when making the decision. In our model, the decision maker's optimal strategy turns out to be either to pay full attention to the sender or to commit to ignore him just often enough to completely discourage him from falsifying evidence.

We identify the circumstances under which it is optimal for the decision maker to commit to ignore the sender with a positive probability. The impact of detectability of falsification on the decision maker's incentives to commit to ignore the sender is twofold. One channel is that, as falsification becomes more difficult to detect, the decision maker's benefit from committing to ignore (and thus deterring falsification altogether) increases. On the other hand, this also drives down the sender's relative cost of falsification (i.e. the cost of falsification relative to the probability of being undetected), and therefore the level of ignorance required to disincentivise falsification goes up and makes commitment to ignorance less attractive for the decision maker. We show that, as falsification becomes very difficult to detect, the former effect dominates and commitment to ignorance is optimal for the decision maker regardless of the cost

of falsification.

Regarding the cost of falsification, commitment to ignorance is optimal when this cost takes moderate values. When it is very high, the sender has no incentive to falsify evidence and therefore there is no benefit to the decision maker from committing to ignore him. When the cost is very small, however, the decision maker needs to commit to ignore the sender with a very high probability if she wants to disincentivise falsification. This makes commitment to ignorance prohibitively costly to the decision maker.

Lastly, we discuss how the decision maker's incentives to acquire an additional private signal interact with her incentives to commit to ignore the sender's message.

The paper is organised as follows. The rest of this section discusses the related literature. Section 3.2 presents the main model and analyses its equilibrium. Section 3.3 investigates the decision maker's incentives to acquire private independent evidence, while Section 3.4 studies her incentives to acquire information from the sender. Section 3.5 concludes.

Related Literature

Although the focus of our paper is on falsification of evidence in the context of interest groups and policy makers, the model can be reinterpreted more generally as one about lying, and hence it naturally belongs to the literature on strategic communication. In cheap talk models (Crawford and Sobel, 1982), the sender can lie arbitrarily without any direct costs and the receiver cannot detect lying. In verifiable disclosure games (Grossman, 1981; Milgrom, 1981), the sender can withhold information but cannot lie, which could be because lying is prohibitively costly for the sender and/or the receiver can always detect it. In our model, we could say that the sender decides whether to tell the truth, which is costless, or to lie, which is costly. The decision maker is able to detect lying with some predetermined probability, and if she does detect a lie, she

immediately observes the true information of the sender.

There are a few theoretical papers which study the impact of the cost and the detectability of lying on communication. Dziuda and Salas (2017) and Balbuzanov (2017) focus on settings where lying has no exogenous cost to the sender. In both papers, it is assumed that the receiver observes the sender's message, and whenever the message is a lie, she observes with positive probability an additional message which says that the sender is lying. In this setting, Dziuda and Salas (2017) focus on equilibria where all lies are treated equally, i.e. the receiver is not allowed to take different actions upon seeing different lies. Thus, effectively, if the receiver detects a lie, she acts as if she only saw that the sender's message is inconsistent. On the other hand, in our model, when the receiver detects a lie, she observes the true information possessed by the sender. Thus, the type of lies which our model analyses is different from those in Dziuda and Salas (2017) and Balbuzanov (2017). Our model is more suited to a particular type of lies which we could call "obvious lies", i.e. those where the receiver can easily determine the truth whenever she detects a lie.[10]

In Dziuda and Salas (2017), like in our model, the receiver wants to take an action that matches the state, whereas the sender wants the receiver to take the highest possible action. She shows that, in any informative equilibrium, the sender types can be divided into three distinct groups. The lowest types lie and pretend to be the highest types. The intermediate and the highest types do not lie; however, the messages of the latter are discounted to a degree because of the untruthful messages sent by the lowest types. Balbuzanov (2017) makes a different assumption about the preferences of the

[10]For example, Donald Trump claimed he had "the biggest electoral college win since Ronald Reagan", while Sean Spicer (White House Press Secretary, January-July 2017) said about Donald Trump's inauguration that "This was the largest audience ever to witness an inauguration, period. Both in person and around the globe." Both statements could have easily been checked to be false:
https://www.nytimes.com/interactive/2016/12/18/us/elections/donald-trump-electoral-college-popular-vote.html?mcubz=0&_r=0;
https://www.theguardian.com/us-news/2017/jan/22/trump-inauguration-crowd-sean-spicers-claims-versus-the-evidence

sender and the receiver in that they are partially aligned. He shows that, in his setting, fully revealing equilibria are possible, and that this remains true even for low levels of detectability of lies.

Kartik, Ottaviani and Squintani (2007) and Kartik (2009) assume that lying has an exogenous cost but the receiver is unable to detect lying. Kartik et al. (2007) analyse a model of strategic communication between an informed sender and one or more uninformed receivers, where the sender has an upward bias. The message sent by the sender directly affects his payoff, which can be interpreted as a consequence of lying being costly or the receiver possibly being naive. Kartik et al. (2007) show that, under broad conditions, there exist separating equilibria with inflated communication, provided that the state space is unbounded above. Kartik (2009) assumes instead that the sender types are bounded above (as well as below). For any given type, it is cheapest for the sender to tell the truth, and as the magnitude of the lie increases, so does the marginal cost of a bigger lie. Kartik (2009) shows that this setting leads to an inflated language and equilibria have the following structure: low types of the sender use inflated language but they still separate, i.e. reveal their types through their equilibrium message, while the high types segment into one or more pools.

Our paper is also related to the literature on manipulation of information by interest groups. Bramoullé and Orset (2017) model how interest groups "manufacture doubt" about scientific research. In their model, firms can, at a cost, fabricate evidence in their favour in order to influence the citizens' beliefs about pollution caused by the firms. Additional evidence about this pollution is provided by scientists. The citizens' beliefs are then taken into account by the government when deciding on the regulation of pollution. Bramoullé and Orset (2017) study the firms' optimal amount of fabrication. The key difference is that they assume that the citizens naively treat the fabricated evidence as independent scientific evidence, while in our paper the receiver is more

sophisticated: she cannot verify whether the sender's message is falsified, but is aware of the sender's incentives to do so.

Shapiro (2016) models how media communicate policy-relevant information to a voter when (i) interest groups can make claims of fact that appear credible to the voter, and (ii) journalists have reputational concerns for appearing neutral. He shows that these two frictions can interact to prevent the voter from learning useful facts. Stone (2011) analyses a simple model of policy making in the presence of an interest group that chooses strategically whether to fund and lobby research. In his model, if research is not lobbied, it enters the public domain and can still be randomly observed by the policy maker. The main result is that, for a range of parameter values, the interest group sometimes funds but never lobbies research. This behaviour effectively "jams" the public signal of the policy maker, making the policy choice worse on average.

3.2 Baseline Model

There are two players: a sender and a decision maker (DM). The state of the world is binary, $\theta \in \{\theta_L, \theta_H\}$, and the prior belief that the state is $\theta = \theta_H$ is $p \in (0, 1)$. The DM chooses between two actions: she can either accept (denoted by $a = A$) or reject (denoted by $a = R$). She receives a payoff of $v(a, \theta) = 1$ if $\theta = \theta_H$ and $a = A$, and if $\theta = \theta_L$ and $a = R$; she receives $v(a, \theta) = 0$ otherwise. In other words, the DM prefers to accept when the state of the world is "high" and to reject when the state is "low". The sender receives a payoff of $u(a) = 1$ if $a = A$ and $u(a) = 0$ if $a = R$. Thus, he wants the DM to accept regardless of the state of the world. Both the DM and the sender are expected payoff maximisers.

The game proceeds as follows. First, the sender privately observes a perfectly infor-

mative signal of the state of the world: $s = i$ if $\theta = \theta_i$, for $i \in \{L, H\}$.[11] He then sends a message $m \in \{L, H\}$ to the DM. We will say that the sender does not falsify evidence if $m = s$, and that he falsifies evidence if $m \neq s$. The message received by the DM is denoted by $\widetilde{m}$. If the sender does not falsify, then the DM receives $\widetilde{m} = m = s$. If the sender falsifies, then with probability $x \in (0, 1]$ the DM receives $\widetilde{m} = m$ (i.e. falsification is successful), and with probability $1 - x$ the DM receives $\widetilde{m} = s$ (i.e. falsification is unsuccessful). The parameter x can be interpreted as a measure of undetectability of falsification, or a measure of the sender's skill in falsification. If the sender falsifies, he pays an exogenous cost $c \in (0, x)$.[12]

After receiving $\widetilde{m}$, the DM forms a belief β about the state of the world and chooses an action, $a \in \{A, R\}$. The DM's posterior belief about the state of the world upon observing $\widetilde{m}$ is denoted by $\beta(\widetilde{m}) = \Pr(\theta = \theta_H \mid \widetilde{m})$. Given the payoff function of the DM, it is straightforward to note that (i) if $\beta(\widetilde{m}) > \frac{1}{2}$, the DM accepts, (ii) if $\beta(\widetilde{m}) < \frac{1}{2}$, the DM rejects, and (iii) if $\beta(\widetilde{m}) = \frac{1}{2}$, then the DM is indifferent between accepting and rejecting.

The sender's strategy, σ, describes the probability with which he falsifies, with mixed strategies being possible. It is a function

$$\sigma : \{L, H\} \to \Delta\{0, 1\}. \tag{3.1}$$

We will denote by σ_s the probability with which the sender falsifies when he observes s.

The DM's strategy, δ, describes the probability with which she accepts, with mixed

[11] More generally, we could assume that the sender privately observes a realisation of a noisy signal of the state of the world, $s \in \{L, H\}$. The conditional probability that the signal correctly reveals the true state is $\Pr(s = i \mid \theta = \theta_i) = q$ for $i \in \{L, H\}$ with $q \in \left(\frac{1}{2}, 1\right]$, where q is the *quality* of the signal. Here, we assume $q = 1$, but we relax this assumption in Section 2.5 to show additional results.

[12] The sender would never have an incentive to falsify if it were that $c > x$.

strategies possible. It is a function

$$\delta : \{L, H\} \to \Delta \{A, R\} . \tag{3.2}$$

We will denote by $\delta_{\widetilde{m}}$ the probability with which the DM accepts when she observes $\widetilde{m}$.

The model is intentionally simple. Our objective is to shed light on how communication is affected by the presence of both (i) an exogenous cost of falsification for the sender and (ii) a possibility that falsification is not detected by the decision maker. Our paper introduces these two features in arguably the simplest possible setting.

3.2.1 Equilibrium

The solution concept is the perfect Bayesian equilibrium, which is defined in the usual way: a triple $(\sigma^*, \delta^*, \beta^*)$ is a perfect Bayesian equilibrium if (i) σ_s^* maximises the expected value of $u(a)$ given the DM's strategy δ^* and the belief β^*, (ii) $\delta_{\widetilde{m}}^*$ maximises the expected value of $v(a, \theta)$ given the sender's strategy σ^* and the belief β^*, and (iii) β^* is derived using $\sigma^*(s)$ by Bayes' rule from the strategy of the sender and the prior distribution over $\{\theta_H, \theta_L\}$, whenever it is possible. Henceforth, we refer to the perfect Bayesian equilibrium simply as an equilibrium. With the exception of knife-edge cases, the game has a unique equilibrium, which we state formally in the following proposition.[13]

Proposition 3.1. *The game has a unique equilibrium, except for knife-edge cases. The sender's and the DM's equilibrium strategies are:*

(i) if $p \in \left(0, \frac{x}{1+x}\right)$, *then* $(\sigma_L^*, \sigma_H^*) = \left(\frac{p}{(1-p)x}, 0\right)$ *and* $(\delta_L^*, \delta_H^*) = \left(0, \frac{c}{x}\right)$;

(ii) if $p \in \left[\frac{x}{1+x}, 1\right)$, *then* $(\sigma_L^*, \sigma_H^*) = (1, 0)$ *and* $(\delta_L^*, \delta_H^*) = (0, 1)$.

Let $\beta(\widetilde{m} \mid \sigma_L)$ denote the DM's belief upon observing $\widetilde{m}$ when the sender's strategy

[13] In knife-edge cases, we select the sender-preferred strategy profile whenever there are multiple equilibria.

is (σ_L, σ_H), noting that the sender never has an incentive to falsify evidence upon observing $s = H$, and hence in equilibrium it must always be that $\sigma_H^* = 0$. For brevity, whenever we say that the sender's equilibrium strategy is to falsify with some probability, we mean that he falsifies with this probability upon observing $s = L$ but does not falsify upon observing $s = H$.

When prior belief p is sufficiently low, $p < \frac{x}{1+x}$, the players play mixed strategies in equilibrium. The condition $p < \frac{x}{1+x}$ can be rewritten as $\beta(\widetilde{m} = H \mid \sigma_L = 1) < \frac{1}{2}$. This condition and $\beta(\widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$ together imply that players cannot play pure strategies in equilibrium. If the sender's strategy were to falsify with probability zero, the DM's optimal response would be to accept with probability 1 upon receiving $\widetilde{m} = H$, which would create incentives for the sender to falsify. But if the sender falsified with probability 1, then the DM would not accept upon receiving $\widetilde{m} = H$, which would erase the sender's incentives to falsify. Thus, there is instead a mixed-strategy equilibrium in which the sender is indifferent between falsifying and not falsifying when he observes $s = L$, while the DM is indifferent between accepting and rejecting when she receives $\widetilde{m} = H$. The sender's equilibrium strategy σ_L^* satisfies the condition for the DM's indifference:

$$\beta(\widetilde{m} = H \mid \sigma_L = \sigma_L^*) = \frac{p}{p + (1-p)\,\sigma_L^* x} = \frac{1}{2}, \tag{3.3}$$

which is equivalent to $\sigma_L^* = \frac{p}{(1-p)x}$. The DM's equilibrium strategy δ_H^* satisfies the condition for the sender's indifference: $\delta_H^* x = c$.

When prior belief p is sufficiently high, $p \geq \frac{x}{1+x}$, the sender falsifies with probability 1, while the DM accepts upon receiving $\widetilde{m} = H$. The condition that $p > \frac{x}{1+x}$ is equivalent to $\beta(\widetilde{m} = H \mid \sigma_L = 1) > \frac{1}{2}$, which implies that the DM's optimal strategy is to accept upon receiving $\widetilde{m} = H$ even when the sender falsifies with probability 1.

It follows from Proposition 3.1 that the probability that the sender falsifies evidence upon observing $s = L$ is strictly increasing in p for $p \in \left(0, \frac{x}{1+x}\right)$ and equals 1 for

$p \in \left[\frac{x}{1+x}, 1\right)$. The probability that the DM accepts upon observing $\widetilde{m} = H$ is weakly increasing in p: it equals $\frac{c}{x}$ for $p \in \left(0, \frac{x}{1+x}\right)$ and 1 for $p \in \left[\frac{x}{1+x}, 1\right)$.

3.2.2 Decision Maker's Welfare

The DM's ex-ante expected payoff in equilibrium is

$$\mathbb{E}\left(v\left(\sigma^*, \delta^*\right)\right) = \begin{cases} 1-p & \text{for } p \in \left(0, \frac{x}{1+x}\right) \\ 1-(1-p)\,x & \text{for } p \in \left[\frac{x}{1+x}, 1\right) \end{cases}. \tag{3.4}$$

When $p < \frac{x}{1+x}$, the DM is indifferent between the two actions when she observes a message $\widetilde{m} = H$, and she rejects otherwise. Thus, her expected payoff would remain the same if she rejected regardless of the message observed. Hence, her expected payoff is $1-p$, which is as if she received a completely uninformative message. On the other hand, when $p \geq \frac{x}{1+x}$, the DM's equilibrium strategy is such that a message $\widetilde{m} = H$ prompts her to accept and a message $\widetilde{m} = L$ prompts her to reject. Since the sender never falsifies when he observes $\theta = \theta_H$, the DM takes an incorrect action only when the state is $\theta = \theta_L$ and the sender successfully falsifies. The DM's expected payoff is then $\mathbb{E}\left(v\left(\sigma^*, \delta^*\right)\right) = 1-(1-p)\,x$.

Thus, the DM's expected payoff is non-monotonic in prior belief p: it is strictly decreasing in p for $p < \frac{x}{1+x}$ and is strictly increasing in p for $p \geq \frac{x}{1+x}$. There are two effects here which lead to this U-shaped relationship. First, when $p < \frac{x}{1+x}$, i.e. when the DM's welfare is the same as if she relied only on her prior, a value of p that is further away from 0 or 1 means that there is more uncertainty about the state of the world and hence the action that the DM takes based on the prior is less likely to match the true state. Second, when $p \geq \frac{x}{1+x}$, i.e. when the DM follows the message received from the sender, a higher value of p means that the sender is ex ante less likely to observe

$s = L$, and hence ex ante less likely to falsify.

Furthermore, the DM's expected payoff is decreasing in the undetectability of falsification, x. Intuitively, as the sender becomes more capable at corrupting the signal about the state of the world, the DM becomes less likely to take an action that matches the true state.

3.2.3 Other Models of Strategic Communication

Although the structure of the model is simple, it is worth noting that the limiting cases with respect to c and x reconcile the equilibrium predictions (in particular, the DM's welfare and the probability of the DM accepting conditional on the state of the world) of three canonical models of strategic communication: verifiable disclosure, cheap talk, and Bayesian persuasion.

When $x \to 0$, the equilibrium approaches $(\sigma_L^*, \sigma_H^*) = (1, 0)$ and $(\delta_L^*, \delta_H^*) = (0, 1)$ for $p \in (0, 1)$. Even though the sender falsifies with probability 1 upon observing $s = L$, the falsification is detectable with probability approaching 1, and so the DM learns the state of the world. Hence, she takes the same action that she would take if she knew the state of the world. This is the same outcome as in a model of verifiable disclosure (Milgrom, 1981; Grossman, 1981). However, it should be pointed out that the mechanism leading to this outcome is different. In verifiable disclosure, the sender cannot report false information to the DM but can suppress information and—by the unravelling result—he fully reveals the information in equilibrium. In contrast, in our model, the sender falsifies in equilibrium but full information revelation is achieved because the falsification is effectively always detected by the DM.[14]

When $x \to 1$ and $c \to 0$, the equilibrium approaches $(\sigma_L^*, \sigma_H^*) = \left(\frac{p}{1-p}, 0\right)$ and

[14]In addition, the outcome in our model for $c > x$ is trivially the same as in a model of verifiable disclosure: since the cost falsification exceeds the potential benefit, there is no falsification at all, and thus full information revelation is achieved in equilibrium.

$(\delta_L^*, \delta_H^*) = (0,0)$ for $p \in \left(0, \frac{1}{2}\right)$, and $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_L^*, \delta_H^*) = (0,1)$ for $p \in \left[\frac{1}{2}, 1\right)$. Thus, when $p < \frac{1}{2}$, the DM effectively always rejects in equilibrium, even upon receiving $\widetilde{m} = H$. When $p \geq \frac{1}{2}$, the DM's equilibrium strategy is to follow the message $\widetilde{m}$ received from the sender, but since $x \to 1$, this message is effectively always $\widetilde{m} = H$ and thus the DM effectively always accepts in equilibrium. In both cases, the DM takes the same action that she would take if she relied only on her prior. This is the same outcome as in a model of cheap talk (Crawford and Sobel, 1982). Yet, again, the mechanism leading to this outcome is different in our model. In cheap talk, the sender sends a costless but unverifiable report to the DM about the state of the world. With no common interest between the sender and the DM, a "babbling" equilibrium arises, where the sender sends an uninformative message while the DM ignores the sender's message and makes a decision based on her prior belief. In contrast, here, when $p < \frac{1}{2}$, the sender's message is actually informative but the DM effectively ignores it in equilibrium, whereas when $p \geq \frac{1}{2}$, the DM does not ignore the message but, due to undetectability, he effectively always receives the same message, which makes it uninformative.

Finally, when $x \to 1$ and $c \to 1$, the equilibrium approaches $(\sigma_L^*, \sigma_H^*) = \left(\frac{p}{1-p}, 0\right)$ and $(\delta_L^*, \delta_H^*) = (0,1)$ for $p \in \left(0, \frac{1}{2}\right)$, and $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_L^*, \delta_H^*) = (0,1)$ for $p \in \left[\frac{1}{2}, 1\right)$. For both $p < \frac{1}{2}$ and $p \geq \frac{1}{2}$, the equilibrium is such that the sender falsifies upon observing $s = L$ with a probability that makes the DM accept with probability 1 upon receiving $\widetilde{m} = H$. This is the same outcome as in a model of Bayesian persuasion (Kamenica and Gentzkow, 2011). One should note, however, that the mechanism in our model is substantially different. In Bayesian persuasion, the sender—before privately observing the state of the world—designs an experiment, i.e. a mapping from each possible state of the world to a distribution over possible signal realisations, and publicly commits to it. In our model, the strategies are determined through equilibrium play

without any commitment.[15]

We summarise the above results in Proposition 3.2, and illustrate them in Figure 3.1.

