Four Essays on Strategic Communication

Uliana Loginova

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

Abstract

Four Essays on Strategic Communication

Uliana Loginova

This dissertation studies patterns of strategic communication in cases, in which the involved parties disagree in their preferences or opinions.

In Chapter 1, I study a model of strategic communication in networks, in which the players diverge in their preferences and information can be communicated either through a costly *verifiable information (hard)* channel or through a low-cost *cheap talk (soft)* channel. I find that the availability of hard links allows each agent to get a weakly greater number of truthful messages compared to the pure cheap talk setting. If only one party bears the cost of a hard link, then introducing hard links increases the total expected welfare. In contrast, if the cost of a hard link is shared by both parties, then allowing for verifiable communication can decrease the total welfare.

In Chapter 2, I consider a model of strategic cheap talk communication in networks, in which the players can disagree in their preferences or their opinions. I find that the information transmission pattern crucially depends on the nature of the disagreement. If the agents diverge in their preferences, then information transmission exhibits a *negative externality effect*: greater information obtained by some agent discourages further information accumulation by harming the credibility of other agents. In contrast, information transmission displays a *positive externality effect* when the agents have divergent opinions: greater information obtained by some agent encourages further information accumulation by improving the credibility of other agents.

Chapter 3 studies a benevolent authority's decision to constrain or inform a population of individuals. It demonstrates that the authority's decision to regulate an activity depends on whether she deems it a matter of preference or opinion. In the former case, the benevolent authority is *libertarian*: she gives truthful advice and safeguards liberty. In the latter case, the benevolent authority is *paternalistic*: believing that she acts in the individuals' best interest, the authority forces another action than the individuals would choose for themselves.

In Chapter 4, I consider communication between an informed Sender and an uninformed Re-

ceiver. The Sender has a preference bias and is *guilt averse* to letting down the Receiver's payoff expectations. I show that no separating equilibrium exists; rather, in case of uniform state of the world and quadratic utilities, I demonstrate that there exist partition equilibria (as in Crawford and Sobel (1982)). An increase in the guilt aversion intensity is akin a decrease in the preference divergence: higher guilt aversion intensity allows for more intervals in the equilibrium partition; and holding the number of elements in the partition fixed, greater guilt aversion intensity results in more balanced intervals.

Table of Contents

List of Figures

Acknowledgments

I am truly grateful to my advisor Navin Kartik for his generous support, invaluable guidance, and immense help at all stages of my research.

I am especially indebted to Alessandra Casella, Yeon-Koo Che, Wouter Dessein, Rajiv Sethi, and Timothy Van Zandt for their creative suggestions, precious comments, and encouragement. I thank Petra Persson for her bright ideas and priceless efforts that made the work in Chapter 3 possible.

I also thank Patrick Bolton, Andrea Galeotti, Samuel Lee, Qingmin Liu, Andrea Prat, Joel Sobel, Giancarlo Spagnolo, Francesco Squintani, Steve Tadelis, and Curtis Taylor for productive discussions and inspirational ideas. I am grateful to Anton Kolotilin, Youngwoo Koh, Sébastien Turban, and participants of the Columbia University theory colloquium for their valuable comments.

Finally, I thank my parents Elena Loginova and Sergey Loginov, my sister Oksana Loginova, and all my friends for their love, encouragement, and sense of humor that helped me a lot throughout this whole process. And of course, I thank my husband Dmitriy Sergeyev, to whom this thesis is dedicated, for his endless support, inspiration, and his confidence in me.

Dissertation Committee

Alessandra Casella

Columbia University, Department of Economics

Yeon-Koo Che

Columbia University, Department of Economics

Wouter Dessein

Columbia University, Graduate School of Business

Navin Kartik

Columbia University, Department of Economics

Rajiv Sethi

Columbia University, Barnard College, Department of Economics

To Dmitriy, who makes me truly enjoy my life

Summary

In many economic situations, people need to make judgements about the uncertain environment and choose the most appropriate actions that influence other agents as well. Oftentimes, these decisions are based not only on the private information of the individuals, but also on the information collected from other agents during preliminary discussions. When people differ in their opinions or in their preferences, these features give rise to *strategic communication*. The goal of this dissertation is to characterize the patterns of information transmission.

The basic framework of Chapter 1 and Chapter 2 is the following. There are n players, each player i receives a private signal s_i about the uncertain state of the world $\theta \in [0,1]$. Conditional on θ, the signals are independent and identically distributed, with each s_i taking a value of 1 with probability θ and 0 with complementary probability $1-\theta$. Agents can simultaneously transmit their private signals to each other according to the communication network. Finally, players pick actions that influence all other agents.

Chapter 1. In Chapter 1, I study the communication in networks, where individuals can choose between soft and hard information transmission. Soft links are very cheap to sustain, and correspond to cheap talk communication (e.g., writing an email with a short summary). Hard links are costly, and correspond to non-strategic direct revelation of the signal (e.g., personal meeting where the receiving party is presented with all supporting evidence that allows to decode the underlying signal). Individuals agree in their beliefs about the uncertain environment, but differ in their preferences, so that player i's payoff is $-\sum_{j=1}^{n}(y_j-\theta-b_i)^2$. That is, player i with the preference bias b_i prefers that his action, as well as the actions of other players, are close to player i's ideal action, $\theta+b_i$.

While both parties always desire the truthful revelation of a signal from the ex-ante perspective, there is a credibility issue in strategic communication via a soft link at the interim stage. Similar to [Galeotti *et al.*, 2011], I find a *congestion effect* in cheap talk communication. Namely, player i is credible in reporting to player j as long as the number of truthful messages player j gets (called the *in-degree* of player j and denoted by k_j) is not that large compared to the preference divergence, $|b_i-b_j|$.[1] At the same time, the incentive to communicate through a hard link does not depend

[1] Intuitively, the magnitude of the effect from an additional signal on j's action decreases with the informativeness

on the preference divergence: a hard link from player i to player j is created whenever the ex-ante benefits of revealing an additional signal (which depend only on player j's in-degree, k_j) outweigh the costs.