Proposition 3.2. *The DM's expected payoff, as well as the probability of her accepting conditional on the state of the world, approach:*

(i) their equivalents in a model of verifiable disclosure when $x \to 0$*;*

(ii) their equivalents in a model of cheap talk when $x \to 1$ *and* $c \to 0$*;*

(iii) their equivalents in a model of Bayesian persuasion when $x \to 1$ *and* $c \to 1$*.*

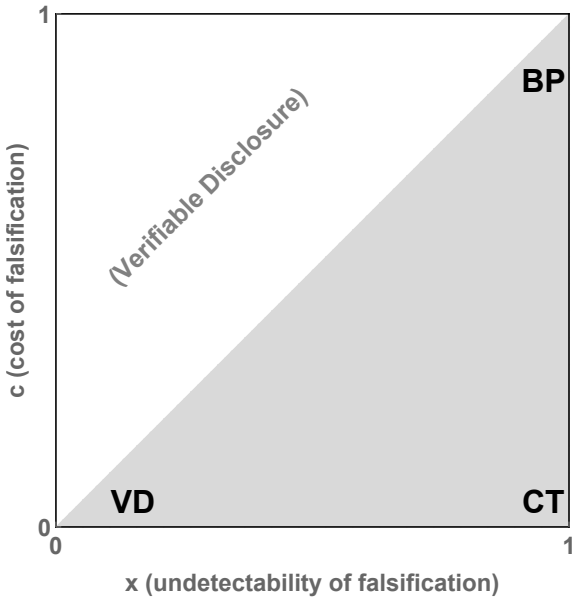


Figure 3.1: The relationship between the predictions of our model (in particular, the DM's welfare and the probability of that the DM accepts conditional on the state of the world) and their equivalents in models of verifiable disclosure (VD), cheap talk (CT), and Bayesian persuasion (BP). The shaded area illustrates the condition $c < x$.

[15]This relationship suggests that there might be some scope for relaxing the commitment assumption in Bayesian persuasion by introducing costly and detectable manipulation of information; however, we leave this for future research.

3.3 Acquisition of Private Independent Evidence

In this section, we analyse a scenario in which the DM receives—in addition to the message from the sender—a private signal about the state of the world. The aim of the analysis is to provide insight into policy makers' incentives to conduct private independent research, e.g., by consulting independent experts or hiring independent advisers.

The game now proceeds as follows. As before, the sender privately observes a perfectly informative signal of the state of the world: $s = i$ if $\theta = \theta_i$, for $i \in \{L, H\}$. However, unlike in the baseline model, the DM now observes a realisation of a private signal of the state of the world. The fact that the DM receives the private signal is public knowledge. We denote the DM's signal realisation by $s_{DM} \in \{L, H\}$. The DM's signal is assumed to be of quality $q_{DM} \in \left(\frac{1}{2}, 1\right)$, where the quality describes the conditional probability $\Pr\left(s_{DM} = i \mid \theta = \theta_i\right) = q_{DM}$.[16] After observing s, the sender sends a message $m \in \{H, L\}$ to the DM. If the sender does not falsify, i.e. he sends $m = s$, then the DM receives $\widetilde{m} = m = s$. If the sender falsifies, i.e. he sends $m \neq s$, then with probability $x \in [0, 1]$ the DM receives $\widetilde{m} = m$, and with probability $1 - x$ the DM receives $\widetilde{m} = s$. Thus, overall, the information observed by the DM is now a pair $(s_{DM}, \widetilde{m}) \in \{(L, L), (L, H), (H, L), (H, H)\}$. After observing $(s_{DM}, \widetilde{m})$, the DM forms a belief β about the state of the world and chooses an action, $a \in \{A, R\}$.

The strategy of the sender is still denoted by (σ_L, σ_H). For the DM's strategy, we will now use $\delta_{s_{DM}\widetilde{m}}$ to denote the probability with which the DM accepts when she observes $(s_{DM}, \widetilde{m})$. The DM's strategy is thus given by $(\delta_{LL}, \delta_{LH}, \delta_{HL}, \delta_{HH})$.

[16] Equivalently, we could think of the baseline model as of one in which the DM's private signal is uninformative, i.e. $q_{DM} = \frac{1}{2}$.

3.3.1 Impact of Private Independent Evidence on Equilibrium

The following proposition characterises the equilibrium in this setup:[17]

Proposition 3.3. *Suppose that, in addition to receiving a message from the sender, the DM observes a noisy private signal of quality* $q_{DM} \in \left(\frac{1}{2}, 1\right)$. *Except for knife-edge cases, there is a unique equilibrium. The equilibrium strategies of the sender and the DM satisfy* $\sigma_H^* = 0$ *and* $\delta_{LL}^* = \delta_{HL}^* = 0$, *and furthermore:*

(i) if $p^- \geq \frac{x}{1+x}$, *then* $\sigma_L^* = 1$ *and* $(\delta_{LH}^*, \delta_{HH}^*) = (1, 1)$;

(ii) if $p^- < \frac{x}{1+x} \leq p^+$ *and* $q_{DM} < 1 - \frac{c}{x}$, *then* $\sigma_L^* = 1$ *and* $(\delta_{LH}^*, \delta_{HH}^*) = (0, 1)$;

(iii) if $p^+ < \frac{x}{1+x}$ *and* $q_{DM} < 1 - \frac{c}{x}$, *then* $\sigma_L^* = \frac{p^+}{(1-p^+)x}$ *and* $(\delta_{LH}^*, \delta_{HH}^*) = \left(0, \frac{c}{(1-q_{DM})x}\right)$;

(iv) if $p^- < \frac{x}{1+x}$ *and* $q_{DM} \geq 1 - \frac{c}{x}$, *then* $\sigma_L^* = \frac{p^-}{(1-p^-)x}$ *and* $(\delta_{LH}^*, \delta_{HH}^*) = \left(\frac{c-(1-q_{DM})x}{q_{DM}x}, 1\right)$,

where $p^- = \beta(\theta = \theta_H \mid s_{DM} = L)$ *and* $p^+ = \beta(\theta = \theta_H \mid s_{DM} = H)$.

The following table restates the results of Proposition 3.3, and compares them with those in the baseline model:

<table>
<tr><td></td><td>$p^+ < \frac{x}{1+x}$</td><td>$p^+ \geq \frac{x}{1+x} > p$</td><td>$p \geq \frac{x}{1+x} > p^-$</td><td>$p^- \geq \frac{x}{1+x}$</td></tr>
<tr><td>Baseline Model</td><td colspan="2">$\sigma_L^* = \frac{p}{(1-p)x}$ and $(\delta_L^*, \delta_H^*) = \left(0, \frac{c}{x}\right)$</td><td colspan="2">$\sigma_L^* = 1$ and $(\delta_L^*, \delta_H^*) = (0, 1)$</td></tr>
<tr><td>Private Signal
$q_{DM} < 1 - \frac{c}{x}$</td><td>$\sigma_L^* = \frac{p^+}{(1-p^+)x}$
$(\delta_{LH}^*, \delta_{HH}^*) = \left(0, \frac{c}{(1-q_{DM})x}\right)$</td><td colspan="2">$\sigma_L^* = 1$
$(\delta_{LH}^*, \delta_{HH}^*) = (0, 1)$</td><td>$\sigma_L^* = 1$
$(\delta_{LH}^*, \delta_{HH}^*) = (1, 1)$</td></tr>
<tr><td>Private Signal
$q_{DM} \geq 1 - \frac{c}{x}$</td><td colspan="3">$\sigma_L^* = \frac{p^-}{(1-p^-)x}$
$(\delta_{LH}^*, \delta_{HH}^*) = \left(\frac{c-(1-q_{DM})x}{q_{DM}x}, 1\right)$</td><td>$\sigma_L^* = 1$
$(\delta_{LH}^*, \delta_{HH}^*) = (1, 1)$</td></tr>
</table>

Table 3.1: A comparison of equilibrium strategy profiles specified in Proposition 3.3 with those in the baseline model (see Proposition 3.1).

Proposition 3.3 tells us that the equilibrium is described by four disjoint and collectively exhaustive regions in the (p, q_{DM}, x, c) parameter space, which correspond to cases (i)-(iv), respectively. We refer to these regions as:

[17] As before, in knife-edge cases, we select the sender-preferred strategy profile whenever there are multiple equilibria.

1. PF_{LH} (pure-strategy falsification with the DM accepting with probability 1 when $(s_{DM}, \widetilde{m}) \in \{(L,H),(H,H)\}$)

2. PF_{HH} (pure-strategy falsification with the DM accepting with probability 1 when $(s_{DM}, \widetilde{m}) = (H,H)$);

3. MF_{HH} (mixed-strategy falsification with the DM accepting with a positive probability when $(s_{DM}, \widetilde{m}) = (H,H)$);

4. MF_{LH} (mixed-strategy falsification with the DM accepting with positive probability when $(s_{DM}, \widetilde{m}) = (L,H)$ and with probability 1 when $(s_{DM}, \widetilde{m}) = (H,H)$).

Figure 3.2 illustrates the four regions in the (p, q_{DM}) parameter space while keeping x and c fixed. The baseline model is captured by $q_{DM} = \frac{1}{2}$.

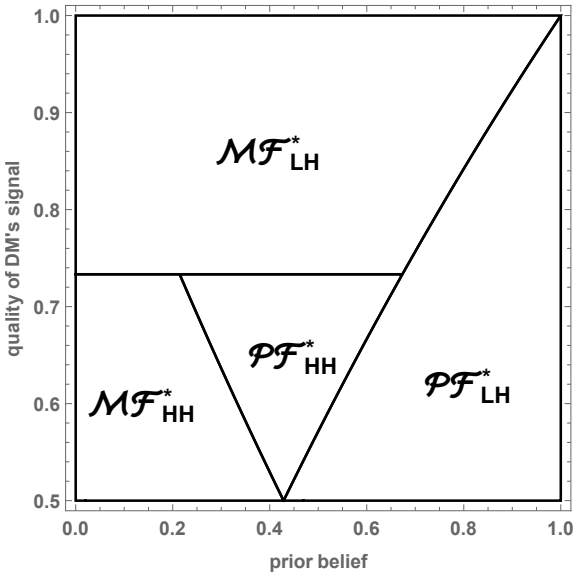


Figure 3.2: An illustration of the regions which describe the equilibrium when the DM obtains a private signal of quality q_{DM}. The simulation uses $c = 0.2$ and $x = 0.75$.

The key difference between this setup and the baseline environment is that, here, the sender needs to make an inference about the DM's belief: the sender knows the

DM's prior belief but he does not know her belief after she has observed her private signal. If the DM observes $s_{DM} = H$, then her belief becomes $\beta(s_{DM} = H) = \frac{pq_{DM}}{pq_{DM}+(1-p)(1-q_{DM})}$, which we denote by p^+. If she observes $s_{DM} = L$, then her belief becomes $\beta(s_{DM} = L) = \frac{p(1-q_{DM})}{p(1-q_{DM})+(1-p)q_{DM}}$, which we denote by p^-. We will say that a DM who has observed $s_{DM} = H$ ($s_{DM} = L$) is "optimistic" ("pessimistic").

When $(p, q_{DM}) \in PF_{LH}$, the prior belief of the DM is so high that both an optimistic and a pessimistic DM's belief is swayed towards acceptance by $\widetilde{m} = H$ even when the sender falsifies with probability 1. Thus, given that $c < x$, the sender has an incentive to falsify with probability 1.

When $(p, q_{DM}) \in PF_{HH}$, the prior belief is at a moderate level, which means that only an optimistic DM's belief is swayed towards acceptance by $\widetilde{m} = H$ when the sender falsifies with probability 1. The quality of the DM's signal is low, which means that the sender realises upon observing $s = L$ that it is quite likely that the DM is optimistic; in other words, when q_{DM} is low, $\Pr(s_{DM} = H \mid s = L) = 1 - q_{DM}$ is high. The sender has an incentive to falsify with probability 1 as the expected benefit of falsification, $(1 - q_{DM})x$, exceeds the cost, c.

When $(p, q_{DM}) \in MF_{HH}$, the prior belief is so low that even an optimistic DM's belief is not swayed towards acceptance when the sender falsifies with probability 1. Again, the quality of the DM's signal is low, so the sender realises upon observing $s = L$ that it is quite likely that the DM is optimistic. In equilibrium, the sender falsifies with a probability that makes an optimistic DM just indifferent between accepting and rejecting upon observing $\widetilde{m} = H$, while a pessimistic DM rejects upon observing $\widetilde{m} = H$.

Finally, when $(p, q_{DM}) \in MF_{LH}$, the quality of the DM's signal is high, so the sender realises upon observing $s = L$ that he likely faces a pessimistic DM. In equilibrium, the sender falsifies with a probability that makes a pessimistic DM just indifferent between accepting and rejecting upon observing $\widetilde{m} = H$, while an optimistic DM accepts upon

observing $\widetilde{m} = H$.

3.3.2 How Is Falsification Affected?

Having described the equilibrium in the setup where the DM acquires a private signal, we can now analyse how the sender's falsification is affected. It turns out that the effect on the sender's falsification is ambiguous: it may increase, decrease, or remain unchanged. The following proposition describes how this effect depends on the parameter values:

Proposition 3.4. *The impact of the DM's acquisition of a private signal of quality q_{DM} on the sender's falsification in equilibrium, σ_L^*, is as follows:*
(i) if $p^- \geq \frac{x}{1+x}$ and if $p^- < \frac{x}{1+x} \leq p$ and $q_{DM} < 1 - \frac{c}{x}$, the sender's falsification remains unchanged;
(ii) if $p < \frac{x}{1+x}$ and $q_{DM} < 1 - \frac{c}{x}$, the sender's falsification increases;
(iii) if $p^- < \frac{x}{1+x}$ and $q_{DM} \geq 1 - \frac{c}{x}$, the sender's falsification decreases.

Figure 3.3 illustrates the effect of the DM's observation of a private signal on the sender's falsification strategy, σ_L^*, in the (p, q_{DM}) parameter space for fixed values of c and x (which are set to be the same as in Figure 3.2).

The DM's acquisition of a private signal has no effect on the sender's falsification when the prior belief is high enough (region labelled as "no effect"). Intuitively, when the prior is high enough, then—regardless of whether the DM observes a private signal or not—she will likely accept even if the sender falsifies with probability 1. Thus, the sender has an incentive to falsify with the same probability (equal to 1) in both cases.

When the quality of the DM's signal is low and the prior belief is low (region labelled as "more falsification"), the DM's acquisition of a private signal results in more falsification by the sender. Since the DM's signal is of low quality, the sender

believes upon observing $s = L$ that he likely faces an optimistic DM, which makes him falsify with a higher probability than if he faced a neutral DM (i.e. a DM who has not observe a private signal and thus has belief p).

Lastly, when the quality of the DM's signal is high and the prior belief is low (region labelled as "less falsification"), the DM's acquisition of a private signal results in less falsification. Here, the sender realises upon observing $s = L$ that he likely faces a pessimistic DM, which makes him falsify with a lower probability than if he faced a neutral DM.

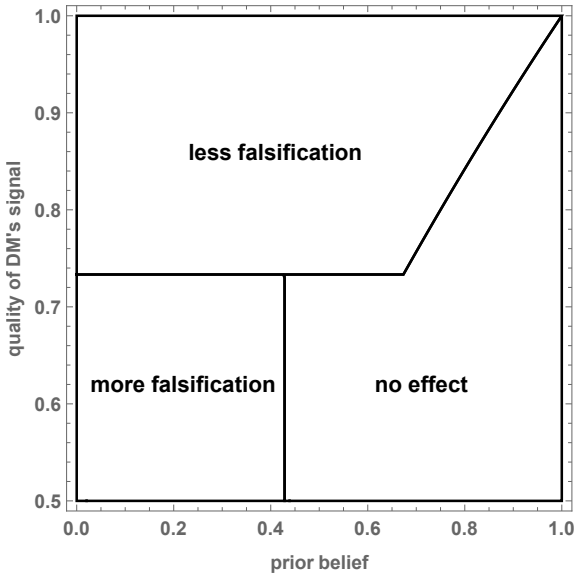


Figure 3.3: An illustration of how the DM's acquisition of a private signal affects the sender's falsification strategy, σ_L^* (compared to the baseline model with no DM's private signal). The simulation uses $c = 0.2$ and $x = 0.75$.

3.3.3 How Is the Decision Maker's Welfare Affected?

In this subsection, we analyse how obtaining a private signal affects the DM's welfare, i.e. her ex-ante expected payoff in equilibrium.

The private signal affects the DM's welfare in two ways. First, observing a private signal obviously means that the DM has additional information, which boosts her ex-

pected payoff. Second, it has an impact on the falsification by the sender and thus on how likely the information received from the sender is manipulated. In the previous section, we have shown that this second effect can work in either direction: the sender's incentives to falsify can be weakened or strengthened. If they are weakened, then a private signal brings an additional benefit—beyond the informational content of the signal itself—in that the sender's message becomes less manipulated. If they are strengthened, the DM's benefit from the informational content of the private signal is at least partly offset by the loss associated with more falsification by the sender.

We denote the value of the private signal by

$$V = \mathbb{E}\left[v\left(\sigma^*\left(q_{DM}\right), \delta^*\left(q_{DM}\right)\right)\right] - \mathbb{E}\left[v\left(\sigma^*\left(\frac{1}{2}\right), \delta^*\left(\frac{1}{2}\right)\right)\right], \tag{3.5}$$

where $\sigma^*\left(q_{DM}\right)$ and $\delta^*\left(q_{DM}\right)$ are the equilibrium strategies under a given value of q_{DM}.

Proposition 3.5. *The DM's acquisition of a private signal of quality q_{DM} strictly increases her welfare if (i) $q_{DM} < 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x} \leq p^+$, or (ii) $q_{DM} \geq 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x}$. Otherwise, her welfare remains unchanged relative to the baseline model.*

The results of Proposition 3.4 and Proposition 3.5 produce five regions in the (p, q) parameter space, which we illustrate in Figure 3.4. We describe these five regions below.

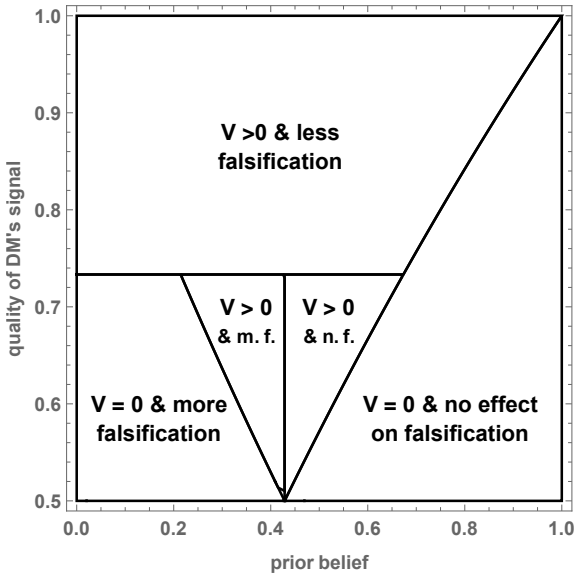


Figure 3.4: An illustration of how the DM's acquisition of a private signal affects the DM's welfare (compared to the baseline model with no DM's private signal). The simulation uses $c = 0.2$ and $x = 0.75$; "m.f." and "n.f." are abbreviations for, respectively, "more falsification" and "no effect on falsification".

Region $V = 0$ and no effect on falsification. This region is identical to region PF^*_{LH} and is defined by $p^- \geq \frac{x}{1+x}$. Here, the private signal does not change the DM's welfare. This is because the sender's equilibrium strategy does not change relative to the baseline model and neither does the DM's decision rule: her equilibrium strategy in the model with a private signal is such that she accepts if and only if $\widetilde{m} = H$. Thus, despite bringing some additional information about the state of the world, the DM's welfare remains the same as in the baseline model.

Region $V > 0$ and no effect on falsification. This region is a part of region PF^*_{HH} and is defined by $q_{DM} < 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x} \leq p$. The private signal strictly increases the DM's welfare. The DM's equilibrium strategy is such that she accepts when $(s_{DM}, \widetilde{m}) = (H, H)$, and rejects otherwise. Thus, her decision-making relies on realisations of both s_{DM} and $\widetilde{m}$, which means that the information from the private

signal is essential to determining her optimal action. At the same time, the sender's falsification strategy remains the same as in the baseline model, which means that the informational benefit from the private signal is not dampened by increased falsification.

Region $V > 0$ and more falsification. This region is a part of region PF^*_{HH} and is defined by $q_{DM} < 1 - \frac{c}{x}$ and $p < \frac{x}{1+x} \leq p^+$. The private signal strictly increases the DM's welfare. The DM accepts only when $(s_{DM}, \widetilde{m}) = (H, H)$, and rejects otherwise. Therefore, again, the information from the private signal is essential to determining her optimal action. There is more falsification than in the baseline model but it does not fully offset the benefits brought by the information contained in the private signal.

Region $V = 0$ and more falsification. This region is identical to region MF^*_{HH} and is defined by $q_{DM} < 1 - \frac{c}{x}$ and $p^+ < \frac{x}{1+x}$. Here, the private signal does not change the DM's welfare. She is indifferent between accepting and rejecting when $(s_{DM}, \widetilde{m}) = (H, H)$, and rejects otherwise. Hence, she would have the same expected payoff if she always rejected. In the baseline model, her equilibrium strategy is also such that she would have the same expected payoff if she always rejected. Therefore, her expected payoff is the same in both cases. The increased falsification thus fully offsets any benefits brought by the additional information in the private signal.

Region $V > 0$ and less falsification. This region is identical to region MF^*_{LH} and is defined by $q_{DM} \geq 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x}$. The private signal strictly increases the DM's welfare. The DM is indifferent between accepting and rejecting when $(s_{DM}, \widetilde{m}) = (L, H)$, accepts when $(s_{DM}, \widetilde{m}) = (H, H)$, and rejects otherwise. Hence, the probability of her accepting upon observing $\widetilde{m} = H$ differs depending on whether $s_{DM} = L$ or $s_{DM} = H$. The benefit brought by the additional information in the private signal is further boosted by the reduced falsification by the sender.

Despite the potentially increased falsification by the sender, the DM is always weakly better off with a private signal in this setting: the fact that the DM obtains a private signal never boosts the sender's incentives to falsify by so much as to make the DM worse off with a signal of quality q_{DM} than without it. In fact, a stronger statement follows from Propositions 3.3 and 3.4:

Corollary 3.1. *The welfare of the decision maker is increasing in the quality of her private signal,* q_{DM}.

As we will show later, the result that the DM is always weakly better off with a private signal than without it still holds when the quality of the sender's private signal is $q < 1$; however, the statement in Corollary 3.1 may then not be satisfied.

3.3.4 Acquiring Private vs. Public Independent Evidence

In this section, we consider how the DM's welfare is affected when the realisation of the additional signal is publicly observed by both players, as opposed to being privately observed only by the DM. The aim here is to shed light on policy makers' benefits from conducting public rather than private independent research.

We will use s_{pub} to denote the realisation of the public signal. Since the sender decides whether to falsify evidence after the public signal is realised, the analysis here amounts to repeating the steps in the baseline model for two cases: (i) the public signal is $s_{pub} = H$ and therefore the new prior is $p^+ = \Pr(\theta = \theta_H \mid s_{pub} = H)$, and (ii) the public signal is $s_{pub} = L$ and therefore the new prior is $p^- = \Pr(\theta = \theta_H \mid s_{pub} = L)$. We will say that the DM is "optimistic" ("pessimistic") if $s_{pub} = H$ ($s_{pub} = L$); however, since the signal is public, the sender now knows whether the DM is optimistic or pessimistic.

The following table shows—using the same format as Table 3.1—the equilibrium strategies when the DM obtains a public signal of quality q_{DM}. For the DM's strategy, we will now use $\delta_{s_{pub}\widetilde{m}}$ to denote the probability with which the DM accepts when she observes $(s_{DM}, \widetilde{m})$. The DM's strategy is thus given by $(\delta_{LL}, \delta_{LH}, \delta_{HL}, \delta_{HH})$. For the sender's strategy, we will now use $\sigma_{s_{pub}\widetilde{m}}$ to denote the probability with which the sender falsifies when he observes $(s_{DM}, \widetilde{m})$. The sender's strategy is thus given by $(\sigma_{LL}, \sigma_{LH}, \sigma_{HL}, \sigma_{HH})$. Naturally, in equilibrium, it must be that $(\sigma^*_{LH}, \sigma^*_{HH}) = (0, 0)$ and $(\delta^*_{LL}, \delta^*_{HL}) = (0, 0)$.

	$p^+ < \frac{x}{1+x}$	$p^+ \geq \frac{x}{1+x} > p^-$	$p^- \geq \frac{x}{1+x}$
Public Signal	$(\sigma^*_{LL}, \sigma^*_{HL}) = \left(\frac{p^-}{(1-p^-)x}, \frac{p^+}{(1-p^+)x}\right)$	$(\sigma^*_{LL}, \sigma^*_{HL}) = \left(\frac{p^-}{(1-p^-)x}, 1\right)$	$(\sigma^*_{LL}, \sigma^*_{HL}) = (1, 1)$
$q_{DM} > \frac{1}{2}$	$(\delta^*_{LH}, \delta^*_{HH}) = \left(\frac{c}{x}, \frac{c}{x}\right)$	$(\delta^*_{LH}, \delta^*_{HH}) = \left(\frac{c}{x}, 1\right)$	$(\delta^*_{LH}, \delta^*_{HH}) = (1, 1)$

Table 3.2: A summary of the equilibrium strategy profiles when the DM obtains a public signal of quality q_{DM}.

The value of a public signal is constructed analogously to (3.5) in the analysis of the private signal. We denote the value of the public signal by V'. We obtain the following proposition:

Proposition 3.6. *The DM's acquisition of a public signal of quality q_{DM} strictly increases her welfare (i.e. $V' > 0$) if $p^+ > \frac{x}{1+x} > p^-$; otherwise, her welfare remains unchanged relative to the baseline model (i.e. $V' = 0$).*

The DM's welfare with a public signal of quality q_{DM} is strictly lower than that with a private signal of quality q_{DM} (i.e. $V > V'$) if $q_{DM} \geq 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x}$; otherwise, the DM's welfare is the same regardless of whether the additional signal is public or private (i.e. $V = V'$).

Proposition 3.6 implies that, by obtaining a public signal, the DM can increase her welfare but the increase is never higher than it would have been with a private signal.

Moreover, for a sufficiently high quality of the signal and a sufficiently low prior, the benefit from obtaining a signal is strictly higher if it is private rather than public.

The results of Proposition 3.6—together with the results of Proposition 3.5—produce regions in the (p, q) parameter space, which we illustrate in Figure 3.5.

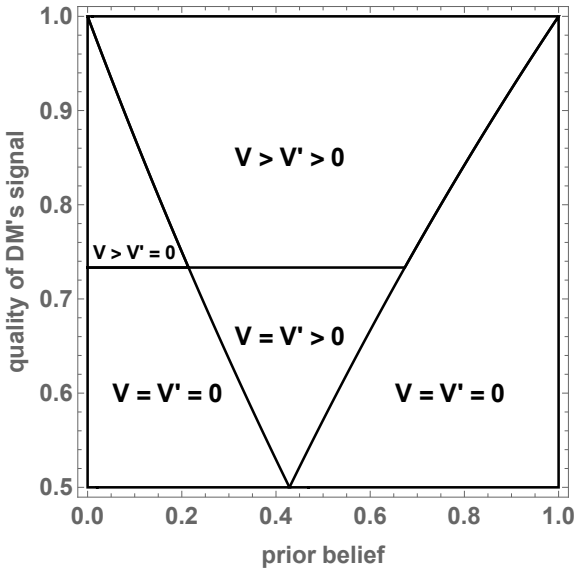


Figure 3.5: An illustration of how the DM's acquisition of a public signal affects the DM's welfare, compared to the acquisition of a private signal and to the baseline model with no DM's private signal. The simulation uses $c = 0.2$ and $x = 0.75$.

It is worth noting that when the signal is public, there is weakly less falsification of evidence than when it is private and $q_{DM} < 1 - \frac{c}{x}$. More precisely, whenever $p^- < \frac{x}{1+x}$, there is strictly less falsification in the former case than in the latter, and whenever $p^- > \frac{x}{1+x}$, the amount of falsification is the same whether the signal is public or private.

As we have seen in Table 3.1, when $p^+ < \frac{x}{1+x}$, $q_{DM} < 1 - \frac{c}{x}$ and the signal is private, the sender falsifies with a probability that makes an optimistic DM indifferent between accepting and rejecting upon observing $\widetilde{m} = H$. On the other hand, when the signal is public, then we can see from Table 3.2 that the probability of falsification depends on the realisation of the public signal: if $s_{pub} = L$, the sender falsifies with a probability that makes a pessimistic DM indifferent (which is lower than the probability needed to

make an optimistic DM indifferent), while if $s_{pub} = H$, he falsifies with a probability that makes an optimistic DM indifferent. For $p^+ \geq \frac{x}{1+x} > p^-$ and $q_{DM} < 1 - \frac{c}{x}$, the sender falsifies with probability 1 when the DM's signal is private; however, when it is public, then again if $s_{pub} = L$, he falsifies with a probability that makes a pessimistic DM indifferent (which is lower than 1), and if $s_{pub} = H$, he falsifies with probability 1.

Despite there being weakly less falsification when the signal is public than when it is private and $q_{DM} < 1 - \frac{c}{x}$, the DM's welfare is the same in both cases. The reason for this is that the reduced falsification is actually on a margin that is irrelevant to the DM's ex ante expected payoff. To see why, consider for example a prior belief such that $p^+ < \frac{x}{1+x}$. When the signal is private and $q_{DM} < 1 - \frac{c}{x}$, the DM's equilibrium strategy is such that a pessimistic DM always rejects and an optimistic DM is indifferent between accepting and rejecting upon observing $\widetilde{m} = H$ and rejects otherwise. When the signal is public, then regardless of whether $s_{pub} = L$ or $s_{pub} = H$, the DM's equilibrium strategy is such that she is indifferent upon receiving $\widetilde{m} = H$ and rejects otherwise. Therefore, when $q_{DM} < 1 - \frac{c}{x}$, no matter whether the signal is public or private, the DM's ex ante expected payoff is the same as if she rejected regardless of the observed information, i.e. as if she effectively relied on her prior only. On the other hand, when $q_{DM} \geq 1 - \frac{c}{x}$, her ex ante expected payoff is strictly higher with a private signal. The DM's equilibrium strategy is then to accept when $(s_{DM}, \widetilde{m}) = (H, H)$, and she is indifferent between the two actions when $(s_{DM}, \widetilde{m}) = (L, H)$, which must yield a strictly higher expected payoff than if she relied on her prior only.

3.3.5 Does Higher Quality of Evidence Always Benefit the Decision Maker?

We have assumed so far that the sender observes a signal which is perfectly informative about the state of the world. In this subsection, we relax this assumption in order

to show additional results: (i) the DM can completely deter falsification by obtaining a private signal, and (ii) a higher quality of the DM's signal may lower her expected payoff.

First, let us consider an adapted baseline model which is the same as the original baseline model in Section 3.2 except that now the sender observes a *noisy* signal of the state of the world, s, where the quality of the signal is $\Pr(s = i \mid \theta = \theta_i) = q$ for $i \in \{L, H\}$ with $q \in \left(\frac{1}{2}, 1\right]$. The following proposition describes the equilibrium:

Proposition 3.7. *Suppose that the sender's private signal is of quality* $q \in \left(\frac{1}{2}, 1\right]$. *The game has a unique equilibrium in which the sender's and the DM's equilibrium strategies are as follows:*

(i) if $p \in (0, 1-q)$, *then* $(\sigma_L^*, \sigma_H^*) = (0, 0)$ and $(\delta_L^*, \delta_H^*) = (0, 0)$;
(ii) if $p \in \left[1-q, \frac{1-(1-x)q}{1+x}\right)$, *then* $(\sigma_L^*, \sigma_H^*) = \left(\frac{p+q-1}{(q-p)x}, 0\right)$ *and* $(\delta_L^*, \delta_H^*) = \left(0, \frac{c}{x}\right)$;
(iii) if $p \in \left[\frac{1-(1-x)q}{1+x}, q\right)$, *then* $(\sigma_L^*, \sigma_H^*) = (1, 0)$ *and* $(\delta_L^*, \delta_H^*) = (0, 1)$;
(iv) if $p \in [q, 1)$, *then* $(\sigma_L^*, \sigma_H^*) = (0, 0)$ *and* $(\delta_L^*, \delta_H^*) = (1, 1)$.

The imperfect quality of the sender's signal means that it is now possible in equilibrium that the sender falsifies with probability zero even upon observing $s = L$. When the sender's signal is noisy and the prior belief is very low or very high, the DM's action does not depend on the received message and thus the sender has no incentive to falsify.

Another new feature is that the higher the quality of the signal observed by the sender, the more likely (at least weakly) he is to falsify upon observing $s = L$. Intuitively, a high quality means that if the message received by the DM happens to be not falsified, it reveals the true state with high precision. Consequently, the DM becomes more willing to make her action dependent on the received message, which boosts the sender's incentive to falsify.

Acquisition of Private Independent Evidence

Suppose now that the DM acquires a private signal of quality $q_{DM} \leq q$. Thus, the sender's signal is not perfectly informative but is more informative than the DM's private signal. In this setting, two new results arise relative to the setting where the sender is perfectly informed about θ.

The first new result is that, by obtaining a private signal, the DM can completely disincentivise falsification of evidence, i.e. she can make the sender falsify evidence with probability zero in equilibrium. We are naturally interested here in values of prior belief p such that $1 - q < p < q$, which guarantees that the sender falsifies with a positive probability in the absence of DM's private signal. The following proposition describes the circumstances under which obtaining a private signal by the DM completely disincentivises falsification of evidence by the sender:

Proposition 3.8. *Suppose that the sender's private signal is of quality* $q \in \left(\frac{1}{2}, 1\right]$ *and that the DM observes a private signal of quality* $q_{DM} \in \left(\frac{1}{2}, q\right)$. *Consider values of prior belief* p *such that* $1-q < p < q$. *Falsification of evidence can be then completely deterred in equilibrium (i.e.* $\sigma_L^* = 0$*) only if* $q < 1$. *Complete deterrence occurs whenever the prior belief is such that:*

(i) $\beta\left(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 0\right) > \frac{1}{2}$, $\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0\right) < \frac{1}{2}$, and $\pi_{H|L} x \leq c$, or

(ii) $\beta\left(s_{DM} = H, \widetilde{m} = L\right) > \frac{1}{2}$, $\beta\left(s_{DM} = L, \widetilde{m} = L\right) < \frac{1}{2}$, *and* $\pi_{L|L} x \leq c$,

where $\pi_{i|j} = \Pr\left(s_{DM} = i \mid s = j\right)$.

The second new result is that now the DM's welfare is not necessarily increasing in the quality of her private signal:

Proposition 3.9. *Suppose that the sender's private signal is of quality* $q \in \left(\frac{1}{2}, 1\right]$ *and that the DM observes a private signal of quality* $q_{DM} \in \left(\frac{1}{2}, q\right)$. *Then the DM's welfare*

is increasing in q_{DM} *if* $p \leq \frac{1}{2}$ *but can be non-monotonic with respect to* q_{DM} *if* $p > \frac{1}{2}$.

We now discuss the results in Propositions 3.8 and 3.9 in more detail. Generally speaking, for the complete deterrence to occur, two elements are necessary. First, it must be that the sender's message is pivotal to the DM's decision only for one realisation of the DM's signal. Second, the sender—upon observing his own signal—must conclude that it is unlikely that the DM has observed that particular signal realisation and hence the sender's message is unlikely to be pivotal. The sender's expected benefit from falsification is then lower than its cost, and so falsification is deterred. Proposition 3.8 demonstrates that there are two instances where this happens.

First, consider part (i) of Proposition 3.8, which requires a prior belief such that

$$\beta\left(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 0\right) > \frac{1}{2} \tag{3.6}$$

and

$$\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0\right) < \frac{1}{2}, \tag{3.7}$$

where the latter together with $q > q_{DM}$ implies that $p < \frac{1}{2}$.

Conditions (3.6) and (3.7) define a range of values of prior belief such that the sender's message can be pivotal to the DM's decision only if the DM's private signal is $s_{DM} = H$. To see this, note that (i) the condition in (3.7) and $q > q_{DM}$ imply that $\beta\left(s_{DM} = H, \widetilde{m} = L\right) < \frac{1}{2}$, and (ii) the condition (3.7) also implies that $\beta\left(s_{DM} = L, \widetilde{m} = L\right) < \frac{1}{2}$. However, the sender chooses to falsify with a positive probability if and only if the expected benefit from falsification exceeds the cost, i.e. $\pi_{H|L}x \geq c$ must be satisfied. If (3.6) and (3.7) are satisfied but $\pi_{H|L}x < c$, the sender realises that her message is not likely enough to be pivotal to the DM's decision, and therefore he chooses not to falsify evidence. Falsification is thus deterred by the fact that the DM

obtains a private signal.[18]

An improvement in the quality of the DM's private signal has two distinct effects on the sender's incentives to falsify evidence. The first effect is that, for any given message from the sender and any given strategy of his, the DM's belief in response to $s_{DM} = L$ decreases and her belief in response to $s_{DM} = H$ increases. In other words, (3.6) and (3.7) are more likely to be satisfied. The second effect of improving the quality of the DM's signal is that the higher q_{DM} is, the less likely it is that the DM's private signal is $s_{DM} = H$ when the sender's signal is $s = L$, i.e. $\pi_{H|L}$ decreases. When q_{DM} becomes high enough, the condition $\pi_{H|L}x \geq c$ is no longer satisfied and the sender has no incentive to falsify.

Thus, both effects of an increase in q_{DM} work to disincentivise falsification. In other words, when the prior belief is less than $\frac{1}{2}$, there may be a twofold benefit to the DM from increasing the quality of her private signal. Not only does it make her private information more precise, but it also unambiguously boosts the chances that the sender will not find it worthwhile to falsify evidence. This result explains the statement in Proposition 3.9 that the DM's welfare is always increasing in q_{DM} when $p < \frac{1}{2}$.

Let us now consider part (ii) of Proposition 3.8, which requires a prior belief such that

$$\beta\left(s_{DM} = H, \widetilde{m} = L\right) > \frac{1}{2} \tag{3.8}$$

and

$$\beta\left(s_{DM} = L, \widetilde{m} = L\right) < \frac{1}{2}, \tag{3.9}$$

where the former together with $q > q_{DM}$ implies that $p > \frac{1}{2}$.

Conditions (3.8) and (3.9) define a range of values of prior belief such that the

[18]When the sender's private signal perfectly reveals the state of the world, i.e. $q = 1$, we have $\beta\left(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 0\right) = 1$ and $\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0\right) = 1$, which implies that only one of (3.6) and (3.7) is satisfied.

sender's message can be pivotal to the DM's decision only if the DM's private signal is $s_{DM} = L$. To see this, note that (i) the condition in (3.8) implies $\beta\left(s_{DM} = H, \widetilde{m} = H\right) > \frac{1}{2}$, and (ii) the condition in (3.8) and $q > q_{DM}$ imply that $\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0\right) > \frac{1}{2}$. The sender chooses to falsify with a positive probability if and only if the expected benefit from falsification exceeds the cost, which means here that $\pi_{L|L}x \geq c$ must be satisfied. If (3.8) and (3.9) are satisfied but $\pi_{L|L}x < c$, the sender realises that her message is not likely enough to be pivotal to the DM's decision, and therefore falsification is deterred when the DM obtains a private signal.[19]

As in part (i) of Proposition 3.8, a higher quality of the DM's private signal has two effects on the sender's incentives to falsify evidence. First, $\beta\left(s_{DM} = H, \widetilde{m} = L\right)$ increases while $\beta\left(s_{DM} = L, \widetilde{m} = L\right)$ decreases, which means that (3.8) and (3.9) are more likely to be satisfied. In other words, we observe an increase in the range of values of the prior belief such that the sender's message is pivotal to the DM's action only if $s_{DM} = L$. However, the second effect works in the opposite direction to what we observed in part (i) of Proposition 3.8. The higher q_{DM} is, the more likely it is that the DM's private signal is $s_{DM} = L$ when the sender's signal is $s = L$, i.e. $\pi_{L|L}$ increases. Thus, an increase in q_{DM} can result in $\pi_{L|L}x \geq c$ being satisfied. Hence, improving the quality of the DM's signal may dramatically (and, in this binary setting, discontinuously) boost the sender's incentives to falsify.