The first—intuitive—result of the chapter describes the effect of introducing the possibility of forming hard links on the scope of information transmission. More precisely, start from the pure cheap talk setting, i.e., verifiable information transmission is prohibitively costly. Then introduce feasible hard links, i.e., lower the cost of hard links such that they can emerge in the equilibrium truthful network. As a result, there is a new equilibrium that generates more intensive information transmission, i.e., weakly larger in-degrees for all players.

The second—and main—result concerns the welfare aspect of introducing the feasible verifiable information transmission channel. The welfare outcome is composed of two effects: the positive effect of an *information improvement* and the negative effect of *crowding out cheap talk communication* with costly verifiable information transmission. The crowding out effect is at the heart of the chapter. As hard links become available, some agents who could not truthfully communicate via cheap talk can find it beneficial to pay the cost and transfer the private information through hard links. The appearance of hard links increases players' in-degrees which, given the congestion result, can render truth-telling through some soft links no longer incentive compatible. Those soft links must be replaced by hard links for the players to remain credible. Thus, even though players get more truthful messages, the number of soft links might decrease, implying that the cheap talk communication is crowded out by the costly verifiable information transmission.

I find that the final welfare outcome of the two forces depends on the cost structure. If only one party bears the cost of a hard link, then there always exists an equilibrium with feasible hard links that generates greater total welfare than in the pure cheap talk case. Interestingly, the welfare can decrease if the cost of a hard link is shared between the players. Indeed, each player does not account for the full cost of a hard link, hence, too many hard links are created. As a result, the negative crowding out effect can dominate the positive informational effect.

Chapter 2. In Chapter 2, I consider patterns of cheap talk communication, where individuals disagree on the proper actions, either because they have *conflicting preferences* or because they

of agent j. Thus, for sufficiently high k_j, the expected effect might become so small that i would prefer to lie in order to shift j's action closer towards i's preferred one.

have *conflicting opinions.* The framework differs from the setting of Chapter 1 in that the state of the world relevant to the decision-making process is redefined to be the sum S of private signals s_i. Under conflicting preferences, the individuals agree on the fundamentals of the uncertain economic environment but simply like different courses of action. Under conflicting opinions, the individuals would prefer the same actions had they known the exact state of the world; however, under incomplete information, they diverge in their opinions about the uncertain environment and, thus, deem different actions optimal. As in Chapter 1, I formalize the interdependence of individuals' payoffs by assuming that each agent faces a loss when his action and the actions of other agents differ from his own ideal action.

Under either type of disagreement between the agents, the two parties, i and j, would ex-ante prefer truthful revelation of s_i to agent j. However, the credibility of communication and the way it is influenced by j's in-degree k_j crucially depends on the nature of disagreement between the individuals. When the two parties diverge in their preferences, I find a *negative externality effect* of information transmission: greater information available to player j—higher k_j—harms the credibility of player i. This corresponds to the congestion effect of Chapter 1 and shares the same intuition.

In contrast, when the parties disagree in their opinions, the communication pattern exhibits *positive externality effect*: greater number of aspects obtained by agent j—higher k_j—improves the credibility of agent i. Intuitively, as k_j increases, two things happen. First, agent i expects the ex-post belief of agent j to become more congruent. Indeed, agent i deems other signals revealed to player j distributed according to i's belief, hence, he expects player j to be persuaded and adjust his ex-post belief in the "right" direction (from i's point of view). Second, the magnitude of the effect of an additional signal on j's action decreases with k_j. However, the additive nature of the state S insures that the rate of decrease is sufficiently low, so that the effect of i's message on j's action remains significant enough to prevent player i from misreporting to player j (given that i expects j to become more congruent).

Chapter 3 (joint with Petra Persson). In Chapter 3, we study how a benevolent authority decides whether to constrain or inform a population of agents (continuum of individuals i). We start by considering an *advisor* who can issue recommendations, but not enact mandates. She privately obtains an imprecise piece of information s about an unknown state of the world $\theta \in \{0, 1\}$, and sends a public message, before each individual i chooses action a_i. Lying entails a cost; modulo this cost, talk is cheap. The advisor's material payoff is maximized by each individual choosing $a_i = \theta$, and individual i's ideal action is $a_i = \theta + b_i$. The advisor's opinion (prior belief) about θ is $\pi_A = \Pr(\theta = 1)$; each individual's opinion is π_i. Under *preference disagreement*, the population's preferences are characterized by a distribution $f(b)$, but their opinions concur with that of the advisor, $\pi_i = \pi_A \ \forall i$. Under *opinion disagreement*, the population's opinions are characterized by a distribution $g(\pi)$, but their preferences concur with those of the advisor, $b_i = 0 \ \forall i$. For each type of disagreement, we analyze how the ability to sustain truthful communication depends on the advisor's *benevolence*, or altruism, modeled as the share of the individuals' material payoffs that she internalizes, $\varphi \geq 0$.

Under preference disagreement, altruism improves communication. As the advisor's altruism becomes stronger, the action that the advisor wants each individual to choose approaches the individual's own preferred action. Higher altruism is thus akin to smaller preference disagreement; as disagreement lessens, truthful communication becomes attainable. By contrast, under conflicting priors, altruism can destroy communication. In this case, the advisor is convinced that her preferred action, given the signal s, maximizes both her own *and* each individual's expected welfare. Each individual, however, would interpret a truthfully revealed signal in light of his own prior, and choose a different action. Even though the advisor represents the median opinion in the population, she may believe that, on average, the individuals' are better off with the action choices they take when she lies. Lying then protects them from misinterpreting a truthful report. When φ increases, the advisor internalizes more of the disutility that she expects each individual to suffer from his (in her view) suboptimal choice of action. Paradoxically, a sufficiently altruistic advisor may therefore lie.