Overall, an increase in q_{DM} has two effects on deterrence of falsification, which—unlike in part (i) of Proposition 3.8—work in opposite directions. The first one widens, while the second one narrows the range of parameter values for which the DM's private information deters falsification. The two opposing effects can lead to a non-monotonic relationship between the quality of the DM's signal and the DM's expected payoff in

[19]When the sender's private signal perfectly reveals the state of the world, i.e. $q = 1$, we have $\beta\left(s_{DM} = H, \widetilde{m} = L\right) = 0$ and $\beta\left(s_{DM} = L, \widetilde{m} = L\right) = 0$, which implies that (3.8) and (3.9) are not simultaneously satisfied.

equilibrium, as stated in Proposition 3.9 for $p > \frac{1}{2}$. In Figure 3.6, we illustrate this possibility with a numerical example:

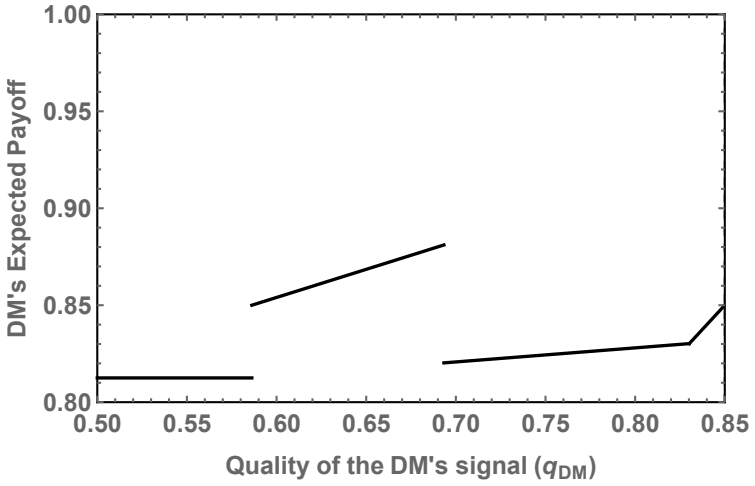


Figure 3.6: The DM's expected payoff in equilibrium as a function of the quality of the DM's signal. The numerical example assumes $x = \frac{3}{4}$, $c = \frac{2}{5}$, $p = \frac{4}{5}$, and $q = \frac{17}{20}$ (depicted with the dashed line).

3.4 Paying Less Attention to the Sender

In the previous section, we have analysed how acquisition of private independent information can improve the DM's decision making. In this section, we study whether she can improve the decision making by committing to pay less attention to the sender. We then briefly discuss the interactions between this strategy and acquisition of private independent evidence.

3.4.1 Commitment to Ignorance

We assume that paying more attention is costless and allow the DM to choose ex ante a level of attention that determines the probability of her receiving the sender's message. More formally, suppose that the DM can commit to pay attention to the sender's message with probability τ and to ignore it with probability $1 - \tau$. If she does so, she

observes the sender's message $\widetilde{m}$ with probability τ, and does not observe it otherwise (and thus she keeps her prior belief about the state of the world). The baseline model corresponds to the DM choosing $\tau = 1$. We refer to the value of τ chosen by the DM as her level of attention, and to $1 - \tau$ as her level of ignorance.[20]

Committing to $\tau < 1$ has two opposing effects on the DM's expected payoff. On the one hand, it means that with probability $1 - \tau$ the DM has to rely only on her prior when making the decision. However, the benefit is that it makes the sender's message less likely to be observed by the DM and thus less likely to influence her decision, which may weaken the sender's incentives to falsify.

In order to disincentivise falsification, the DM needs to commit to ignore the sender's message with a sufficiently high probability. To be more precise, τ must be below $\frac{c}{x}$. To see why, note that, from the sender's perspective, when the DM commits to ignore his message with probability $1 - \tau$, the expected benefit from falsifying when he observes that $\theta = \theta_L$ is at most τx (which happens when the DM's strategy is $\delta_H^* = 1$ and $\delta_L^* = 0$), whereas the cost is c. If $\tau < \frac{c}{x}$, the cost of falsification exceeds the expected benefit, and therefore it is optimal for the sender not to falsify.

Therefore, from the DM's point of view, if committing to $\tau < 1$ is optimal, the DM should choose $\tau = \frac{c}{x} - \varepsilon$, where ε is arbitrarily close to zero. We will henceforth assume that when $\tau = \frac{c}{x}$, the sender does not falsify. If the DM chose $\tau < \frac{c}{x}$, she would disincentivise falsification, but she would be receiving no information other than her prior more often than necessary; if she chose $\frac{c}{x} < \tau < 1$, she would not disincentivise falsification and would be sometimes making her decision based on the prior rather than on a possibly falsified—but still to some degree informative—message from the sender.

These observations yield the following lemma:

[20] In the appendix, we consider a model with a continuous—rather than a binary—choice of falsification by the sender, which yields additional intuition.

Lemma 3.1. *The DM's optimal strategy is either (a) to never ignore the sender's message, i.e. to set* $\tau = 1$*, or (b) to commit to ignore the sender's message just often enough to discourage the sender from falsifying evidence, i.e. to set* $\tau = \frac{c}{x}$*.*

3.4.2 Optimal Level of Attention and Ignorance

Let $\mathbb{E}\left(v_\tau\left(\sigma^*, \delta^*\right)\right)$ denote the DM's expected payoff when she pays attention with probability τ and the equilibrium strategies are (σ^*, δ^*). If the DM chooses not to ignore any information, i.e. she sets $\tau = 1$, her expected payoff is the same as in the equilibrium of the baseline model:

$$\mathbb{E}\left(v_{\tau=1}\left(\sigma^*, \delta^*\right)\right) = \begin{cases} 1-p & \text{for } p \in \left(0, \frac{x}{1+x}\right) \\ 1-(1-p)\,x & \text{for } p \in \left[\frac{x}{1+x}, 1\right) \end{cases}. \tag{3.10}$$

On the other hand, if she commits to pay attention to the sender's message with probability $\tau = \frac{c}{x}$, it is as if she received a perfectly informative signal with probability $\frac{c}{x}$ and a completely uninformative signal with probability $1 - \frac{c}{x}$. Thus, with probability $\frac{c}{x}$, she makes the correct decision, and with probability $1 - \frac{c}{x}$, she makes the decision that is dictated by the prior. Her expected payoff is then

$$\mathbb{E}\left(v_{\tau=\frac{c}{x}}\left(\sigma^*, \delta^*\right)\right) = \begin{cases} 1 \times \frac{c}{x} + (1-p) \times \left(1 - \frac{c}{x}\right) & \text{for } p \in \left(0, \frac{1}{2}\right) \\ 1 \times \frac{c}{x} + p \times \left(1 - \frac{c}{x}\right) & \text{for } p \in \left[\frac{1}{2}, 1\right) \end{cases}. \tag{3.11}$$

The following table summarises the DM's expected payoff in the two scenarios:

	$0 < p < \frac{x}{1+x}$	$\frac{x}{1+x} < p < \frac{1}{2}$	$\frac{1}{2} < p$
$\mathbb{E}\left(v_{\tau=1}\left(\sigma^*, \delta^*\right)\right)$	$1-p$	$p + (1-p)(1-x)$	
$\mathbb{E}\left(v_{\tau=c/x}\left(\sigma^*, \delta^*\right)\right)$	$\frac{c}{x} + (1-p)\left(1-\frac{c}{x}\right)$		$\frac{c}{x} + p\left(1-\frac{c}{x}\right)$

Table 3.3: The DM's expected payoff when she does not ignore the sender's message (i.e. $\tau = 1$) and when she commits to ignore it with probability $1 - \frac{c}{x}$ (i.e. $\tau = \frac{c}{x}$).

By comparing the DM's expected payoffs when $\tau = 1$ and when $\tau = \frac{c}{x}$, we can pin down the circumstances under which it is optimal for the DM to set $\tau = 1$ or $\tau = \frac{c}{x}$. The DM's optimal choice of τ is denoted by τ^*.

Proposition 3.10. *It is optimal for the DM to commit to ignore the sender's message with probability $\tau^* = \frac{c}{x}$ if and only if*

$$\min\left\{x(1-x), x\left(1 - \frac{1-p}{p}x\right)\right\} \leq c; \tag{3.12}$$

otherwise it is optimal for the DM not to ignore the sender's message, i.e. to set $\tau^* = 1$.

Naturally, if it were that $c > x$, the sender would never falsify evidence, so the DM would have no incentive to commit to ignore his message and hence would set $\tau^* = 1$. Together with Proposition 3.10, this implies that committing to ignore the sender's message is optimal for the DM when $\min\left\{x(1-x), x\left(1 - \frac{1-p}{p}x\right)\right\} \leq c \leq x$, i.e. when the cost of falsification c takes moderate values.

Table 3.4 summarises the DM's optimal choices of τ as defined by Proposition 3.10. The analysis is performed for three different cases in which we compare the values of x and $1 - \frac{c}{x}$. As we will see in the next section—where we provide more intuition—x can be seen as capturing the DM's "benefit" from commitment to ignorance, while $1 - \frac{c}{x}$ reflects the DM's "cost" of commitment to ignorance. When the undetectability of falsification is such that $x > 1 - \frac{c}{x}$, it is always optimal for the DM to commit to ignore by setting $\tau = \frac{c}{x}$. On the other hand, when $x < 1 - \frac{c}{x}$, the DM finds it optimal to pay

full attention, i.e. to set $\tau = 1$, as long as prior belief p is high enough. Furthermore, this threshold value of p above which the DM optimally chooses to pay full attention, $\tilde{p} = \frac{x}{1+x-\frac{c}{x}}$, decreases as the difference between $1 - \frac{c}{x}$ (the "cost" of ignorance) and x (the "benefit" from ignorance) increases.

	$0 < p < \frac{x}{1+x}$	$\frac{x}{1+x} \leq p < \frac{1}{2}$	$\frac{1}{2} \leq p$
$x < 1 - \frac{c}{x}$	$\tau^* = \frac{c}{x}$	$\tau^* = \frac{c}{x}$ for $p < \tilde{p}$, $\tau^* = 1$ for $p > \tilde{p}$, where $\tilde{p} = \frac{x}{1+x-\frac{c}{x}}$	$\tau^* = 1$
$x = 1 - \frac{c}{x}$	$\tau^* = \frac{c}{x}$	$\tau^* = \frac{c}{x}$	indifferent between $\tau^* = \frac{c}{x}$ and $\tau^* = 1$
$x > 1 - \frac{c}{x}$	$\tau^* = \frac{c}{x}$	$\tau^* = \frac{c}{x}$	$\tau^* = \frac{c}{x}$

Table 3.4: The DM's optimal choices of τ as a function of p, c, and x, when she can commit to ignore the sender's message.

3.4.3 The Role of Parameters

Figure 3.7 illustrates, in the (x, c) parameter space, the regions where the DM chooses not to ignore the sender's message and where she chooses to commit to ignore it with probability $1 - \tau^*$. There are two regions in which it is optimal for the DM not to ignore the sender's message: the top-left triangular region and the bottom semicircular region. The former corresponds to the values of c and x which satisfy $c > x$. The latter corresponds to the values of c and x which satisfy $\min\left\{x\left(1-x\right), x\left(1-\frac{1-p}{p}x\right)\right\} > c$.

We now discuss in more detail how the DM's incentives to commit to ignore the sender's message are affected by parameters c, x, and p.

Role of the cost of falsification, c. It is optimal to commit to ignore the sender's message only if the cost of falsification is moderate. As mentioned earlier, when $c > x$, the cost of falsification is so high that the sender never has an incentive to falsify, and

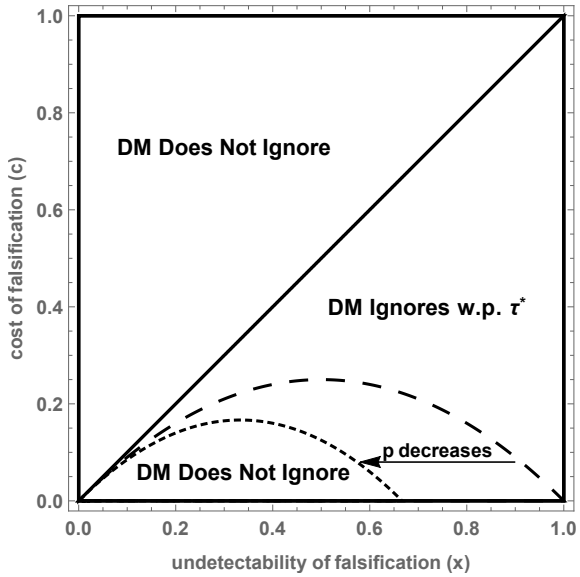


Figure 3.7: An illustration of the DM's optimal commitment choices as a function of parameters c, $\bar{x}$, and p. The dashed line shows $c = \bar{x}\,(1 - \bar{x})$; the dotted line shows $c = \bar{x}\left(1 - \frac{1-p}{p}\bar{x}\right)$ for $p = \frac{2}{5}$.

hence the DM naturally prefers not to ignore his message. The more surprising result is that the DM optimally chooses not to ignore the sender's message also when falsification is sufficiently cheap, i.e. when the condition in Proposition 3.10 is not satisfied. This occurs because a decrease in c lowers the sender's relative cost of falsification, $\frac{c}{x}$, which means that the DM needs to commit to ignore the sender's message with a higher probability if she wants to disincentivise falsification. Eventually, when c is sufficiently low, the DM prefers to observe a possibly falsified message from the sender rather than to commit to almost always ignore his message.

Role of the undetectability of falsification, x. It can be optimal for the DM to pay full attention to the sender's message both under high and low undetectability of falsification relative to its cost.

Naturally, it is optimal to pay full attention to the sender's message when falsification has low undetectability relative to the cost, i.e. when $x < c$. In that case,

falsification is so easily detectable by the DM that the sender has no incentive to falsify and thus the DM prefers not to ignore the sender's message.

However, even when $x \geq c$, Proposition 3.10 shows that it may still be optimal for the DM to pay full attention to the sender's message. This is because an increase in the undetectability of falsification, x, has two conflicting effects on the DM's incentives to commit to ignore the sender's message. First, a higher value of x means that falsification is less detectable. Since the DM's commitment to ignorance can deter falsification by the sender, this channel drives up the DM's relative benefit from commitment to ignorance. Second, as x increases, the relative cost of falsification incurred by the sender decreases, and therefore the level of ignorance required to disincentivise falsification, $1-\frac{c}{x}$, goes up. Thus, this channel makes commitment to ignorance less attractive to the DM. When x is sufficiently high, the channel which makes commitment to ignorance more attractive to the DM dominates and therefore the condition in Proposition 3.10 is satisfied. As a result, when x is sufficiently high, the DM optimally chooses to ignore the sender's message with the required probability for all $x \geq c$.

Role of the prior belief, p. As for the impact of prior belief p on the DM's incentives to commit to ignore the sender's message, note that the condition $c \geq x\left(1-\frac{1-p}{p}x\right)$ in Proposition 3.10 is binding if and only if $p < \frac{1}{2}$. This means that when $p \geq \frac{1}{2}$, the DM prefers to commit to ignore the sender's message only if $c \geq x\,(1-x)$. Thus, when $p \geq \frac{1}{2}$, the DM optimally chooses $\tau = 1$ whenever c is below the dashed line in Figure 3.7.

On the other hand, when $p < \frac{1}{2}$, the condition in Proposition 3.10 becomes tighter as p decreases—the cost of falsification c must then be lower and lower for it to be satisfied. Therefore, a lower value of p decreases the size of the region where the DM chooses not to ignore. Figure 3.7 shows that as p decreases to $\frac{2}{5}$, the value of c must be

below the dotted line for the DM to prefer $\tau = 1$. Intuitively, a low value of p means that the state is ex ante likely to be $\theta = \theta_L$, and hence it is likely that the sender is put in a position where he might want to falsify. Therefore, as p decreases, falsification becomes more likely ex ante, which in turn makes paying full attention to the sender's message less attractive than ignoring the sender with a probability that disincentivises falsification.

3.4.4 How Does the Ability to Commit to Ignore Affect the Incentives to Acquire Independent Evidence?

In Section 3.3, we have identified the circumstances under which the acquisition of private independent evidence leads to more or less falsification by the sender. But if the DM is able to commit to ignore the sender's message, she may disincentivise falsification via this instrument, which then means that the benefit from acquiring independent evidence changes. With falsification already deterred, the benefit from acquiring a private signal derives only from its informational value and not from how it affects the sender's falsification strategy.

Thus, the ability to commit to ignore the sender's message may boost or reduce the DM's incentives to acquire independent evidence. It is useful to consider here Figure 3.3. In this example, we have $c = 0.2$ and $x = 0.75$, which means that the condition $x > 1 - \frac{c}{x}$ from Table 3.4 is satisfied, and therefore it is optimal for the DM to commit to ignore the sender with probability $\frac{c}{x}$ for all values of prior belief p. Whenever the acquisition of a private signal by the DM leads to more falsification by the sender (see part (ii) of Proposition 3.4 and the region labelled "more falsification" in Figure 3.3), the DM's ability to commit to ignore is likely to increase the DM's benefit from acquiring a private signal. On the other hand, when the acquisition of a private signal by the DM weakens the sender's incentives to falsify (see part (iii) of Proposition 3.4

and the region labelled "less falsification" in Figure 3.3), the DM's ability to commit to ignore is likely to decrease the DM's benefit from acquiring a private signal. In other words, the two measures discussed in Sections 3.3 and 3.4 are likely to be complements in the former case, and substitutes in the latter case.[21]

3.5 Conclusion

Falsification of scientific evidence can lead to a significant delay, or even prevent, regulatory action in a way that is beneficial to a specific interest group but harms the wider public. A prominent example is the impact of the tobacco industry on the general acknowledgement that smoking has a detrimental effect on human health; however, similar measures taken by interest groups have also been observed in other areas of scientific research.

In this paper, we have investigated the incentives of interest groups to falsify scientific evidence in a communication game between a sender and a decision maker where falsification is costly and can be detected with some predetermined probability. We have also analysed the measures the policymakers could take in order to prevent that. In particular, we have looked at the benefits from acquiring independent evidence and from controlling how much information to receive from the sender.

The obvious benefit from conducting one's own independent research is that it brings additional information that can be used in the decision making process; however, it also affects the sender's incentives to falsify evidence. As a result, the value of acquiring independent evidence may be smaller or larger than the pure information it contains. We have identified the conditions for both possibilities by first determining the circumstances under which the sender's incentives to falsify are strengthened or weakened.

[21] A more complete analysis of the interactions between the DM's incentives to acquire a private signal and her incentives to commit to ignore the sender is, however, beyond the scope of this paper.

The former is more likely when the quality of the decision maker's research is low and the prior belief is against the sender's interests, while the latter occurs when the evidence acquired by the decision maker is of high quality and the prior belief is not too much in favour of the sender's agenda. Finally, we have demonstrated how acquisition of independent evidence may completely deter falsification, and have shown that better quality of one's own research may not always benefit the decision maker.

In the second part of the paper, we have assumed that the decision maker can commit to ignore the sender's message with some probability. By doing so, the decision maker can decrease the chances that the sender's message will be pivotal to the decision, and hence may decrease his incentives to falsify evidence. We have fully characterised the conditions under which such a strategy is optimal for the sender. In general, for it to be optimal, the cost of falsification cannot be too high or too low if falsification is easy to detect, and it cannot be too high if falsification is difficult to detect. Moreover, other things being equal, the prior belief must be to a sufficient extent against the sender's interests. Finally, we have suggested how this commitment ability of the decision maker can affect her incentives to acquire independent evidence.

Future research could investigate in more detail how the decision maker's ability to control the amount of information she receives from an interest group interacts with her incentives to conduct independent research. This analysis would shed more light on the circumstances under which the two measures analysed in this paper are substitutes or complements.