We then consider an *authority* who, after observing the signal s, either can send a public message, or incur some cost $q \geq 0$ to mandate one action for all individuals. Under preference disagreement, enacting a mandate is unattractive to a benevolent authority, for two reasons: First, truthful communication can be sustained. Second, if the authority lets each individual i choose his action, then i implements an action that is close to the action that the highly altruistic authority would want him to choose. Consequently, while a self-interested authority may enact a mandate, a benevolent authority instead communicates truthfully, and gives each individual the liberty to choose. By contrast, under opinion disagreement, mandating an action is attractive to the altruistic authority for two reasons: First, she may not be credible; if she would allow the individuals to choose their actions, they would thus base their choices on less information than she would do. Second, when the authority is sufficiently altruistic, she would enact a mandate even if truthful communication were possible, because she knows that the actions that the individuals would take differ from the action that she deems optimal for them. Consequently, while a self-interested authority may communicate truthfully, a benevolent authority instead mandates an action, believing that she acts in the population's interest.

Chapter 4. In Chapter 4, I study strategic communication setting of [Crawford and Sobel, 1982] with guilt averse Sender, who suffers from letting down the Receiver's payoff expectation. That is, the Sender faces the cost $k \cdot c(x, e)$ (where k is the cost intensity parameter), if the Receiver's actual payoff x after the message m appears to be lower than the Receiver's expected payoff e (conditional on the message m). It is straightforward to see, that the psychological cost of lying arising from guilt aversion is *not* a specific case of the literal cost approach, in which the lying cost depends only on the state of the world and the message sent ([Kartik, 2009], [Kartik *et al.*, 2007]). Indeed, in the guilt aversion case, the Sender suffers from the cost when the Receiver gets a lower payoff than the Receiver expected. Hence, the level of the cost depends on how the Receiver interprets the message and forms his payoff expectation, and not on the literal difference between the actual and the reported states.

Recall, that the outcome of the classic [Crawford and Sobel, 1982] model is a partition equilibrium, in which the Sender indicates only the interval where the state of the world lies. Introducing literal cost of lying ([Kartik, 2009]) results in the ability of the lower type senders to separate themselves through an inflated language. On the contrary, I show that introducing quilt aversion cost of lying precludes the existence of equilibria with separating intervals of types. Instead, under the assumptions of uniform state distribution and quadratic payoff functions, there exist partition equilibria like in [Crawford and Sobel, 1982].

Assume that the Sender has a persistent positive preference bias, i.e., the Sender wants a higher action than the Receiver, given any state of the world. I demonstrate that as in [Crawford and Sobel, 1982] model, in any partition equilibrium, the higher type senders transmit less information than the lower type senders: the intervals in the equilibrium partition expand as the Sender's type increases. Increasing the level of guilt aversion (or decreasing the preference divergence), while holding the number of partition elements fixed, does not change this pattern, but makes the intervals more balanced. Further, an increase in the cost intensity (or a decrease in the preference bias) allows for a greater number of intervals in the equilibrium partition.

Chapter 1

Strategic communication in networks: The choice between soft and hard information transmission

Uliana Loginova

Abstract

I study a model in which every agent needs to take an action that matches his preferences given the state of the world and that action affects the payoffs of all other agents. Before deciding upon the action, agents can choose to whom and how to reveal their private information about the state. There are two ways in which the information can be communicated: either through a costly *verifiable information (hard)* channel or through a low-cost *cheap talk (soft)* channel. The information transmission pattern is characterized by a strategic communication network whose links represent truthful information transmission via either channel. I characterize and compare the pure cheap talk setting with the setting in which verifiable information transmission is feasible. I find that the availability of hard links allows each agent to get a weakly greater number of truthful messages compared to the pure cheap talk setting. If only one party bears the cost of a hard link, then introducing hard links increases the total expected welfare. In contrast, if the cost of a hard link is shared by both parties, then allowing for verifiable communication can decrease the total welfare.

1.1 Introduction

In many economic situations, people are required to make judgements about the uncertain state of the world and pick the most appropriate actions that influence other participants as well. Oftentimes, these decisions are based not only on the private information of the individuals, but also on the information collected from other agents during preliminary discussions. Clearly, the stage of information transmission plays an important role in such settings. The literature that studies possible communication patterns can generally be classified by the type of information transmission technology into two categories. One approach follows [Grossman, 1981] and [Milgrom, 1981] in assuming a verifiable (or hard) type of information, i.e., information that can be withheld by the agent but not lied about (*hard talk*). The second approach complies with [Crawford and Sobel, 1982] and [Green and Stokey, 2007] in considering an unverifiable (or soft) type of information, i.e., information that can be arbitrarily misreported by the agent at no cost (*cheap talk*). The overwhelming majority of theoretical studies share the assumption that communication mode, either cheap talk or hard talk, is a feature of the environment, and hence is a fixed characteristic of communication stage.[1]

While a predetermined communication mode might be natural in many cases, there are other situations in which the participants not only choose with whom to communicate, but also get to determine the way in which the information is transferred. Indeed, in reality communication of the same piece of private information can take different forms, from sending one-sentence emails to prolonged discussions. With the extreme options, it is natural to assume that the email communication is cheap and the discussion is costly, because sending a one-sentence email requires considerably less time and effort than participating in long discussions. Moreover, one would expect that a one-sentence email gives a decision-relevant summary of the sending party's private information without providing supplementary materials and explanations that will allow the re-

[1] One exception is [Eső and Galambos, 2008] who extend the model of [Crawford and Sobel, 1982] to allow for costly provision of hard evidence, and show that contrary to the pure cheap talk case, a greater preference bias can lead to more informative communication. Other exceptions are [Kartik *et al.*, 2007] and [Kartik, 2009] who consider the case when it is costly to misreport the information and show that the equilibrium outcome involves separation with inflated language. [Dessein and Santos, 2006] assume that the information transmission takes the form of a noisy hard talk and allow the organization to choose the level of communication precision. [Dewatripont and Tirole, 2005], [Calvó-Armengol *et al.*, 2011] and [Persson, 2011] depart from the predetermination of communication mode by considering the setting where the trustworthiness of information transmission is endogenously defined by the communication efforts of the parties.