Appendix to Chapter 3

C.1 Commitment to Ignorance in a Model with a Continuous Choice of Falsification by the Sender

In order to gain a better intuition for the results on the DM's incentives to commit to ignore the sender's message (which are the subject of Section 3.4), it is useful to consider a model with a continuous—rather than a binary—choice of falsification by the sender. Suppose that the sender can choose the "intensity" of falsification, which we denote by $x \in [0,1]$. When he falsifies with intensity x, the probability that the DM observes $\widetilde{m} \neq s$ from the sender is x, and the probability that she observes $\widetilde{m} = s$ is $1-x$.[22] Since we assume here that $q = 1$, the signal received by the sender perfectly reveals the state of the world. The cost of falsifying with intensity x is $c(x)$, where the marginal cost is $c'(x) = MC(x) > 0$ with $MC(0) = 0$ and $MC(1) \geq 1$. We also assume that $p \geq \frac{x}{1+x}$ (or, equivalently, $x \leq \frac{p}{1-p}$), so that the DM optimally accepts with probability 1 upon seeing $\widetilde{m} = H$.

The sender's strategy, σ_s, now effectively describes the probability with which the sender falsifies upon observing a signal s, with mixed strategies allowed. It is a function

$$\sigma : \{L, H\} \to [0,1]. \tag{3.13}$$

As before, the DM's strategy, $\delta_{\tilde{m}}$, describes the probability with which she accepts. It is a function

$$\delta : \{L, H\} \to \Delta\{A, R\}. \tag{3.14}$$

Proposition 3.11. *Suppose $p \geq \frac{x}{1+x}$. In the model with a continuous choice of falsification intensity by the sender, it is optimal for the DM to commit to ignore the sender's*

[22] Thus, parameter x has now a slightly different interpretation than before.

message more often (i.e. to decrease τ) if and only if

$$\frac{\partial \left[MC^{-1}(\tau)\,\tau\right]}{\partial \tau} \geq \min\left\{\frac{p}{1-p}, 1\right\}. \tag{3.15}$$

Consider first the case where $p < \frac{1}{2}$. If the DM commits to pay attention to the sender's message with probability τ, her expected payoff is:

$$\mathbb{E}\left(v_\tau\left(\sigma^*, \delta^*\right)\right) = \left(p + (1-p)\left(1 - MC^{-1}(\tau)\right)\right)\tau + (1-p)(1-\tau). \tag{3.16}$$

With probability τ, the DM makes her decision based on the message received from the sender, who—upon seeing that the state is $\theta = \theta_L$—falsifies evidence with intensity x such that $MC(x) = \tau$, i.e. with intensity $x^*(\tau) = MC^{-1}(\tau)$. Thus, when the state is $\theta = \theta_L$, the DM receives $\widetilde{m} = H$ with probability $x^* = MC^{-1}(\tau)$, and $\widetilde{m} = L$ otherwise. When the state is $\theta = \theta_H$, the DM receives $\widetilde{m} = H$ with probability 1.

With probability $1 - \tau$, the DM ignores the sender's message and therefore makes her decision solely based on her prior. Given that $p < \frac{1}{2}$, she then chooses to reject, which is the "correct" action with probability $1 - p$.

The trade-off faced by the DM is as follows. By committing to ignore the sender's message with some probability, she ensures that the signal she receives is more precise—since the sender is falsifying with less intensity. This is because the sender's optimal choice of the intensity, $x^*(\tau)$, is increasing in τ. However, ignoring the sender's message more often also has a cost: the DM needs to make the decision based only on her prior information more often. If the rate at which the signal becomes more precise as τ falls is sufficiently high, it is optimal for the DM to commit to ignore the sender's message more often.

To be more concrete, for any given level of τ, it is optimal for the DM to commit to ignore the sender's message more often if and only if $\partial \mathbb{E}\left(v_\tau\left(\sigma^*, \delta^*\right)\right)/\partial \tau \leq 0$, which

yields

$$\frac{\partial \left[MC^{-1}(\tau)\,\tau\right]}{\partial \tau} \geq \frac{p}{1-p}. \tag{3.17}$$

Figure 3.8 helps visualise this. When the DM commits to ignore the sender's message slightly more often, she marginally decreases the value of τ. This leads to a decrease in the area of the grey rectangle (the area is equal to $MC^{-1}(\tau) \times \tau$). How large this decrease is depends on the slope of the marginal cost function, $MC(x)$. If $MC(x)$ has a steep slope, i.e. $MC'(x)$ is high, then a small decrease in τ results in a small decrease in $MC^{-1}(\tau)$, and so the area of the rectangle falls by little. If the slope of the $MC(x)$ curve is shallow, i.e. $MC'(x)$ is low, then a small decrease in τ results in a large decrease in $MC^{-1}(\tau)$, and thus the area of the rectangle falls significantly. Hence, committing to ignore more often is likely to be effective from the DM's point of view when the sender's marginal cost of increasing x does not rise too fast as x goes up.

An analogous analysis holds for the case where $p \geq \frac{1}{2}$. If the DM commits to pay attention to the sender's message with probability τ, her expected payoff is:

$$\mathbb{E}\left(v_\tau\left(\sigma^*, \delta^*\right)\right) = \left(p + (1-p)\left(1 - MC^{-1}(\tau)\right)\right)\tau + p\,(1-\tau)\,. \tag{3.18}$$

As before, with probability τ, the DM makes her decision based on the message received from the sender, who falsifies evidence with intensity $x^*(\tau) = MC^{-1}(\tau)$ upon seeing that the state is $\theta = \theta_L$. However, when the DM ignores the sender's message (which happens with probability $1-\tau$) and makes her decision based on the prior only, she now chooses to accept, which is the "correct" action with probability p. Given (3.18), $\partial \mathbb{E}\left(v_\tau\left(\sigma^*, \delta^*\right)\right)/\partial \tau \leq 0$ holds if and only if

$$\frac{\partial \left[MC^{-1}(\tau)\,\tau\right]}{\partial \tau} \geq 1. \tag{3.19}$$

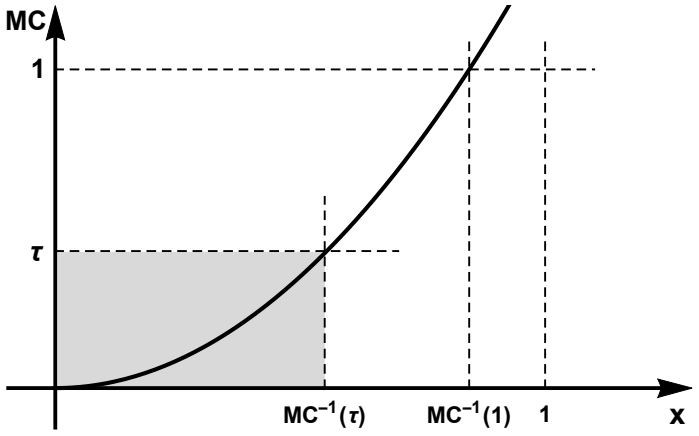


Figure 3.8: An illustration of the DM's problem of whether to commit to ignore the sender's message more often when the sender's choice of x is continuous.

Consistency with the binary choice model. The results in Proposition 3.11 are consistent with those for the model with a binary choice by the sender. To see this, note that when the DM considers whether to increase τ from $\frac{c}{x}$ to 1, we have

$$\frac{\Delta\left[x^*\left(\tau\right)\tau\right]}{\Delta\tau} = \frac{x^*\left(1\right)\times 1 - x^*\left(\frac{c}{x}\right)\times\frac{c}{x}}{1-\frac{c}{x}} = \frac{x\times 1 - 0\times\frac{c}{x}}{1-\frac{c}{x}} = \frac{x}{1-\frac{c}{x}}. \tag{3.20}$$

Hence, it follows from Proposition 3.11 that if $p < \frac{1}{2}$, the DM finds it optimal to choose $\tau = 1$ rather than $\tau = \frac{c}{x}$ if

$$\frac{\Delta\left[x^*\left(\tau\right)\tau\right]}{\Delta\tau} = \frac{x}{1-\frac{c}{x}} \geq \frac{p}{1-p}, \tag{3.21}$$

which can be rearranged to yield $c \leq x\left(1-\frac{1-p}{p}x\right)$.

Furthermore, Proposition 3.11 implies that if $p \geq \frac{1}{2}$, the DM finds it optimal to set $\tau = 1$ rather than $\tau = \frac{c}{\bar{x}}$ if

$$\frac{\Delta\left[x^*\left(\tau\right)\tau\right]}{\Delta\tau} = \frac{x}{1-\frac{c}{x}} \geq 1, \tag{3.22}$$

which is equivalent to $c \leq x\,(1-x)$. We thus obtain exactly the same result as the one in Proposition 3.10.

C.2 Omitted Proofs

Proof of Proposition 3.1

It is straightforward to show that it must be that $\sigma_H^* = 0$ in any equilibrium.

Since $\beta\,(\widetilde{m} = L \mid \sigma_L = 0) = 0$ for any σ_L, the DM's sequentially rational (henceforth, abbreviated to "s.r.") strategy satisfies $\delta_L = 0$. Thus, it must be $\delta_L^* = 0$ in any equilibrium. It remains to derive σ_L^* and δ_H^*. To do so, we consider two mutually exclusive and exhaustive regions of parameter values, and show that there is a unique equilibrium strategy profile in each of them.

1. Consider parameter values such that $\beta\,(\widetilde{m} = H \mid \sigma_L = 1) \geq \frac{1}{2}$, which is equivalent to $p \geq \frac{x}{1+x}$. For any $\sigma_L \in [0,1]$ and $\sigma_H = 0$, given that $\beta\,(\widetilde{m} = H \mid \sigma_L = 1) \geq \frac{1}{2}$, the DM's s.r. strategy satisfies $\delta_H = 1$. Now, given that $c < x$, $\delta_L = 0$, and $\delta_H = 1$, the sender's expected payoff from falsifying upon observing $s = L$ is $x\,(\delta_H - \delta_L) - c = x - c > 0$, and thus the sender's s.r. strategy satisfies $\sigma_L = 1$. Thus, for $p \geq \frac{x}{1+x}$, the unique equilibrium strategies are $\sigma_L^* = 1$ and $\delta_H^* = 1$.

2. Consider parameter values such that $\beta\,(\widetilde{m} = H \mid \sigma_L = 1) < \frac{1}{2}$, which is equivalent to $p < \frac{x}{1+x}$.

   (a) If we suppose that $\sigma_L = 1$, the DM's s.r. strategy satisfies $\delta_H = 0$, which makes $\sigma_L = 1$ not s.r., a contradiction.

   (b) If we suppose that $\sigma_L = 0$, the DM's s.r. strategy satisfies $\delta_H = 1$, which makes $\sigma_L = 0$ not s.r., a contradiction.

(c) Suppose now that $\sigma_L \in (0,1)$. If $\beta(\widetilde{m} = H \mid \sigma_L) > \frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_H = 1$, which makes $\sigma_L \in (0,1)$ not s.r., a contradiction. If $\beta(\widetilde{m} = H \mid \sigma_L) < \frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_H = 0$, which makes $\sigma_L \in (0,1)$ not s.r., a contradiction. Suppose then that $\beta(\widetilde{m} = H \mid \sigma_L) = \frac{1}{2}$, which pins down $\sigma_L \in (0,1)$ to $\sigma_L = \frac{p}{(1-p)x}$. If $x(\delta_H - \delta_L) - c > 0$, where $\delta_L = 0$, then $\sigma_L \in (0,1)$ is not s.r. ($\sigma_L = 1$ is). If $x(\delta_H - \delta_L) - c < 0$, then $\sigma_L \in (0,1)$ is not s.r. ($\sigma_L = 0$ is). If $x(\delta_H - \delta_L) - c = 0$, then the sender is indifferent between falsifying and not falsifying upon observing $s = L$, so $\sigma_L = \frac{p}{(1-p)x} \in (0,1)$ is s.r. This uniquely pins down $\delta_H = \frac{c}{x}$. Thus, for $p < \frac{x}{1+x}$, the unique equilibrium strategies are $\sigma_L^* = \frac{p}{(1-p)x}$ and $\delta_H^* = \frac{c}{x}$.

Proof of Proposition 3.2

Part (i). When $x \to 0$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_L^*, \delta_H^*) = (0,1)$ for all $p \in (0,1)$. Given that $x \to 0$ and these equilibrium strategies, we have $\mathbb{E}(v(\sigma^*, \delta^*)) \to p \cdot 1 + (1-p) \cdot 1 = 1$, $\Pr(a = A \mid \theta = \theta_H) \to 1$, and $\Pr(a = A \mid \theta = \theta_L) \to 0$.

In a model of verifiable disclosure, in equilibrium, the sender effectively truthfully reveals $s = H$ and $s = L$, so the DM learns s and takes action accordingly. This gives us $\mathbb{E}(v(\sigma^*, \delta^*)) = p \cdot 1 + (1-p) \cdot 1 = 1$, $\Pr(a = A \mid \theta = \theta_H) = 1$, and $\Pr(a = A \mid \theta = \theta_L) = 0$.

Part (ii). When $x \to 1$ and $c \to 0$, the equilibrium strategies are $\sigma_L^* \to \frac{p}{1-p}$, $\sigma_H^* = 0$, $\delta_L^* = 0$, $\delta_H^* \to 0$ for $p \in \left(0, \frac{1}{2}\right)$, and $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_L^*, \delta_H^*) = (0,1)$ for $p \in \left[\frac{1}{2}, 1\right)$. Given that $x \to 1$ and $c \to 0$ and these equilibrium strategies, $\mathbb{E}(v(\sigma^*, \delta^*)) \to 1 - p$ for $p \in \left(0, \frac{1}{2}\right)$, $\mathbb{E}(v(\sigma^*, \delta^*)) \to p$ for $p \in \left[\frac{1}{2}, 1\right)$, $\Pr(a = A \mid \theta = \theta_H) \to 0$ for $p \in \left(0, \frac{1}{2}\right)$, $\Pr(a = A \mid \theta = \theta_H) \to 1$ for $p \in \left[\frac{1}{2}, 1\right)$, $\Pr(a = A \mid \theta = \theta_L) \to 0$ for $p \in \left(0, \frac{1}{2}\right)$, and

$\Pr(a = A \mid \theta = \theta_L) \to 1$ for $p \in \left[\frac{1}{2}, 1\right)$.

In a model of cheap talk, given the misaligned preferences of the sender and the DM, in equilibrium, the sender sends an uninformative ("babbling") message to the DM regardless of the observed s. Effectively, the DM does not learn anything about s and keeps her prior belief, and takes action accordingly. This gives us $\mathbb{E}(v(\sigma^*, \delta^*)) = 1 - p$ for $p \in \left(0, \frac{1}{2}\right)$, $\mathbb{E}(v(\sigma^*, \delta^*)) = p$ for $p \in \left[\frac{1}{2}, 1\right)$, $\Pr(a = A \mid \theta = \theta_H) = 0$ for $p \in \left(0, \frac{1}{2}\right)$, $\Pr(a = A \mid \theta = \theta_H) = 1$ for $p \in \left[\frac{1}{2}, 1\right)$, $\Pr(a = A \mid \theta = \theta_L) = 0$ for $p \in \left(0, \frac{1}{2}\right)$, and $\Pr(a = A \mid \theta = \theta_L) = 1$ for $p \in \left[\frac{1}{2}, 1\right)$.

Part (iii). When $x \to 1$ and $c \to 1$, the equilibrium strategies are $\sigma_L^* \to \frac{p}{1-p}$, $\sigma_H^* = 0$, $\delta_L^* = 0$, $\delta_H^* \to 1$ for $p \in \left(0, \frac{1}{2}\right)$, and $(\sigma_L^*, \sigma_H^*) = (1, 0)$ and $(\delta_L^*, \delta_H^*) = (0, 1)$ for $p \in \left[\frac{1}{2}, 1\right)$. Given that $x \to 1$ and $c \to 1$ and these equilibrium strategies, $\mathbb{E}(v(\sigma^*, \delta^*)) \to 1 - p$ for $p \in \left(0, \frac{1}{2}\right)$, $\mathbb{E}(v(\sigma^*, \delta^*)) \to p$ for $p \in \left[\frac{1}{2}, 1\right)$, $\Pr(a = A \mid \theta = \theta_H) \to 1$ for $p \in (0, 1)$, $\Pr(a = A \mid \theta = \theta_L) \to \frac{p}{1-p}$ for $p \in \left(0, \frac{1}{2}\right)$, and $\Pr(a = A \mid \theta = \theta_L) \to 1$ for $p \in \left[\frac{1}{2}, 1\right)$.

In a model of Bayesian persuasion, note that the receiver is willing to take action $a = A$ if $\Pr(\theta = \theta_H \mid m = H) \geq \frac{1}{2}$ (we assume here that the DM takes action $a = A$ when indifferent). Therefore, the sender (the designer) commits to the following signal structure: $\Pr(m = H \mid \theta = \theta_H) = 1$ for $p \in (0, 1)$, $\Pr(m = H \mid \theta = \theta_L) = \frac{p}{1-p}$ for $p \in \left(0, \frac{1}{2}\right)$, and $\Pr(m = H \mid \theta = \theta_L) = 1$ for $p \in \left[\frac{1}{2}, 1\right)$. Given this strategy of the sender, the DM's posterior beliefs are $\Pr(\theta = \theta_H \mid m = H) = \frac{1}{2}$ when $p < \frac{1}{2}$, and $\Pr(\theta = \theta_H \mid m = H) = p$ when $p \geq \frac{1}{2}$, which means that (i) the DM's strategy is to take action $a = A$ upon observing $m = H$ and $a = R$ upon observing $m = L$, and (ii) the probability of the DM accepting conditional on the state of the world is maximised for any given p. This gives us $\mathbb{E}(v(\sigma^*, \delta^*)) = 1 - p$ for $p \in \left(0, \frac{1}{2}\right)$, $\mathbb{E}(v(\sigma^*, \delta^*)) = p$ for $p \in \left[\frac{1}{2}, 1\right)$, $\Pr(a = A \mid \theta = \theta_H) = 1$ for $p \in (0, 1)$, $\Pr(a = A \mid \theta = \theta_L) = \frac{p}{1-p}$ for $p \in \left(0, \frac{1}{2}\right)$, and $\Pr(a = A \mid \theta = \theta_L) = 1$ for $p \in \left[\frac{1}{2}, 1\right)$.

Proof of Proposition 3.3

Let p^- denote the DM's belief upon observing $s_{DM} = L$, i.e. $p^- = \frac{p(1-q_{DM})}{p(1-q_{DM})+(1-p)q_{DM}}$, and let p^+ denote the DM's belief upon observing $s_{DM} = H$, i.e. $p^+ = \frac{pq_{DM}}{pq_{DM}+(1-p)(1-q_{DM})}$. It is straightforward to show that it must be that $\sigma_H^* = 0$ in any equilibrium. Since $\beta\left(s_{DM} = L, \widetilde{m} = L \mid \sigma_L\right) = \beta\left(s_{DM} = H, \widetilde{m} = L \mid \sigma_L\right) = 0$ for any σ_L and $\sigma_H^* = 0$, the DM's sequentially rational (henceforth, abbreviated to "s.r.") strategy satisfies $\delta_{LL} = \delta_{HL} = 0$. Thus, it must be $\delta_{LL}^* = \delta_{HL}^* = 0$ in any equilibrium. It remains to derive σ_L^*, δ_{LH}^*, and δ_{HH}^*. To do so, we consider four mutually exclusive regions of parameter values, and show that there is a unique equilibrium strategy profile in each of them. In the remaining knife-edge cases, we select the sender-preferred strategy profile whenever there are multiple equilibria.

1. Consider parameter values such that $\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 1\right) > \frac{1}{2}$, which is equivalent to $p^- > \frac{x}{1+x}$. For any $\sigma_L \in [0,1]$ and $\sigma_H = 0$, given that we have $\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 1\right) > \frac{1}{2}$ (and thus $\beta\left(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 1\right) > \frac{1}{2}$), the DM's s.r. strategy satisfies $\delta_{LH} = 1$ and $\delta_{HH} = 1$. Now, given that $c < x$, the sender's expected payoff from falsifying upon observing $s = L$ is $q_{DM}x\left(\delta_{LH} - \delta_{LL}\right) + \left(1 - q_{DM}\right)x\left(\delta_{HH} - \delta_{HL}\right) - c = x - c > 0$, and thus the sender's s.r. strategy satisfies $\sigma_L = 1$. Thus, for $p^- > \frac{x}{1+x}$, the unique equilibrium strategy profile satisfies $\sigma_L^* = 1$, $\delta_{LH}^* = 1$, and $\delta_{HH}^* = 1$.