ceiving party to verify the message. Such limitation leaves the composition of the message entirely up to the sender's discretion. At the same time, prolonged discussions can give an opportunity for the reporting party to present the supporting data and provide the necessary justifications, so that the receiving agent is able to uncover the decision-relevant piece of information himself. From this point of view, sending a one-sentence email is closer to cheap talk communication, while prolonged meetings resemble verifiable information transmission; and often it is up to the participants which type of communication to engage in. Clearly, in such cases, the assumption about a predetermined communication mode might be substantially limiting.

The aim of this chapter is to analyze communication patterns that arise when agents are allowed to choose the way they communicate, soft or hard, and to evaluate the effect of providing the individuals with such a choice compared to the predetermined cheap talk communication mode setting. The framework I study builds on [Galeotti *et al.*, 2011] with the possibility to form not only soft links, but also hard links. In the model, there are n players characterized by preference biases $b_1, ..., b_n$. Each player i receives a private signal s_i about the unknown state of the world θ, which has a common prior of Beta distribution with parameters (α, β). The signals are independent and identically distributed, with each s_i taking a value of 1 with probability θ and a value of 0 with probability $1-\theta$. Players can simultaneously transmit their private signals to each other according to the communication network, which is set prior to the signals' realization. The network is described by a directed graph of hard and soft links. Player i can send messages to other player that he has links to in compliance with the link type: if the link is soft, then reporting takes the form of cheap talk, while if the link is hard, then communication is a non-strategic direct revelation of the signal s_i to the other party.

While soft links are very cheap to sustain, hard links are costly. The way the cost of a hard link is split between the involved parties is defined by a specific cost structure that is the same for all pairs of agents. That is, the costs of outgoing and incoming hard links are fixed across agents. While considering a general cost structure where the cost is split arbitrarily between the parties, I distinguish between the following two cases. The first case is one in which only one party, either the sender or the receiver, bears the cost of a hard link. The structure with only the sender paying the cost naturally arises when it takes considerable effort to to develop argumentation and to report the findings of the analysis, while it is very easy for the receiver to uncover the underlying signal after being presented with the collected materials. Similarly, the cost structure where only the receiver incurs the cost corresponds to a situation in which it is much easier to formulate the message and provide the material, than it is to decode the underlying signal. The second case is one in which both parties bear strictly positive costs of a hard link and represents situations in which both the sender and the receiver are required to put in considerable effort to transmit and understand the information ([Dewatripont and Tirole, 2005] and [Persson, 2011]).

After the communication stage, players simultaneously choose actions $\{y_1, ..., y_n\}$ that influence each other's payoffs, such that every player i obtains $-\sum_{j=1}^{n}(y_j - \theta - b_i)^2$. That is, player i's payoff depends on how close his action, as well as the actions of other players, are to player i's ideal action, $\theta + b_i$.

As an example of this setting, consider a group of managers of an international corporation, who have to decide on the strategies for their respective divisions. Managers are located in different countries, each being responsible for the corporation's performance in the manager's respective location. The effect of chosen strategies on the firm's well-being depends on the economic environment, which is uncertain. However, each manager holds some private information about the characteristics of the economic environment. Corporation divisions might be endowed with different goals and interests depending on the location, which implies preference divergence among the managers. Nevertheless, different divisions are parts of one corporation, meaning that the strategy chosen by some manager has an effect on other divisions' payoffs. Before making their decisions, managers can communicate their private information to each other. Communication can take one of the two forms: a short email or a personal meeting. An email is easy to compose and send, but its content can not be verified by the receiver. At the same time, personal meetings require traveling and spending extra time and effort on preparation, but necessarily reveal the underlying private information. The burden of personal meetings—the cost and the way it is divided between the parties—depends on the corporate policy. In particular, the corporation can either encourage personal discussions by appropriate financing and making traveling a pleasant experience, or can discourage managers from meeting with each other by providing limited traveling assistance.

To study the model, I define an *equilibrium* to be an extension of pure strategies Perfect Bayesian Equilibrium (PBE). Namely, communication and action strategies form the usual pure strategies PBE, holding the communication network fixed. At the same time, the communication network must be such that no player prefers to delete a link, and no two players prefer to form a hard link in the absence of truthful communication, given rationally updated communication and decision making strategies. The resulting communication pattern is described by a directed truthful network, in which every link corresponds to truthful communication through a soft or hard link.

While both parties always desire the truthful revelation of a signal from the ex-ante perspective, there is a credibility issue in strategic communication via a soft link at the interim stage. Similar to [Galeotti *et al.*, 2011], I find that a decreasing marginal effect of an additional truthful message leads to a *congestion effect* in cheap talk communication: the willingness of player i to report truthfully to player j decreases with the number of truthful messages player j gets, called the *in-degree* of player j and denoted by k_j. Moreover, the incentive to misreport increases with the preference divergence between the players. In contrast, the incentive to communicate through the hard link does not depend on the preference divergence. The benefit of forming a hard link ij from player i to player j depends only on the number of truthful messages that j gets, k_j, and is strictly decreasing in k_j. This implies that hard links are directed towards only those players who have a sufficiently low number of incoming truthful soft links.

The first—intuitive—result of the chapter describes the effect of introducing the possibility of forming hard links on the scope of information transmission. More precisely, start from the pure cheap talk setting, i.e., verifiable information transmission is prohibitively costly and all communication is performed through soft links. Then introduce feasible hard links, i.e., lower the cost of hard links such that they can emerge in the equilibrium truthful network. As a result, there is a new equilibrium that generates more intensive information transmission. Using the introduced notation, it means that a new equilibrium with hard links generates a truthful network with weakly larger in-degrees for all players.