2. Consider parameter values such that $\beta\left(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 1\right) < \frac{1}{2}$ (which is equivalent to $p^- < \frac{x}{1+x}$), $\beta\left(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 1\right) > \frac{1}{2}$ (which is equivalent to $p^+ > \frac{x}{1+x}$), and $q_{DM} < 1 - \frac{c}{x}$. For any $\sigma_L \in [0,1]$ and $\sigma_H = 0$, given that $\beta\left(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 1\right) > \frac{1}{2}$, the DM's s.r. strategy satisfies $\delta_{HH} = 1$. Now, given that $q_{DM} < 1 - \frac{c}{x}$, the sender's expected payoff from falsifying upon observing $s = L$ is $q_{DM}x(\delta_{LH} - \delta_{LL}) + (1 - q_{DM})x(\delta_{HH} - \delta_{HL}) - c = q_{DM}x\delta_{LH} +$

$(1-q_{DM})x-c>0$, and thus the sender's s.r. strategy satisfies $\sigma_L=1$. Thus, for $p^-<\frac{x}{1+x}$, $p^+>\frac{x}{1+x}$, and $c<(1-q_{DM})x$, the unique equilibrium strategy profile satisfies $\sigma_L^*=1$, $\delta_{LH}^*=0$, and $\delta_{HH}^*=1$.

3. Consider parameter values such that $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L=1\right)<\frac{1}{2}$ (which is equivalent to $p^-<\frac{x}{1+x}$), $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L=1\right)<\frac{1}{2}$ (which is equivalent to $p^+<\frac{x}{1+x}$), and $q_{DM}<1-\frac{c}{x}$. If we suppose that $\sigma_L=1$, the DM's s.r. strategy satisfies $\delta_{LH}=0$ and $\delta_{HH}=0$, which makes $\sigma_L=1$ not s.r., a contradiction. If we suppose that $\sigma_L=0$, the DM's s.r. strategy satisfies $\delta_{LH}=1$ and $\delta_{HH}=1$, which makes $\sigma_L=0$ not s.r., a contradiction. Suppose now that $\sigma_L\in(0,1)$. There are then five possibilities to consider.

   (a) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=1$ and $\delta_{HH}=1$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x(\delta_{LH}-\delta_{LL})+(1-q_{DM})x(\delta_{HH}-\delta_{HL})-c=x-c>0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

   (b) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=0$ and $\delta_{HH}=1$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x(\delta_{LH}-\delta_{LL})+(1-q_{DM})x(\delta_{HH}-\delta_{HL})-c=(1-q_{DM})x-c>0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

   (c) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=0$ and $\delta_{HH}=0$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x(\delta_{LH}-\delta_{LL})+(1-q_{DM})x(\delta_{HH}-\delta_{HL})-c=-c<0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

(d) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)=\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{HH}=1$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x\delta_{LH}+\left(1-q_{DM}\right)x-c>0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

(e) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)=\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=0$ and σ_L is pinned down to $\sigma_L=\frac{p^+}{(1-p^+)x}$. If either $\left(1-q_{DM}\right)x\delta_{HH}-c<0$ or $\left(1-q_{DM}\right)x\delta_{HH}-c>0$, then $\sigma_L\in(0,1)$ is not s.r., a contradiction. If $\left(1-q_{DM}\right)x\delta_{HH}-c=0$, the sender is indifferent between falsifying and not falsifying upon observing $s=L$, so $\sigma_L=\frac{p^+}{(1-p^+)x}\in(0,1)$ is s.r. This uniquely pins down $\delta_{HH}=\frac{c}{(1-q_{DM})x}$. Thus, for $p^-<\frac{x}{1+x}$, $p^+<\frac{x}{1+x}$, and $q_{DM}<1-\frac{c}{x}$, the unique equilibrium strategy profile satisfies $\sigma_L^*=\frac{p^+}{(1-p^+)x}$, $\delta_{LH}^*=0$, and $\delta_{HH}^*=\frac{c}{(1-q_{DM})x}$.

4. Consider parameter values such that $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L=1\right)<\frac{1}{2}$, which is equivalent to $p^-<\frac{x}{1+x}$, and $q_{DM}>1-\frac{c}{x}$. If we suppose that $\sigma_L=1$, the DM's s.r. strategy satisfies $\delta_{LH}=0$, so the sender's expected payoff from falsifying upon observing $s=L$ is $\left(1-q_{DM}\right)x\delta_{HH}-c<0$, which makes $\sigma_L=1$ not s.r., a contradiction. If we suppose that $\sigma_L=0$, the DM's unique s.r. action is $\delta_{LH}=1$, so the sender's expected payoff from falsifying upon observing $s=L$ is $x-c>0$, which makes $\sigma_L=0$ not s.r., a contradiction. Suppose now that $\sigma_L\in(0,1)$. There are then five possibilities to consider.

(a) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=1$ and $\delta_{HH}=1$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x(\delta_{LH}-\delta_{LL})+(1-q_{DM})x(\delta_{HH}-\delta_{HL})-c=x-c>0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

(b) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=0$ and $\delta_{HH}=1$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x(\delta_{LH}-\delta_{LL})+(1-q_{DM})x(\delta_{HH}-\delta_{HL})-c=(1-q_{DM})x-c<0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

(c) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{LH}=0$ and $\delta_{HH}=0$, so the sender's expected payoff from falsifying upon observing $s=L$ is $q_{DM}x(\delta_{LH}-\delta_{LL})+(1-q_{DM})x(\delta_{HH}-\delta_{HL})-c=-c<0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

(d) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)<\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)=\frac{1}{2}$, the DM's s.r. strategy satisfies $\delta_{LH}=0$, so the sender's expected payoff from falsifying upon observing $s=L$ is $(1-q_{DM})x\delta_{HH}-c<0$, which makes $\sigma_L\in(0,1)$ not s.r., a contradiction.

(e) If $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L\right)=\frac{1}{2}$ and $\beta\left(s_{DM}=H,\widetilde{m}=H\mid\sigma_L\right)>\frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_{HH}=1$ and σ_L is pinned down to $\sigma_L=\frac{p^-}{(1-p^-)x}$. If either $qx\delta_{LH}-c<0$ or $qx\delta_{LH}-c>0$, then $\sigma_L\in(0,1)$ is not s.r., a contradiction. If $qx\delta_{LH}-c=0$, the sender is indifferent between falsifying and not falsifying upon observing $s=L$, so $\sigma_L=\frac{p^-}{(1-p^-)x}\in(0,1)$ is s.r. This uniquely pins down $\delta_{LH}=\frac{c}{q_{DM}x}$. Thus, for $p^-<\frac{x}{1+x}$ and $q_{DM}>1-\frac{c}{x}$, the unique equilibrium strategy profile satisfies $\sigma_L^*=\frac{p^-}{(1-p^-)x}$, $\delta_{LH}^*=\frac{c}{q_{DM}x}$, and $\delta_{HH}^*=1$.

Proof of Proposition 3.4

Using the results from Proposition 3.1 and 3.3, we can compare the sender's equilibrium strategy in the baseline model and in the model in which the DM observes a private

signal, for all possible regions of parameter values. The results in Proposition 3.4 follow from Table 3.1. Note here that, since $p^- < p < p^+$, which follows from $q_{DM} > \frac{1}{2}$, we also have $\frac{p^-}{(1-p^-)x} < \frac{p}{(1-p)x} < \frac{p^+}{(1-p^+)x}$.

Proof of Proposition 3.5

Using the results from Proposition 3.1 and 3.3, we can compare the DM's welfare in the baseline model and in the model in which the DM observes a private signal of quality q_{DM}, for all possible regions of parameter values. For brevity, let $W\left(\frac{1}{2}\right) = \mathbb{E}\left[v\left(\sigma^*\left(\frac{1}{2}\right), \delta^*\left(\frac{1}{2}\right)\right)\right]$ be the DM's welfare in the baseline model, and let $W(q_{DM}) = \mathbb{E}\left[v\left(\sigma^*\left(q_{DM}\right), \delta^*\left(q_{DM}\right)\right)\right]$ be the DM's welfare in the model where she observes a private signal of quality q_{DM}. Thus, $V = W\left(q_{DM}\right) - W\left(\frac{1}{2}\right)$. We consider five mutually exclusive and exhaustive regions of parameter values.

1. Consider parameter values such that $p^- \geq \frac{x}{1+x}$. When $q_{DM} = \frac{1}{2}$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = (1, 0)$ and $(\delta_L^*, \delta_H^*) = (0, 1)$. When $q_{DM} > \frac{1}{2}$, they are $(\sigma_L^*, \sigma_H^*) = (1, 0)$ and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = (0, 0, 1, 1)$. We can then derive $W\left(\frac{1}{2}\right) = p + (1-p)(1-x)$ and $W(q_{DM}) = p + (1-p)(1-x)$. Hence, $V = 0$.

2. Consider parameter values such that $q_{DM} < 1 - \frac{c}{x}$ and $p^+ < \frac{x}{1+x}$. When $q_{DM} = \frac{1}{2}$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = \left(\frac{p}{(1-p)x}, 0\right)$ and $(\delta_L^*, \delta_H^*) = \left(0, \frac{c}{x}\right)$. When $q_{DM} > \frac{1}{2}$, they are $(\sigma_L^*, \sigma_H^*) = \left(\frac{p^+}{(1-p^+)x}, 0\right)$ and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = \left(0, 0, 0, \frac{c}{(1-q_{DM})x}\right)$. We can then derive $W\left(\frac{1}{2}\right) = 1 - p$ and $W(q_{DM}) = 1 - p$. Hence, $V = 0$.

3. Consider parameter values such that $q_{DM} < 1 - \frac{c}{x}$ and $p < \frac{x}{1+x} \leq p^+$. When $q_{DM} = \frac{1}{2}$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = \left(\frac{p}{(1-p)x}, 0\right)$ and $(\delta_L^*, \delta_H^*) = \left(0, \frac{c}{x}\right)$. When $q_{DM} > \frac{1}{2}$, they are $(\sigma_L^*, \sigma_H^*) = (1, 0)$ and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = (0, 0, 0, 1)$. We can then derive $W\left(\frac{1}{2}\right) = 1 - p$, $W(q_{DM}) = p\left[q_{DM}\right] + (1 -$

$p)\left[(1-x)+q_{DM}x\right]$. The condition $p^+ \geq \frac{x}{1+x}$ implies that $V = W(q_{DM}) - W\left(\frac{1}{2}\right) = pq_{DM} - (1-p)(1-q_{DM})x \geq 0$, with the inequality being strict whenever $p^+ > \frac{x}{1+x}$.

4. Consider parameter values such that $q_{DM} < 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x} \leq p$. When $q_{DM} = \frac{1}{2}$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_L^*, \delta_H^*) = (0,1)$. When $q_{DM} > \frac{1}{2}$, they are $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = (0,0,0,1)$. We can then derive $W\left(\frac{1}{2}\right) = p + (1-p)(1-x)$ and $W(q_{DM}) = p\left[q_{DM}\right] + (1-p)\left[(1-x)+q_{DM}x\right]$. The condition that $p^- < \frac{x}{1+x}$ implies that $V = W(q_{DM}) - W\left(\frac{1}{2}\right) = -p(1-q_{DM}) + (1-p)q_{DM}x > 0$.

5. Consider parameter values such that $q_{DM} \geq 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x}$.

   (a) Suppose that $p^- < \frac{x}{1+x} \leq p$. When $q_{DM} = \frac{1}{2}$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = (1,0)$ and $(\delta_L^*, \delta_H^*) = (0,1)$. When $q_{DM} > \frac{1}{2}$, they are $(\sigma_L^*, \sigma_H^*) = \left(\frac{p^-}{(1-p^-)x}, 0\right)$ and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = \left(0, 0, \frac{c-(1-q_{DM})x}{q_{DM}x}, 1\right)$. We can then find $W\left(\frac{1}{2}\right) = p + (1-p)(1-x)$ and $W(q_{DM}) = p + (1-p)\left[(1-\sigma_L^*) + \sigma_L^*(1-x)\right]$, where $\sigma_L^* = \frac{p^-}{(1-p^-)x}$. It follows that $V = W(q_{DM}) - W\left(\frac{1}{2}\right) = (1-p)(1-\sigma_L^*)x > 0$, where $\sigma_L^* = \frac{p^-}{(1-p^-)x}$.

   (b) Suppose that $p < \frac{x}{1+x}$. When $q_{DM} = \frac{1}{2}$, the equilibrium strategies are $(\sigma_L^*, \sigma_H^*) = \left(\frac{p}{(1-p)x}, 0\right)$ and $(\delta_L^*, \delta_H^*) = \left(0, \frac{c}{x}\right)$. When $q_{DM} > \frac{1}{2}$, they are $(\sigma_L^*, \sigma_H^*) = \left(\frac{p^-}{(1-p^-)x}, 0\right)$ and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = \left(0, 0, \frac{c-(1-q_{DM})x}{q_{DM}x}, 1\right)$. We can then find $W\left(\frac{1}{2}\right) = 1-p$ and $W(q_{DM}) = p+(1-p)\left[(1-\sigma_L^*) + \sigma_L^*(1-x)\right]$, where $\sigma_L^* = \frac{p^-}{(1-p^-)x}$. It follows that $V = W(q_{DM}) - W\left(\frac{1}{2}\right) = p - \frac{1-p}{1-p^-}p^- > 0$.

Proof of Corollary 3.1

First, we consider the DM's expected payoff in five mutually exclusive and exhaustive regions of the parameter values. Let $W(q_{DM})$ denote the DM's welfare in the model in

which she observes a private signal of quality q_{DM}, $\mathbb{E}\left[v\left(\sigma^*\left(q_{DM}\right), \delta^*\left(q_{DM}\right)\right)\right]$.

1. Consider parameter values such that $p^- \geq \frac{x}{1+x}$. The DM's welfare is $W(q_{DM}) = p + (1-p)(1-x)$. Then, $\frac{\partial W(q_{DM})}{\partial q_{DM}} = 0$.

2. Consider parameter values such that $q_{DM} < 1 - \frac{c}{x}$ and $p^+ < \frac{x}{1+x}$. The DM's welfare is $W(q_{DM}) = 1 - p$. Then $\frac{\partial W(q_{DM})}{\partial q_{DM}} = 0$.

3. Consider parameter values such that $q_{DM} < 1 - \frac{c}{x}$ and $p < \frac{x}{1+x} \leq p^+$. The DM's welfare is $W(q_{DM}) = p\left[q_{DM}\right] + (1-p)\left[(1-x) + q_{DM}x\right]$. Then, $\frac{\partial W(q_{DM})}{\partial q_{DM}} = p + (1-p)x > 0$.

4. Consider parameter values such that $q_{DM} < 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x} \leq p$. The DM's welfare is $W(q_{DM}) = p\left[q_{DM}\right] + (1-p)\left[(1-x) + q_{DM}x\right]$. Then, $\frac{\partial W(q_{DM})}{\partial q_{DM}} = p + (1-p)x > 0$.

5. Consider parameter values such that $q_{DM} \geq 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x}$. The DM's welfare is $W(q_{DM}) = p + (1-p)\left[(1-\sigma_L^*) + \sigma_L^*(1-x)\right]$, where $\sigma_L^* = \frac{p^-}{(1-p^-)x}$. Then, $\frac{\partial W(q_{DM})}{\partial q_{DM}} = (1-p)(-x)\frac{\partial \sigma_L^*}{\partial q_{DM}}$, where $\frac{\partial \sigma_L^*}{\partial q_{DM}} = \frac{p(p-q_{DM})x}{[(1-p)q_{DM}x]}$. From $p^- < \frac{x}{1+x}$ it follows that $p^- < \frac{1}{2}$, which is equivalent to $p < q_{DM}$. Hence, $\frac{\partial \sigma_L^*}{\partial q_{DM}} < 0$, and so $\frac{\partial W(q_{DM})}{\partial q_{DM}} > 0$.

Second, we consider whether the DM's expected payoff is increasing in q_{DM} at the boundaries of the above five regions. It suffices to check here (as can also be seen from Figure 3.4) whether

$$\lim_{q_{DM} \to \left(1-\frac{c}{x}\right)^+} W(q_{DM}) \geq \lim_{q_{DM} \to \left(1-\frac{c}{x}\right)^-} W(q_{DM}) \tag{3.23}$$

is satisfied when $p^- < \frac{x}{1+x} < p^+$. To start with, note that the DM's equilibrium strategy is $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = (0,0,0,1)$ when $q_{DM} < 1 - \frac{c}{x}$, and that it is $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) =$

$\left(0, 0, \frac{c-(1-q_{DM})x}{q_{DM}x}, 1\right)$ when $q_{DM} \geq 1 - \frac{c}{x}$. Because in the latter case the DM is indifferent between the two actions upon observing $s_{DM} = L$ and $\widetilde{m} = H$, the DM's expected payoff would remain unchanged if she took action $a = A$ with prob. 1 if and only if $s_{DM} = H$ and $\widetilde{m} = H$, in which case the mapping from s_{DM} and $\widetilde{m}$ to actions would be exactly the same as when $q_{DM} < 1 - \frac{c}{x}$. However, the sender's equilibrium strategy when $q_{DM} < 1 - \frac{c}{x}$ is $(\sigma_L^*, \sigma_H^*) = (1, 0)$, which means that it involves more falsification than the sender's equilibrium strategy when $q_{DM} \geq 1 - \frac{c}{x}$, which is $(\sigma_L^*, \sigma_H^*) = \left(\frac{p^-}{(1-p^-)x}, 0\right)$. Hence, (3.23) holds.

Proof of Proposition 3.6

Let us denote by $\sigma_{s_{pub}s}$ the sender's strategy for given realisations of the public signal, s_{pub}, and the sender's private signal, s. Similarly, let us denote by $\delta_{s_{pub}s}$ the sender's strategy for given realisations of s_{pub} and s. We consider three mutually exclusive regions of parameter values; as before, in the knife-edge cases, we select the sender-preferred strategy profile whenever there are multiple equilibria.

1. Consider parameter values such that $p^+ < \frac{x}{1+x}$. The equilibrium strategies are $(\sigma_{LL}^*, \sigma_{HL}^*) = \left(\frac{p^-}{(1-p^-)x}, \frac{p^+}{(1-p^+)x}\right)$ for the sender and $(\delta_{LL}^*, \delta_{HL}^*, \delta_{LH}^*, \delta_{HH}^*) = \left(0, 0, \frac{c}{x}, \frac{c}{x}\right)$ for the DM. Since the DM either strictly prefers to reject or is indifferent between accepting and rejecting, her expected payoff would remain unchanged if she rejected regardless of s_{pub} and $\widetilde{m}$. Hence, her expected payoff is

$$W_{pub}(q_{DM}) = \mathbb{E}\left[v\left(\sigma^*\left(q_{DM}\right), \delta^*\left(q_{DM}\right)\right)\right] = 1 - p. \tag{3.24}$$

   The algebraic expression for $W_{pub}\left(q_{DM}\right)$ is thus the same as $W\left(q_{DM}\right)$ for $q_{DM} < 1 - \frac{c}{x}$ and $p^+ < \frac{x}{1+x}$.

2. Consider parameter values such that $p^- < \frac{x}{1+x} \leq p^+$. The equilibrium strate-

gies are $(\sigma^*_{LL}, \sigma^*_{HL}) = \left(\frac{p^-}{(1-p^-)x}, 1\right)$ for the sender, and $(\delta^*_{LL}, \delta^*_{HL}, \delta^*_{LH}, \delta^*_{HH}) = \left(0, 0, \frac{c}{x}, 1\right)$ for the DM. The DM's expected payoff is then

$$W_{pub}(q_{DM}) = pq_{DM} + (1-p)\left[(1-q_{DM})(1-x) + q_{DM}\right] \quad (3.25)$$

$$= p\left[q_{DM}\right] + (1-p)\left[(1-x) + q_{DM}x\right]. \quad (3.26)$$

The algebraic expression for $W_{pub}(q_{DM})$ is thus the same as $W(q_{DM})$ for $q_{DM} < 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x} \leq p^+$.