The second—and main—result concerns the welfare aspect of introducing the feasible verifiable information transmission channel. The fact that there are almost always multiple equilibria presents some difficulties in comparing the welfare across different settings. I resolve the multiplicity issue by introducing the notion of pairwise stability—a natural equilibrium refinement. This notion is an analogue to the pairwise stability condition commonly used in network theory, but adapted to the defined equilibrium concept. More precisely, an equilibrium is called *pairwise stable*, if no two players can find an incentive compatible way to improve interaction between them by communicating truthfully through a soft link instead of a costly hard link or instead of not communicating at all, holding other communication and decision making strategies fixed.

The welfare outcome of introducing feasible hard links in the initial pure cheap talk setting is composed of two effects: the positive effect of an *information improvement* and the negative effect of *crowding out cheap talk communication* with costly verifiable information transmission. The information improvement effect is the substance of the informational result described above: availability of verifiable information transmission leads to greater and more evenly distributed in-degrees. The crowding out effect is at the heart of the chapter. As hard links become available, some agents who could not truthfully communicate via cheap talk can find it beneficial to pay the cost and transfer the private information through hard links. The appearance of hard links increases players' in-degrees which, given the congestion result, can render truth-telling through some soft links no longer incentive compatible.[2] Those soft links should be replaced by hard links for the players to remain credible. Thus, even though players get more truthful messages, the number of soft links might decrease, implying that the cheap talk communication is crowded out by the costly verifiable information transmission.

I find that the final welfare outcome of the two forces depends on the cost structure. In the case where only one party bears the cost of a hard link, there always exists a pairwise stable equilibrium with feasible hard links that generates greater total welfare than in the pure cheap talk case. Interestingly, the positive welfare result no longer holds when the cost of a hard link is shared between the players. In particular, I follow [Galeotti *et al.*, 2011] in analyzing communication between two communities composed of players with the same preferences. In this setting, I find that introducing feasible hard links does not alter soft intra-group communication—people with the same preferences can always be credible to each other through a cheap talk. At the same time, newly created hard links can make cross-group cheap talk communication no longer credible, wipe out all cross-group soft links, and replace them with costly hard links. As a result, the total welfare can decrease relative to the pure cheap talk case.[3] As another extreme, I consider the case of a

[2] Clearly, the vulnerability of a soft link increases with the divergence in preferences.

[3] The observation that communication between the individuals with different characteristics tends to be more substantial and proof-oriented than between "similar" individuals, adds to a new perspective to the study of homophily presented in [Galeotti *et al.*, 2011], i.e., "the tendency of individuals to associate and exchange information with others who are similar to themselves." Also on this topic see, for example, [Lazarsfeld and Merton, 1954], [McPherson *et al.*, 2001], [Moody, 2001], [Currarini *et al.*, 2009].

diverse group of people who have equidistant preference biases, and show that adding the possibility of forming hard links can decrease the total welfare only when the number of players is 3.

Finally, I study two natural extensions of the model. In the first extension, I allow the players to negotiate how to split the cost of a hard link between them. Surprisingly, I find that endogenizing of the cost shares does not imply aggregate efficiency and introducing hard links can still lead to lower total welfare—which reinforces the main result of the chapter.

In the second extension, I allow the costs of hard links differ across the pairs of players. This might happen for various reasons, e.g., people working at the same location might face a lower cost of verifiable information transmission that takes the form of personal meetings, than people from different locations. I illustrate that when the difference in costs is substantial, introducing hard links likely results in *localization* of communication—pairs of players with a low cost communicate with each other via hard links. At the same time, the amount of information accumulated by every player can remain the same. This implies a decrease in the total welfare compared to the pure cheap talk case, even when only one party bears the cost.

The chapter is organized as follows. Section 1.2 presents the model and discusses the solution concept. Section 1.3 provides the incentive for truthful reporting via a soft link, the benefit of forming a hard link, and their implications for the equilibrium communication networks. The informational result of introducing verifiable information transmission is presented in Section 1.4, and the welfare results are analyzed in Section 1.5. The extensions of the model with endogenous cost shares and heterogeneous hard links costs are considered in Section 1.6. Section 1.7 concludes the chapter. Finally, Appendix 1.8 provides some additional details on characterization of the pairwise stable equilibria, and Appendix 1.9 contains all the proofs omitted in the main text.

1.1.1 Related literature

This chapter contributes to the literature that considers settings with a fixed communication mode: cheap talk or verifiable information transmission. While this literature is very rich, in this section I mention only the few of most related studies. The most closely related work is [Galeotti *et al.*, 2011], and as already discussed, the current chapter departs from their framework mainly by incorporating additional means of verifiable communication. Another closely related paper is [Hagenbach and Koessler, 2010], who study strategic cheap talk communication in networks. Their model differs in two important ways. First, in their model the underlying state of the world is determined as a sum of signals sampled from known distributions, which implies a constant marginal value of every additional signal. As a result, there is no congestion effect in their model. Second, they consider a different type of coordination, i.e., each player not only wants others' actions to be closer to his action (as is the case in my model), but also also wants to match his action with the actions of others.

Other related papers with cheap talk communication include settings with multiple senders (e.g. [Morgan and Stocken, 2008]) and multiple receivers (e.g. [Caillaud and Tirole, 2007]). While in this chapter I focus on private communication, there are studies that compare private with public information transmission, e.g. [Farrell and Gibbons, 1989], [Goltsman and Pavlov, 2011] and [Galeotti *et al.*, 2011].

Results on how the structure of a hard link cost influences communication patterns contribute to the literature on organizational design with cheap talk (e.g. [Alonso *et al.*, 2008], [Rantakari, 2008]), verifiable information (e.g. [Bolton and Dewatripont, 1994], [Radner, 1992], [Radner, 1993], [Sah and Stiglitz, 1986], [Van Zandt and Radner, 2001]) and noisy hard talk, where the information is transmitted perfectly with some probability less than 1 (e.g. [Dessein and Santos, 2006]). A finding that allowing for verifiable communication can lead to a welfare decreasing excessive hard link formation is related to [Dessein, 2007]. In his model, a committee of privately informed members needs to make a decision and the communication process combines soft and hard information. Dessein shows that authoritative decision-making can outperform majority decision-making, because it avoids costly rent-seeking discussions and reduces lobbying for mediocre alternatives.

Also related are studies that focus on questions of coordination and adaptation with verifiable information transmission in communication networks. In particular, [Chwe, 2000] studies a collective action problem with preliminary communication regarding participation activity in a deterministic exogenous network. [Calvó-Armengol and de Martí, 2007] and [Calvó-Armengol and de Martí, 2009] analyze how the communication pattern affects individual behavior and aggregate welfare in a setting in which the agents not only want to coordinate their actions, but also adapt to an unknown state of the world. Instead, [Calvó-Armengol *et al.*, 2011] consider local uncertainty regarding the state and study information transmission patterns that arise when the agents are allowed to alter the communication precision.

Finally, the analysis of the communication patterns arising in equilibria contributes to the literature of strategic network formation, which includes [Bala and Goyal, 2000], [Goyal, 2007], [Jackson, 2008], [Jackson and Wolinsky, 1996].

1.2 Model

Let the set of players be $N = \{1, ..., n\}$ with $n \geq 2$. Each player i has preference bias b_i; the preference profile $\{b_1, ..., b_n\}$ is common knowledge. The state of the world θ is unknown and has a density of Beta distribution with commonly known parameters (α, β): [4]

$$f(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Each player i receives a private signal $s_i \in \{0, 1\}$ about θ, where $s_i = 1$ with probability θ and $s_i = 0$ with complementary probability $1 - \theta$. Private signals are assumed to be independent and identically distributed.

The communication network is set prior to the signals' realization and is described by a directed graph $g \in \{0, s, h\}^{n \times n}$, where $g_{ij} = s$ if and only if a soft link ij is present, meaning that i reports to j via a cheap talk channel; $g_{ij} = h$ if and only if a hard link ij is present, meaning that i communicates his signal to j via a verifiable information channel; $g_{ij} = 0$ if and only if there is no link from player i to player j.[5] It is assumed that hard links are costly and the way the cost of a hard link $C \geq 0$ is split between the players depends on the specific cost structure. Once the specific cost structure is fixed, it is the same for all players. For the main part of the chapter, I assume cost C of a hard link to be the same for all pairs of players; the case in which cost can vary with the players' identities is considered in the Extensions section. At the same time, soft links are cheap to sustain: each involved party bears just an infinitesimal cost $\varepsilon > 0$.

While the players are aware of each other's existence and preference biases, the communication

[4] Considering a general Beta distribution is a generalization of the [Galeotti *et al.*, 2011] model, where they use a uniform on $[0, 1]$ prior, which is a particular case of Beta distribution with parameters $(1, 1)$.

[5] The assumption that the communication network is defined before the realization of the signals, for instance, can be justified by the necessity of forming the communication schedule beforehand. Another case where this assumption is appropriate corresponds to repeated interactions, where each period there is a new draw of the state θ, then signals arrive and decisions are made, while the communication stage is a part of the routine schedule that stays constant across periods.

network g is not common knowledge. Rather, each player i knows only the structure of his respective neighborhood. That is, player i observes the set of players to whom he has soft links, denoted as $N_i^s(g) = \{j \in N : g_{ij} = s\}$; and the set of players to whom he has hard links, denoted as $N_i^h(g) = \{j \in N : g_{ij} = h\}$. Similarly, player i observes the set of players who have soft links directed to him, $N_i^{-1,s}(g) = \{j \in N : g_{ji} = s\}$; and the set of players who have hard links directed to him, $N_i^{-1,h}(g) = \{j \in N : g_{ji} = h\}$.[6]

Cost structure. If there is a hard link ij from player i (the sender) to player j (the receiver), then player i bears a share $\alpha \in [0, 1]$ and player j bears a share $1 - \alpha$ of the total link cost C. I distinguish between two cost structures:

(i) *Only one party bears the cost of a hard link, C.* $\alpha = 1$ corresponds to a case in which the sender has to exert some effort to create and deliver a message, which then allows the receiver to easily extract the signal. On the other hand, $\alpha = 0$ represents a case in which it is costless for the sender to elaborate the message, but then the receiver has to make a costly effort in order to understand it and reveal the underlying signal.

(ii) *The sender and the receiver share the cost of a hard link, C,* with the weights $\alpha \in (0, 1)$ and $(1 - \alpha)$, respectively. This happens when player i, as a sender, needs to exert some effort to elaborate and deliver the message about the signal, while player j, as a receiver, needs to make an effort in order to understand the message and learn what the underlying signal is.

Communication. Each player i sends private messages to the players that he has links to according to the respective links' nature in the communication network g: if a link ij is hard, then the message sent to player j truthfully reveals the signal, $m_{ij}^g = s_i$; if a link ij is soft, then any message $m_{ij}^g \in \{0, 1\}$ can be sent. It is assumed that the messages are sent simultaneously and are observed only by the sending and the receiving parties. Note, that communication via hard links is non-strategic—the message perfectly reveals the signal.[7] In contrast, communication through soft

[6] The assumption that the players observe only the structure of their respective neighborhoods, while being more realistic, is not without loss of generality. If instead the network were commonly observed, then the players could condition their strategies on the whole network structure and more communication patterns—including very unlikely ones—could have be induced.