3. Consider parameter values such that $p^- \geq \frac{x}{1+x}$. The equilibrium strategies are $(\sigma^*_{LL}, \sigma^*_{HL}) = (1, 1)$ for the sender, and $(\delta^*_{LL}, \delta^*_{HL}, \delta^*_{LH}, \delta^*_{HH}) = (0, 0, 1, 1)$ for the DM. The DM's expected payoff is then $W_{pub}(q_{DM}) = p + (1-p)(1-x)$. The algebraic expression for $W_{pub}(q_{DM})$ is thus the same as $W(q_{DM})$ for $p^- \geq \frac{x}{1+x}$.

To show that the DM's expected payoff is strictly lower with a public signal than with a private signal when $q_{DM} \geq 1 - \frac{c}{x}$ and $p^- < \frac{x}{1+x}$, consider two cases:

1. Consider parameter values such that $q_{DM} \geq 1 - \frac{c}{x}$ and $p^+ < \frac{x}{1+x}$. The DM's expected payoff is then $W_{pub}(q_{DM}) = 1 - p$ with a public signal, and $W(q_{DM}) = p + (1-p)\left[(1-\sigma^*_L) + \sigma^*_L(1-x)\right] = p + (1-p)\left(1 - \frac{p^-}{1-p^-}\right)$ with a private signal. It can be easily shown that the latter is strictly higher than the former (using the property that $p^- < p$).

2. Consider parameter values such that $q_{DM} \geq 1 - \frac{c}{x}$ and $p^+ \geq \frac{x}{1+x} > p^-$. The DM's expected payoff is then $W_{pub}(q_{DM}) = p\left[q_{DM}\right] + (1-p)\left[(1-x) + q_{DM}x\right]$ with a public signal, and $W(q_{DM}) = p + (1-p)\left[(1-\sigma^*_L) + \sigma^*_L(1-x)\right] = p + (1-p)\left(1 - \frac{p^-}{1-p^-}\right)$ with a private signal. Note that $W(q_{DM})$ can be rearranged to $p\left[q_{DM}\right] + (1-p)\left[1 - (1-q_{DM})\frac{p^-}{1-p^-}\right]$. It can then be easily shown that $W_{pub}(q_{DM}) > 0$, and also that $W(q_{DM}) > W_{pub}(q_{DM})$ for these parameter values.

Proof of Proposition 3.7

It is straightforward to show that in any equilibrium it must be that $\sigma_H^* = 0$. It remains to derive σ_L^*, δ_L^*, and δ_H^*. To do so, we consider four mutually exclusive regions of parameter values, and show that there is a unique equilibrium strategy profile in each of them. As before, in the knife-edge cases, we select the sender-preferred strategy profile whenever there are multiple equilibria.

1. Consider parameter values such that $\beta\left(\widetilde{m} = H \mid \sigma_L = 0\right) < \frac{1}{2}$, which is equivalent to $p < 1 - q$. For any $\sigma_L \in [0, 1]$ and $\sigma_H = 0$, given that $\beta\left(\widetilde{m} = H \mid \sigma_L = 0\right) < \frac{1}{2}$, the DM's sequentially rational (henceforth, abbreviated to "s.r.") strategy satisfies $\delta_H = 0$. From $\beta\left(\widetilde{m} = H \mid \sigma_L = 0\right) < \frac{1}{2}$, it also follows that $\beta\left(\widetilde{m} = L \mid \sigma_L = 0\right) < \frac{1}{2}$, so the DM's s.r. strategy satisfies $\delta_L = 0$. Given that $c > 0$, $\delta_L = 0$, and $\delta_H = 0$, the sender's expected payoff from falsifying upon observing $s = L$ is $x(\delta_H - \delta_L) - c = -c < 0$, so the sender's s.r. strategy satisfies $\sigma_L = 0$. Thus, for $p < 1 - q$, the unique equilibrium strategies are $\sigma_L^* = 0$, $\delta_L^* = 0$, and $\delta_H^* = 0$.

2. Consider $\beta\left(\widetilde{m} = H \mid \sigma_L = 0\right) > \frac{1}{2}$ and $\beta\left(\widetilde{m} = H \mid \sigma_L = 1\right) < \frac{1}{2}$. These two conditions yield parameter values satisfying $p \in \left(1 - q, \frac{1-(1-x)q}{1+x}\right)$.

   (a) If we suppose that $\sigma_L = 0$, then the DM's s.r. strategy satisfies $\delta_H = 1$, which makes $\sigma_L = 0$ not s.r., a contradiction.

   (b) If we suppose that $\sigma_L = 1$, then the DM's s.r. strategy satisfies $\delta_H = 0$, which makes $\sigma_L = 1$ not s.r., a contradiction.

   (c) Suppose now that $\sigma_L \in (0, 1)$. If $\beta\left(\widetilde{m} = H \mid \sigma_L\right) > \frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_H = 1$, which makes $\sigma_L \in (0, 1)$ not s.r., a contradiction. If $\beta\left(\widetilde{m} = H \mid \sigma_L\right) < \frac{1}{2}$, then the DM's s.r. strategy satisfies $\delta_H = 0$, which makes $\sigma_L \in (0, 1)$ not s.r., a contradiction. Therefore, suppose now that

$\beta(\widetilde{m} = H \mid \sigma_L) = \frac{1}{2}$, which pins down $\sigma_L \in (0,1)$ to $\sigma_L = \frac{p+q-1}{(q-p)x}$. If $x\delta_H - c > 0$, then $\sigma_L \in (0,1)$ is not s.r. ($\sigma_L = 1$ is). If $x\delta_H - c < 0$, then $\sigma_L \in (0,1)$ is not s.r. ($\sigma_L = 0$ is). If $x\delta_H - c = 0$, then the sender is indifferent between falsifying and not falsifying upon observing $s = L$, so $\sigma_L = \frac{p+q-1}{(q-p)x} \in (0,1)$ is s.r. This uniquely pins down $\delta_H = \frac{c}{x}$. Thus, for $p \in \left(1-q, \frac{1-(1-x)q}{1+x}\right)$, the unique equilibrium strategies satisfy $\sigma_L^* = \frac{p+q-1}{(q-p)x}$, $\delta_L^* = 0$, and $\delta_H^* = \frac{c}{x}$.

3. Consider parameter values such that $\beta(\widetilde{m} = H \mid \sigma_L = 1) > \frac{1}{2}$ and $\beta(\widetilde{m} = L) < \frac{1}{2}$, which together are equivalent to $p \in \left(\frac{1-(1-x)q}{1+x}, q\right)$. For any $\sigma_L \in [0,1]$ and $\sigma_H = 0$, given that $\beta(\widetilde{m} = H \mid \sigma_L = 1) > \frac{1}{2}$, the DM's s.r. strategy satisfies $\delta_H = 1$. Given that $\beta(\widetilde{m} = L) < \frac{1}{2}$, the DM's s.r. strategy satisfies $\delta_L = 0$. Given that $c \in (0,x)$, $\delta_L = 0$, and $\delta_H = 1$, the sender's expected payoff from falsifying upon observing $s = L$ is $x(\delta_H - \delta_L) - c = x - c > 0$, and thus the sender's s.r. strategy satisfies $\sigma_L = 1$. Thus, for $p \in \left(\frac{1-(1-x)q}{1+x}, q\right)$, the unique equilibrium strategies satisfy $\sigma_L^* = 1$, $\delta_L^* = 0$, and $\delta_H^* = 1$.

4. Consider parameter values such that $\beta(\widetilde{m} = L) > \frac{1}{2}$, which is equivalent to $p \in (q,1)$. For any $\sigma_L \in [0,1]$ and $\sigma_H = 0$, given that $\beta(\widetilde{m} = L) > \frac{1}{2}$, the DM's s.r. strategy satisfies $\delta_L = 1$ and $\delta_H = 1$. Given that $c > 0$, $\delta_L = 1$, and $\delta_H = 1$, the sender's expected payoff from falsifying upon observing $s = L$ is $x(\delta_H - \delta_L) - c = -c < 0$, so the sender's unique s.r. action is $\sigma_L = 0$. Thus, for $p \in (q,1)$, the unique equilibrium strategies satisfy $\sigma_L^* = 0$, $\delta_L^* = 1$, and $\delta_H^* = 1$.

Proof of Proposition 3.8

We consider here $p \in (1-q, q)$, i.e. the values of prior belief such that $\sigma_L^* > 0$ in the baseline model. First, we show that full deterrence can occur in the following two cases:

1. Consider parameter values such that

(a) $\beta\left(s_{DM}=H,\,\widetilde{m}=H\mid\sigma_L=0\right)>\frac{1}{2}$,

(b) $\beta\left(s_{DM}=L,\,\widetilde{m}=H\mid\sigma_L=0\right)<\frac{1}{2}$,

(c) $\beta\left(s_{DM}=H,\,\widetilde{m}=L\right)<\frac{1}{2}$, and

(d) $\beta\left(s_{DM}=L,\,\widetilde{m}=L\right)<\frac{1}{2}$.

Note that condition (b) and the assumption $q>q_{DM}$ imply that $p<\frac{1}{2}$. Given that $p<\frac{1}{2}$, $q>q_{DM}$ implies condition (c). Finally, condition (d) is implied by condition (b). Thus, conditions (c) and (d) are redundant. We need to analyse two subcases here:

- Consider $\beta\left(s_{DM}=H,\,\widetilde{m}=H\mid\sigma_L=1\right)>\frac{1}{2}$. Suppose that $\sigma_L=0$. The DM's s.r. strategy is $(\delta_{LL},\delta_{HL},\delta_{LH},\delta_{HH})=(0,0,0,1)$ regardless of σ_L, and hence the sender's expected benefit from falsification upon observing $s=L$ is $\pi_{H|L}x-c$. If $c<\pi_{H|L}x$, the sender's sequentially rational ("s.r.") strategy is $\sigma_L=1$, a contradiction. However, otherwise, the sender's s.r. strategy is $\sigma_L=0$ (full deterrence). In the knife-edge case of $c=\pi_{H|L}x$, we assume that the sender does not falsify, i.e. $\sigma_L=0$.
- Consider now $\beta\left(s_{DM}=H,\,\widetilde{m}=H\mid\sigma_L=1\right)<\frac{1}{2}$. Suppose that $\sigma_L=0$. Given condition (a), the DM's s.r. strategy then satisfies $\delta_{HH}=1$, so the sender's expected benefit from falsification upon observing $s=L$ is $\pi_{H|L}x-c$. If $c<\pi_{H|L}x$, the sender's s.r. strategy is $\sigma_L>0$, a contradiction. If $c>\pi_{H|L}x$, the sender's s.r. strategy is $\sigma_L=0$, and therefore the equilibrium strategies of the DM and the sender satisfy $\sigma_L=0$ (full deterrence) and $\delta_{HH}=1$. In the knife-edge case of $c=\pi_{H|L}x$, we assume that the sender does not falsify, i.e. $\sigma_L=0$.

Finally, we need to show that there exist values of p which satisfy (a) and (b)

as well as $p > 1 - q$, which is a necessary condition for $\sigma_L^* > 0$ when the DM does not acquire a private signal. Note here that $p > 1 - q$ is equivalent to $\beta(\widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$. Furthermore, $\beta(\widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$ implies $\beta(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$, and $\beta(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$ implies $\beta(\widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$. Hence, there exists a range of p such that (a) and (b) are satisfied as well as $p > 1 - q$.

2. Consider parameter values such that

   (a) $\beta(s_{DM} = H, \widetilde{m} = H \mid \sigma_L) > \frac{1}{2}$,

   (b) $\beta(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$,

   (c) $\beta(s_{DM} = H, \widetilde{m} = L) > \frac{1}{2}$, and

   (d) $\beta(s_{DM} = L, \widetilde{m} = L) < \frac{1}{2}$.

   Note that condition (c) implies that condition (a) holds for any σ_L. Furthermore, condition (c) and the assumption $q > q_{DM}$ imply that $p > \frac{1}{2}$. Finally, given that $p > \frac{1}{2}$, $q > q_{DM}$ implies condition (b). Thus, conditions (a) and (b) are redundant here. We need to analyse two subcases here:

   - Consider $\beta(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 1) > \frac{1}{2}$. Suppose that $\sigma_L = 0$. The DM's s.r. strategy is then $(\delta_{LL}, \delta_{HL}, \delta_{LH}, \delta_{HH}) = (0, 1, 1, 1)$ regardless of σ_L, and hence the sender's expected benefit from falsification upon observing $s = L$ is $\pi_{L|L}x - c$. If $c < \pi_{L|L}x$, the sender's s.r. strategy is $\sigma_L = 1$, a contradiction. However, otherwise, the sender's s.r. strategy is $\sigma_L = 0$ (full deterrence). In the knife-edge case of $c = \pi_{L|L}x$, we assume that the sender does not falsify.
   - Consider $\beta(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 1) < \frac{1}{2}$. The analysis proceeds analogously to that in the second bullet point of (1).

Finally, we need to show that there exist values of p which satisfy (c) and (d) as well as $p < q$, which is a necessary condition for $\sigma_L^* > 0$ when the DM does not acquire a private signal. Note here that $p < q$ is equivalent to $\beta(\widetilde{m} = L) < \frac{1}{2}$. Furthermore, $\beta(\widetilde{m} = L) < \frac{1}{2}$ implies $\beta(s_{DM} = L, \widetilde{m} = L) < \frac{1}{2}$, and $\beta(s_{DM} = H, \widetilde{m} = L) < \frac{1}{2}$ implies $\beta(\widetilde{m} = L) < \frac{1}{2}$. Hence, there exists a range of p such that (c) or (d) are satisfied as well as $p < q$.

In addition to this, we show that there are no other cases where full deterrence can occur.

First, $\beta(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 0) < \frac{1}{2}$ contradicts $p > 1-q$, as the latter is equivalent to $\beta(\widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$ and therefore implies $\beta(s_{DM} = H, \widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$. Similarly, the condition $\beta(s_{DM} = L, \widetilde{m} = L) > \frac{1}{2}$ contradicts the condition $p < q$, as the latter is equivalent to $\beta(\widetilde{m} = L) < \frac{1}{2}$ and thus implies $\beta(s_{DM} = L, \widetilde{m} = L) < \frac{1}{2}$.

Second, consider parameter values which satisfy the following conditions:

(a) $\beta(s_{DM} = H, \widetilde{m} = H \mid \sigma_L^*) > \frac{1}{2}$,

(b) $\beta(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 0) > \frac{1}{2}$ and $\beta(s_{DM} = L, \widetilde{m} = H \mid \sigma_L = 1) < \frac{1}{2}$,

(c) $\beta(s_{DM} = H, \widetilde{m} = L) < \frac{1}{2}$, and

(d) $\beta(s_{DM} = L, \widetilde{m} = L) < \frac{1}{2}$.

Suppose that $\sigma_L^* = 0$. Then the DM's s.r. strategy satisfies $\delta_{LH} = 1$. The sender's expected payoff from falsifying upon seeing $s = L$ is $\pi_{H|L}x + \pi_{L|L}x - c = x - c > 0$, and so his s.r. strategy is $\sigma_L = 1$, a contradiction. Instead, there is a mixed-strategy equilibrium with $\delta_{LH}^* = \frac{c - \pi_{H|L}x}{\pi_{L|L}x}$ and σ_L^* such that $\beta(s_{DM} = H, \widetilde{m} = H \mid \sigma_L^*) = \frac{1}{2}$.

Finally, note that parameter values which satisfy $\beta(s_{DM} = L, \widetilde{m} = L) = \frac{1}{2}$ or $\beta(s_{DM} = H, \widetilde{m} = L) = \frac{1}{2}$ correspond to knife-edge cases.

Proof of Proposition 3.9

The proof follows from the Proof of Proposition 3.8. Consider parameter values such that conditions (a)-(d) in point 2 in the Proof of Proposition 3.8 are satisfied. Again, conditions (a) and (b) are redundant here. Let us consider the equilibrium for these parameter values:

- Consider $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L=1\right)>\frac{1}{2}$. The DM's equilibrium strategy is then $(\delta^*_{LL},\delta^*_{HL},\delta^*_{LH},\delta^*_{HH})=(0,1,1,1)$ regardless of σ^*_L, and hence the sender's expected benefit from falsification upon observing $s=L$ is $\pi_{L|L}x-c$. If $c\leq\pi_{L|L}x$, the sender's equilibrium strategy is $\sigma^*_L=1$. Otherwise, the sender's equilibrium strategy is $\sigma^*_L=0$. In the knife-edge case of $c=\pi_{L|L}x$, we assume that the sender does not falsify. Since $\pi_{L|L}$ is increasing in q_{DM}, an increase in q_{DM} can result in the condition $c\leq\pi_{L|L}x$ not being satisfied and hence can lead to an increase in falsification from $\sigma^*_L=0$ to $\sigma^*_L=1$. By means of an example one can show that it can lead to a (discontinuous) decrease in the DM's ex ante expected payoff.

- Consider $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma_L=1\right)<\frac{1}{2}$. Suppose that $\sigma_L=0$. Then the DM's s.r. strategy would satisfy $\delta_{LH}=1$. The sender's expected benefit from falsifying upon seeing $s=L$ is then $\pi_{L|L}x-c$. If $\pi_{L|L}x-c\leq 0$, then the sender has no incentive to falsify, and the conjectured pair of strategies is an equilibrium (i.e. $\sigma^*_L=0$ and $\delta^*_{LH}=1$). On the other hand, if $\pi_{L|L}x-c>0$, then the sender has an incentive to falsify with probability 1, which means that the conjectured pair of strategies is not an equilibrium. Instead, there is a mixed-strategy equilibrium with $\delta^*_{LH}=\frac{c}{\pi_{L|L}x}$ and σ^*_L such that $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma^*_L\right)=\frac{1}{2}$. Since $\pi_{L|L}$ is increasing in q_{DM}, an increase in q_{DM} can lead to an increase in falsification from $\sigma^*_L=0$ to σ^*_L such that $\beta\left(s_{DM}=L,\widetilde{m}=H\mid\sigma^*_L\right)=\frac{1}{2}$, and by means of an example one can show that it can lead to a (discontinuous) decrease in the DM's

ex ante expected payoff.

For parameter values other than those satisfying conditions (a)-(d) in point 2 in the Proof of Proposition 3.8, the above effect is absent and hence an increase in q_{DM} can never lead to an increase in σ_L^*, and thus there cannot be a decrease in the DM's ex ante expected payoff. In particular, note that (i) in point 1 of the Proof of Proposition 3.8 the sender's equilibrium strategy is $\sigma_L = 0$ if and only if $c \geq \pi_{H|L}x$, and (ii) $\pi_{H|L}$ is decreasing in q_{DM}, which means that an increase in q_{DM} makes it more likely that $c \geq \pi_{H|L}x$ is satisfied.

Proof of Lemma 3.1

1. Suppose that the DM sets $\tau = \tau^{'}$ where $\tau^{'} \leq \frac{c}{x}$, and solve for the equilibrium strategies of the DM and the sender. Consider any $\delta_H \in [0,1]$. When $\tau^{'} < \frac{c}{x}$, the sender's expected payoff from falsifying upon observing $s = L$ is $\tau x \delta_H - c < 0$, so the sender's s.r. strategy satisfies $\sigma_L = 0$. In the knife-edge case of $\tau^{'} = \frac{c}{x}$, we assume that whenever the sender is indifferent between falsifying and not falsifying, he chooses not to falsify. Given that the sender's equilibrium strategy satisfies $\sigma_L^* = 0$, the DM's welfare is $(1-\tau^{'})\mathbb{E}\left[v\left(\sigma^*, \delta^* = (0,0)\right)\right] + \tau^{'}\mathbb{E}\left[v\left(\sigma^* = (0,0), \delta^*\right)\right]$ for $p \leq \frac{1}{2}$ and $(1-\tau^{'})\mathbb{E}\left[v\left(\sigma^*, \delta^* = (1,1)\right)\right] + \tau^{'}\mathbb{E}\left[v\left(\sigma^* = (0,0), \delta^*\right)\right]$ for $p > \frac{1}{2}$. The DM's welfare is increasing in $\tau^{'}$. When $\tau^{'} = \frac{c}{x}$, the DM's welfare is

$$\left(1 - \frac{c}{x}\right)\mathbb{E}\left[v\left(\sigma^*, \delta^* = (0,0)\right)\right] + \frac{c}{x}\mathbb{E}\left[v\left(\sigma^* = (0,0), \delta^*\right)\right] \tag{3.27}$$

for $p \leq \frac{1}{2}$, and

$$\left(1 - \frac{c}{x}\right)\mathbb{E}\left[v\left(\sigma^*, \delta^* = (1,1)\right)\right] + \frac{c}{x}\mathbb{E}\left[v\left(\sigma^* = (0,0), \delta^*\right)\right] \tag{3.28}$$

for $p > \frac{1}{2}$.