[7] Note that, due to the unraveling argument, the perfect revelation of the underlying signal is also the outcome in the usual verifiable information setting with the option to withhold information. Indeed, assume that the message

links is in the form of cheap talk, and hence is strategic. A *communication strategy* of player i with the private signal s_i defines a vector

$$\mu_i^g(s_i) = \left\{ \{\mu_{ij}^g(s_i)\}_{j\in N_i^s(g)}, \{\mu_{ij}^g(s_i)\}_{j\in N_i^h(g)} \right\},$$

where $\{\mu_{ij}^g(s_i)\}_{j\in N_i^s(g)} \in \{0,1\}^{|N_i^s(g)|}$ and $\mu_{ij}^g(s_i) = s_i$ for every $j \in N_i^h(g)$. A communication strategy profile is denoted by $\mu^g = \{\mu_1^g, \ldots, \mu_n^g\}$. The messages actually sent by player i are denoted by vector $\widehat{m}_i^g$, while the profile of all sent messages is $\widehat{m}^g = \{\widehat{m}_1^g, \ldots, \widehat{m}_n^g\}$. The superscript g signifies the dependence of the communication strategies on the network structure. I use the same superscript g for strategies of different players to simplify the notation, however, one needs to keep in mind that every player i conditions his communication strategy only on the available information about the communication network—the structure of player i's neighborhood.

Decision making. After the communication stage, each player i chooses an action $y_i^g \in \mathbb{R}$. Denote the set of all players who have a link to i as $N_i^{-1}(g) = \{j \in N : g_{ji} = s \text{ or } g_{ji} = h\}$. Because the information set of player i consists of his own signal, s_i, and the messages he gets from $N_i^{-1}(g)$, $\widehat{m}^g_{N_i^{-1}(g),i}$, the *action strategy* of player i is a function $y_i^g : \{0,1\} \times \{0,1\}^{|N_i^{-1}(g)|} \to \mathbb{R}$. Let $y^g = \{y_1^g, ..., y_n^g\}$ denote the action strategy profile.[8] Conditional on the state of the world θ, if the chosen action profile is $\hat{y}^g = \{\hat{y}_1^g, ..., \hat{y}_n^g\}$, then the realized payoff (utility) of player i is

$$u_i(\hat{y}^g|\theta) = -\sum_{j=1}^{n} (\hat{y}_j^g - \theta - b_i)^2.$$

Player i's payoff depends on how close his own action and the actions of other players are to player i's ideal action, $\theta + b_i$.[9]

Time notation. For further analysis, it is useful to introduce the following time notation to distinguish among periods with different scopes of information available to the agents: *"ex-ante"*

space is $\{0, 1, \{0, 1\}\}$. If player i reports to player j with the bias $b_j > b_i$ via a hard link, then player i would always choose to reveal the signal 0. This, in turn, leads to the unique full revelation outcome.

[8] Superscript g emphasizes the dependence of the action strategies on the communication network: each player i knows the structure and the type of links in his neighborhood.

[9] The chosen functional form of the payoff simplifies the technical analysis; all the main results can be extended to more general settings as well.

to denote the stage prior to when the signals are realized (such that the only information about the state of the world each player has is the common prior), *"interim"* to refer to the time period after the signals' realization but prior to communication (such that each player knows the common prior and his private signal), and finally, *"ex-post"* - for the period after communication has occurred but before the actions are taken (such that each player knows the common prior, his private signal, and reported messages).

Solution concept. I solve the model using the concept of pure strategies Perfect Bayesian Equilibrium (PBE). The restriction to pure strategies simplifies the analysis and implies that cheap talk communication can take two forms: *truthful*, where the message reflects the signal perfectly, or *uninformative*, where, for any signal s_i, player i sends the same message, either 0 or 1. In the latter case, I assume that when a player gets a message which is off the equilibrium path, he ignores it and does not update his belief. Because of this simplification, equilibrium beliefs are defined in a straightforward way: any message received through a hard link or in truthful communication through a soft link induces perfect knowledge about the underlying signal, while any message received in uninformative communication through a soft link leaves the prior belief about the underlying signal unchanged.

Holding the communication network g fixed, it is natural to determine the communication and action strategy profile $(\mu^g, y^g) = (\{\mu_i^g\}_{i \in N}, \{y_i^g\}_{i \in N})$ by using the standard PBE solution concept. However, conditional on a particular choice of g and equilibrium (μ^g, y^g), some soft and costly hard links in g might be ex-ante undesired by at least one party involved in the link (i.e., a sender or a receiver with respect to this link). In particular, all soft links with uninformative communication are ex-ante unprofitable to both parties.[10] On the other hand, it might as well be the case that the two players, i and j, would prefer to have a costly hard link ij in order to be able to directly transmit the signal s_i to j, rather than having no link or having a soft link with uninformative communication, while holding all other strategies fixed. One of the underlying ideas for the model is that people can to some extent manage the links themselves and might object to existence of soft links with uninformative communication and to some hard links. To account for this, I define

[10] Note, that no player will object to a cheap soft link with truthful communication. Indeed, as it is shown in Lemma 2, destroying the soft link with truthful communication will strictly harm the ex-ante expected payoffs of both parties involved in the link (provided that the cost ε is infinitesimal).

an equilibrium as a communication network g coupled with a strategy profile (μ^g, y^g) such that (i) the pair (μ^g, y^g) forms a PBE given g, (ii) no player would strictly prefer to break some incoming or outgoing link from an ex-ante perspective, and (iii) no two players, i and j, would strictly prefer to form a costly hard link $g_{ij} = h$ from an ex-ante perspective, holding other communication and action strategies fixed.[11]