2. Suppose that the DM sets $\tau = \tau^{'}$ where $\tau^{'} > \frac{c}{x}$, and solve for the equilibrium strategies of the DM and the sender.

   (a) Consider parameter values such that $\beta\left(\widetilde{m} = H \mid \sigma_L = 1\right) \geq \frac{1}{2}$, which is equivalent to $p \geq \frac{x}{1+x}$. Then, for any $\sigma_L \in [0,1]$, the DM's s.r ("sequentially rational") strategy satisfies $\delta_H = 1$. The sender's expected payoff from falsifying upon observing $s = L$ is $\tau x - c > 0$, so the sender's s.r. strategy satisfies $\sigma_L = 1$. Thus, for $p \geq \frac{x}{1+x}$, the sender's equilibrium strategy is $\sigma_L^* = 1$. The DM's welfare for $p \leq \frac{1}{2}$ is then $(1 - \tau^{'})\mathbb{E}\left[v\left(\sigma^*, \delta^* = (0,0)\right)\right] + \tau^{'}\mathbb{E}\left[v\left(\sigma^* = (1,0), \delta^* = (0,1)\right)\right]$, and for $p > \frac{1}{2}$ it is $(1 - \tau^{'})\mathbb{E}\left[v\left(\sigma^*, \delta^* = (1,1)\right)\right] + \tau^{'}\mathbb{E}\left[v\left(\sigma^* = (1,0), \delta^* = (0,1)\right)\right]$. The DM's welfare is increasing in $\tau^{'}$. When $\tau^{'} = 1$, there is a unique equilibrium and the DM's welfare is $\mathbb{E}\left[v\left(\sigma^* = (1,0), \delta^* = (0,1)\right)\right] = 1 - (1-p)\,x$ for all $p \in \left(\frac{x}{1+x}, 1\right)$.

   (b) Consider parameter values such that $\beta\left(\widetilde{m} = H \mid \sigma_L = 1\right) < \frac{1}{2}$, which is equivalent to $p < \frac{x}{1+x}$.

      i. If $\sigma_L = 1$, then the DM's s.r strategy satisfies $\delta_H = 0$, so the sender's expected payoff from falsifying upon observing $s = L$ is $\tau^{'} x \delta_H - c = -c < 0$, which makes $\sigma_L = 1$ not s.r., a contradiction.

      ii. If $\sigma_L = 0$, then the DM's s.r strategy satisfies $\delta_H = 1$, so the sender's expected payoff from falsifying upon observing $s = L$ is $\tau^{'} x \delta_H - c = \tau^{'} x - c > 0$, which makes $\sigma_L = 0$ not s.r., a contradiction.

      iii. Suppose now that $\sigma_L \in (0,1)$. If $\beta\left(\widetilde{m} = H \mid \sigma_L\right) > \frac{1}{2}$ or $\beta\left(\widetilde{m} = H \mid \sigma_L\right) < \frac{1}{2}$, then $\sigma_L \in (0,1)$ is not s.r. If $\beta\left(\widetilde{m} = H \mid \sigma_L\right) = \frac{1}{2}$, then any $\delta_H \in [0,1]$ is s.r for the DM. If either $\tau^{'} x \delta_H - c > 0$ or $\tau^{'} x \delta_H - c < 0$, then $\sigma_L \in (0,1)$ is not s.r. If $\tau^{'} x \delta_H - c = 0$, then $\sigma_L \in (0,1)$ is s.r. for the

sender. This uniquely pins down $\delta_H = \frac{c}{\tau' x}$. Also, $\beta(\widetilde{m} = H \mid \sigma_L) = \frac{1}{2}$ uniquely pins down $\sigma_L = \frac{p}{(1-p)x}$. Thus, for $p < \frac{x}{1+x}$, the equilibrium strategies satisfy $\sigma_L^* = \frac{p}{(1-p)x}$ and $\delta_H^* = \frac{c}{\tau' x}$. For $p \leq \frac{1}{2}$, the DM's welfare is $(1-\tau')\mathbb{E}[v(\sigma^*, \delta^* = (0,0))] + \tau'\mathbb{E}[v(\sigma^*, \delta^* = (0,0))] = \mathbb{E}[v(\sigma^*, \delta^* = (0,0))]$, which is smaller than (3.27).

Proof of Proposition 3.10

Using the results from Table 3.3:

1. Consider $0 < p < \frac{x}{1+x}$. Then $\mathbb{E}\left(v_{\tau=c/x}(\sigma^*, \delta^*)\right) = \frac{c}{x} + (1-p)\left(1 - \frac{c}{x}\right)$ and $\mathbb{E}(v_{\tau=1}(\sigma^*, \delta^*)) = 1-p$. It can be easily verified that $\frac{c}{x} + (1-p)\left(1 - \frac{c}{x}\right) > 1-p$ for all parameter values.

2. Consider $\frac{x}{1+x} < p < \frac{1}{2}$. Then $\mathbb{E}\left(v_{\tau=c/x}(\sigma^*, \delta^*)\right) = \frac{c}{x} + (1-p)\left(1 - \frac{c}{x}\right)$ and $\mathbb{E}(v_{\tau=1}(\sigma^*, \delta^*)) = p + (1-p)(1-x)$. The condition $\frac{c}{x} + (1-p)\left(1 - \frac{c}{x}\right) \geq p + (1-p)(1-x)$ can be rearranged to $c \geq x\left(1 - \frac{1-p}{p}x\right)$, where the right-hand side of the inequality is increasing in p.

3. Consider $\frac{1}{2} < p$. Then $\mathbb{E}\left(v_{\tau=c/x}(\sigma^*, \delta^*)\right) = \frac{c}{x} + p\left(1 - \frac{c}{x}\right)$ and $\mathbb{E}(v_{\tau=1}(\sigma^*, \delta^*)) = p + (1-p)(1-x)$. The condition $\frac{c}{x} + p\left(1 - \frac{c}{x}\right) \geq p + (1-p)(1-x)$ can be rearranged to $c \geq x(1-x)$.

Note that $x\left(1 - \frac{1-p}{p}x\right) \leq x(1-x)$ if and only if $p \leq \frac{1}{2}$, and that $x\left(1 - \frac{1-p}{p}x\right) \leq 0$ if and only if $p \leq \frac{x}{1+x}$. This yields the condition in (3.12) and can be used to construct all the cases in Table 3.4.

Proof of Proposition 3.11

The proof follows from the main text.

References

[1] Alchian, Armen A., and Harold Demsetz (1972): "Production, Information Costs, and Economic Organization," *American Economic Review*, 62 (5): 777–795.

[2] Ananyev, Maxim, Maria Petrova, and Galina Zudenkova (2017): "Information and Communication Technologies, Protests, and Censorship," Working Paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2978549

[3] Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan (2006): "Signaling in a Global Game: Coordination and Policy Traps," *Journal of Political Economy*, 114(3): 452-484.

[4] ——— (2007): "Dynamic Global Games of Regime Change: Learning, Multiplicity, and the Timing of Attacks," *Econometrica*, 75 (3): 711-756.

[5] Baker, George P. (1992): "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100 (3): 598-614.

[6] Balbuzanov, Ivan (2017): "Lies and Consequences: The Effect of Lie Detection on Communication Outcomes," Working Paper; https://sites.google.com/site/ibalbuzanov/research

[7] Bannier, Christina E., and Frank Heinemann (2005): "Optimal Transparency and Risk-Taking to Avoid Currency Crises," *Journal of Institutional and Theoretical Economics*, 161(3): 374-391.

[8] Bénabou, Roland, and Jean Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70 (3): 489-520.

[9] ——— (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96 (5): 1652-1678.

[10] Besley, Tim, and Andrea Prat (2006): "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability," *American Economic Review*, 96 (3): 720-736.

[11] Bester, Helmut, and Daniel Krähmer (2008): "Delegation and Incentives," *RAND Journal of Economics*, 39 (3): 664-682.

[12] Boix, Carles, and Milan W. Svolik (2013): "The Foundations of Limited Authoritarian Government: Institutions and Power Sharing in Dictatorships," *Journal of Politics*, 75 (2): 300–316.

[13] Bramoullé, Yann, and Caroline Orset (2017): "Manufacturing Doubt," Working Paper; https://sites.google.com/site/bramoulley/

[14] Bueno De Mesquita, Ethan (2010): "Regime Change and Revolutionary Entrepreneurs," *American Political Science Review*, 104 (3): 446–466.

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

[16] Carrell, Scott E., and James E. West (2010): "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 118 (3): 409-432.

[17] Casper, Brett A., and Scott A. Tyson (2014): "Popular Protest and Elite Coordination in a Coup d'état" *Journal of Politics*, 76 (2): 548-564.

[18] Chassang, Sylvain, and Gerard Padró-i-Miquel (2010): "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics*, 125 (4): 1821-1858.

[19] Chen, Huan, and Wing Suen (2017): "Competition for Attention in the News Media Market," Working Paper: http://www.sef.hku.hk/∼wsuen/research/wppdf/media-8.pdf

[20] Colombo, Massimo G., and Marco Delmastro (2004): "Delegation of Authority in Business Organization: An Empirical Test," *Journal of Industrial Economics*, 52 (1): 53–80.

[21] Corsetti, Giancarlo, Amil Dasgupta, Stephen Morris, and Hyun S. Shin (2004): "Does One Soros Make a Difference? A Theory of Currency Crises with Small and Large Traders," *Review of Economic Studies*, 71 (1): 87-113.

[22] Crawford, Vince, and Joel Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50 (6), 1431-1451.

[23] Dellarocas, Chrysanthos (2006): "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science*, 52 (10): 1577-1593.

[24] De Varo, Jed, and Fidan A. Kurtulus (2010): "An Empirical Analysis of Risk, Incentives and the Delegation of Worker Authority," *Industrial and Labor Relations Review*, 63 (4): 641–661.

[25] De Varo, Jed, and Suraj Prasad (2015): "The Relationship between Delegation and Incentives Across Occupations: Evidence and Theory," *Journal of Industrial Economics*, 63 (2): 279-312.

[26] Dziuda, Wioletta, and Christian Salas (2017): "Communication with Detectable Deceit," Working Paper; https://sites.google.com/site/dziudawiola/research

[27] Ederer, Florian, Richard Holden, and Margaret Meyer (2017): "Gaming and Strategic Opacity in Incentive Provision," Working Paper; http://www.nuff.ox.ac.uk/Users/Meyer/EHM06July2017.pdf

[28] Edmond, Chris (2013): "Information Manipulation, Coordination, and Regime Change," *Review of Economic Studies*, 80 (4): 1422-1458.

[29] Egorov, Georgy, Sergei Guriev, and Konstantin Sonin (2009): "Why Resource-poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data," *American Political Science Review*, 103 (4): 645-668.

[30] Enikopolov, Ruben, Alexey Makarin, and Maria Petrova (2016): "Social Media and Protest Participation: Evidence from Russia," Working Paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2696236

[31] Enikopolov, Ruben, Maria Petrova, and Konstantin Sonin (2017): "Social Media and Corruption," forthcoming in *American Economic Journal: Applied Economics*.

[32] Enikopolov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya (2011): "Media and Political Persuasion: Evidence from Russia," *American Economic Review,* 101 (7): 3253-3285.

[33] Fields, Thomas D., Thomas Z. Lys, and Linda Vincent (2001): "Empirical Research on Accounting Choice," *Journal of Accounting and Economics*, 31 (1-3): 255–307.

[34] Foss, Nicolai J., and Keld Laursen (2005): "Performance Pay, Delegation and Multitasking under Uncertainty and Innovativeness: An Empirical Investigation," *Journal of Economic Behavior and Organization*, 58 (2): 246–276.

[35] Gehlbach, Scott, and Konstantin Sonin (2014): "Government Control of the Media," *Journal of Public Economics*, 118: 163-171.

[36] Giebe, Thomas, and Oliver Gürtler (2012): "Optimal Contracts for Lenient Supervisors," *Journal of Economic Behavior and Organization*, 81 (2): 403-420.

[37] Grossman, Sanford J. (1981): "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics*, 24 (3): 461-483.

[38] Guriev, Sergei, and Daniel Treisman (2015): "How Modern Dictators Survive: Cooptation, Censorship, Propaganda, and Repression," CEPR Discussion Paper No. 10454.

[39] Han, Rongbin (2015): "Manufacturing Consent in Cyberspace: China's 'Fifty-Cent Army'," *Journal of Current Chinese Affairs*, 44 (2): 105-134.

[40] Hellwig, Christian (2002): "Public Information, Private Information, and the Multiplicity of Equilibria in Coordination Games," *Journal of Economic Theory*, 107 (2): 191-222.

[41] Holmström, Bengt, and Paul Milgrom (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7: 24-52.

[42] Hong, Bryan, Lorenz Kueng, and Mu-Jeung Yang (2016): "Complementarity of Performance Pay and Task Allocation," Working Paper; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2746478

[43] Iachan, Felipe S., and Plamen T. Nenov (2015): "Information Quality and Crises in Regime-Change Games," *Journal of Economic Theory*, 158: 739-768.

[44] Jia, Yuping, and Paula M. G. van Veen-Dirks (2015): "Control System Design at the Production Level," Working Paper; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2426148

[45] Kamenica, Emir, and Matthew Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101 (6): 2590-2615.

[46] Kartik, Navin (2009): "Strategic Communication with Lying Costs," *Review of Economic Studies*, 76 (4): 1359-1395.

[47] Kartik, Navin, Marco Ottaviani, and Francesco Squintani (2007): "Credulity, Lies, and Costly Talk," *Journal of Economic Theory*, 134 (1): 93-116.

[48] Kerr, Steven (1975): "On the Folly of Rewarding A, While Hoping for B," *Academy of Management Journal*, 18 (4): 769-783.

[49] King, Gary, Jennifer Pan, and Margaret E. Roberts (2013): "How Censorship in China Allows Government Criticism but Silences Collective Expression," *American Political Science Review*, 107 (2): 1-18.

[50] King, Gary, Jennifer Pan, and Margaret E. Roberts (2014): "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation," *Science*, 345 (6199): 1-10.

[51] King, Gary, Jennifer Pan, and Margaret E. Roberts (2017): "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument," *American Political Science Review*, 111 (3): 484-501.

[52] Lando, Henrik (2004): "Are Delegation and Incentives Complementary Instruments?", Working Paper; http://openarchive.cbs.dk/bitstream/handle/10398/6813/lefic%202004-02.pdf?sequence=1

[53] Letina, Igor, Shuo Liu, and Nick Netzer (2016): "Delegating Performance Evaluation," Working Paper; http://www.econ.uzh.ch/dam/jcr:46691a36-c2ad-443c-8a12-3d2caab0b8d3/DelegatingEvaluation.pdf

[54] Little, Andrew (2016): "Communication Technology and Protest," *Journal of Politics*, 78 (1): 152-166.

[55] Lo, Desmond, Wouter Dessein, Mrinal Ghosh, and Francine Lafontaine (2016): "Price Delegation and Performance Pay: Evidence from Industrial Sales Forces," *Journal of Law, Economics, and Organizatio*n, 32 (3): 508-544.

[56] Loeper, Antoine, Jakub Steiner, and Colin Stewart (2014): "Influential Opinion Leaders," *Economic Journal*, 124 (581): 1147-1167.

[57] Lorentzen, Peter (2014): "China's Strategic Censorship," *American Journal of Political Science*, 58 (2): 402-414.

[58] MacLeod, W. Bentley, and Daniel Parent (1999): "Job Characteristics and the Form of Compensation," in Polacheck S. (ed.), Research in Labor Economics, JAI: Stamford, Connecticut.

[59] Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014): "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8): 2421-55.

[60] Metz, Christina E. (2002): "Private and Public Information in Self-Fulfilling Currency Crises," *Journal of Economics*, 76 (1): 65-85.

[61] Milgrom, Paul R., and John Roberts (1992): *Economics, Organization, and Management*, Pearson.

[62] Milgrom, Paul R. (1981): "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 12 (2): 380-391.

[63] Mondria, Jordi, and Climent Quintana-Domeque (2013): "Financial Contagion and Attention Allocation," *Economic Journal*, 123 (568): 429-454.

[64] Morris, Stephen, and Hyun S. Shin (1998): "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *American Economic Review*, 88 (3): 587–597.

[65] ——— (2001): "Rethinking Multiple Equilibria in Macroeconomics," in NBER Macroeconomics Annual 2000, ed. by Ben S. Bernanke and Kenneth S. Rogoff, Cambridge, MA: MIT Press, pp. 139-161.

[66] ——— (2002): "Social Value of Public Information," *American Economic Review*, 92 (5): 1521-1534.

[67] ——— (2003): "Global Games—Theory and Applications," in "Advances in Economics and Econometrics," 8th World Congress of the Econometric Society, ed. by Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Cambridge, U.K.: Cambridge University Press, pp. 56-114.

[68] Myatt, David, and Chris Wallace (2012): "Endogenous Information Acquisition in Coordination Games," *Review of Economic Studies*, 79 (1): 340-374.

[69] Nagar, Venky (2002): "Delegation and Incentive Compensation," *The Accounting Review*, 77 (2): 379–395.

[70] Oreskes, Naomi, and Erik M. Conway (2010): *Merchants of Doubt. How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*, New York: Bloosmbury Press.

[71] Petrova, Maria (2008): "Inequality and Media Capture," *Journal of Public Economics*, 92: 183-212.

[72] Prendergast, Canice (2002): "The Tenuous Trade-Off between Risk and Incentives," *Journal of Political Economy*, 110 (5): 1071-1102.

[73] Prendergast, Canice, and Robert H. Topel (1996): "Favoritism in Organizations," *Journal of Political Economy*, 104 (5): 958-978.

[74] Proctor, Robert N. (2012): *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition*, University of California Press.

[75] Rahman, David (2012): "But Who Will Monitor the Monitor?", *American Economic Review*, 102 (6): 2767-2797.

[76] Rohlfing-Bastian, Anna, and Anja Schöttner (2017): "Incentives and the Delegation of Task Assignment," Working Paper; https://www.wiwi.hu-berlin.de/de/professuren/vwl/microeconomics/people/aschoettner

[77] Rose, Colin, and Murray D. Smith (1996): "The Multivariate Normal Distribution," *Mathematica Journal*, 6: 32-37.

[78] Rose, Colin, and Murray D. Smith (2002): "The Bivariate Normal," §6.4 A in "Mathematical Statistics with Mathematica," New York: Springer-Verlag, pp. 216-226.

[79] Segal, Ilya (1999): "Contracting with Externalities," *Quarterly Journal of Economics*, 114 (2): 337-388.

[80] Shadmehr, Mehdi, and Dan Bernhardt (2011): "Collective Action with Uncertain Payoffs: Coordination, Public Signals and Punishment Dilemmas," *American Political Science Review*, 105 (4): 829–851.

[81] ——— (2015): "State Censorship," *American Economic Journal: Microeconomics*, 7 (2): 280-307.

[82] Shapiro, Jesse M. (2016): "Special Interests and the Media: Theory and an Application to Climate Change," *Journal of Public Economics*, 144: 91-108.

[83] Sibuya, Masaaki (1960): "Bivariate Extreme Statistics, I," *Annals of the Institute of Statistical Mathematics*, 11 (2): 195-210.

[84] Stone, Daniel F. (2011): "A Signal-Jamming Model of Persuasion: Interest Group Funded Policy Research," *Social Choice and Welfare*, 37 (3): 397-424.

[85] Stuart, Alan, and Keith Ord (1998): "Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory," 6th ed. New York: Oxford University Press.

[86] Sungur, Engin A. (1990): "Dependence Information in Parameterized Copulas," *Communications in Statistics - Simulation and Computation*, 19 (4): 1339-1360.

[87] Szkup, Michal, and Isabel Trevino (2015): "Information Acquisition in Global Games of Regime Change," *Journal of Economic Theory*, 160: 387-428.

[88] Watts, Ross L., and Jerold L. Zimmerman (1986): *Positive Accounting Theory*, Englewood Cliffs, NJ: Prentice-Hall.

[89] Wulf, Julie (2007): "Authority, Risk, and Performance Incentives: Evidence from Division Manager Positions inside Firms," *Journal of Industrial Economics*, 55 (1): 169–196.