To formally state the equilibrium definition, it is useful to denote by $\mathbb{E}u_l\left(g, \mu^g, y^g\right)$ the ex-ante expected utility of agent l, where (μ^g, y^g) are communication and strategy profiles given some communication network g.[12] Let $g(g_{ij} = 0)$ be the communication network with the same set of links as in g, except that there is no link from i to j; let $\mu^{g(g_{ij}=0)}$ be the profile of communication strategies which coincides with μ^g everywhere, except that now there is no communication from i to j; and let $y^{g(g_{ij}=0)}$ be the same action profile as y^g for all players but j, while player j's action is now optimally defined conditional on the lower number of truthful messages. Similarly, let $g(g_{ij} = h)$ denote the communication network with the same set of links as in g, except that there is a hard link from i to j. Then $\mu^{g(g_{ij}=h)}$ is the profile of communication strategies which coincides with μ^g for all links but g_{ij}, through which the communication is truthful. Finally, $y^{g(g_{ij}=h)}$ is the same action profile as y^g for all players but j, while player j's action is now optimally defined given the new information structure. Using this notation, below is the formal equilibrium definition:

Definition. *Equilibrium* $\{g, (\mu^g, y^g)\}$ consists of a communication network g and a strategy profile $(\mu^g, y^g) = (\{\mu_i^g\}_{i \in N}, \{y_i^g\}_{i \in N})$, such that the following properties hold:

(i) The pair (μ^g, y^g) forms a PBE given the communication network g.

(ii) For any i and j such that $g_{ij} = h$ or $g_{ij} = s$:

$$\mathbb{E}u_l\left(g, \mu^g, y^g\right) \geq \mathbb{E}u_l\left(g(g_{ij} = 0), \mu^{g(g_{ij}=0)}, y^{g(g_{ij}=0)}\right), \quad \text{for } l = i, j.$$

[11] So far I don't consider a similar desire to form a soft link, because communication through a soft link is strategic, and the mere existence of a soft link does not guarantee informativeness of the communication pattern. Adding a soft link with truthful communication requires more coordination at the communication stage, otherwise the communication might be completely uninformative. The possibility of adding a soft link and coordinating on the respective communication pattern is considered later in the context of pairwise stability.

[12] Note that the ex-ante expected utility $\mathbb{E}u_l\left(g, \mu^g, y^g\right)$ takes into account all link costs that accrue to agent l.

(iii) For any i and j such that $g_{ij} = 0$ or $g_{ij} = s$:

$$\mathbb{E}u_l\left(g, \mu^g, y^g\right) \geq \mathbb{E}u_l\left(g(g_{ij} = h), \mu^{g(g_{ij}=h)}, y^{g(g_{ij}=h)}\right), \quad \text{for either } l = i \text{ or } l = j.$$

Remark. In any equilibrium $\{g, (\mu^g, y^g)\}$, communication network g is *truthful*, i.e., all links of g represent truthful revelation of private signals. Different equilibria correspond to different communication networks.

1.3 Analysis

1.3.1 Optimal action

I start the analysis by deriving the optimal action choice holding fixed the communication network g. Given that agent i got the private signal s_i and received messages $\widehat{m}^g_{N_i^{-1}(g),i}$, he chooses an action $y_i^g(s_i, \widehat{m}^g_{N_i^{-1}(g),i})$ to maximize his expected payoff,

$$\mathbb{E}\left(-\sum_{j=1}^{N}(y_j^g - \theta - b_i)^2 \,\middle|\, s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right),$$

which means that the agent chooses

$$\begin{aligned} y_i^g(s_i, \widehat{m}_{N_i^{-1}(g),i}) &= \arg\max_{y_i^g}\left\{\mathbb{E}\left(-(y_i^g - \theta - b_i)^2 \,\middle|\, s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right)\right\} \\ &= b_i + \mathbb{E}\left(\theta | s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right). \end{aligned} \tag{1.1}$$

Given that the cheap talk communication is assumed to be either truthful or completely unrevealing, the information set of player i, $\left(s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right)$, can be equivalently represented as the set of revealed signals. In particular, assume that player i gets to know k signals, l of which are 1. The conditional probability that l signals out of k are 1 given θ is

$$f(l|\theta, k) = \frac{k!}{l!(k-l)!}\theta^l(1-\theta)^{k-l}.$$

The posterior of θ, $f(\theta|l, k)$, is proportional to the product of the prior and the conditional probability, $f(\theta)f(l|\theta,k)$. Hence, the posterior has Beta distribution with parameters $(\alpha+l, \beta+k-l)$:

$$f(\theta|l,k) = \frac{1}{B(\alpha+l,\beta+k-l)}\theta^{\alpha+l-1}(1-\theta)^{\beta+k-l-1},$$

and the expected value of θ then is $\mathbb{E}(\theta|l,k) = \frac{\alpha+l}{\alpha+\beta+k}$.

1.3.2 Equilibrium networks

Assume that the strategy profile (μ^g, y^g) is such that the communication network g is truthful: communication through each link of g, either soft or hard, leads to a signal revelation. Define $k_j(g)$ to be the number of other players who report truthfully to j via either channel, and refer to it as the *in-degree* of player j.

Benefits of signal revelation. For the analysis of equilibrium networks, it is useful to study the ex-ante expected additional benefit that accrues to some player $i \in N$ from an extra signal revealed to player j. In order to do this, assume that k_j players report truthfully to player j. This means that, together with his own private signal, player j gets to know a set of k_j+1 signals, summarized by a vector s_R. Based on the information s_R, j chooses action $y_{s_R} = b_j + \mathbb{E}(\theta|s_R)$. The ex-ante expected input from player j into player i's payoff in this case is

$$-\int_0^1 \sum_{s_R\in\{0,1\}^{k_j+1}} (y_{s_R}-\theta-b_i)^2 f(\theta,s_R)d\theta.$$

The following lemma shows that this input consists of the two components: the expected residual variance of the state θ and the square of the preference divergence.

Lemma 1. *Fix a truthful network g and consider player j with the in-degree of $k_j = k_j(g)$. Let s_R be the set of signals that player j gets to know and y_{s_R} be the corresponding action chosen by j. For any $i \in N$ the ex-ante expected input from player j into i's utility is given by*

$$-\int_0^1 \sum_{s_R\in\{0,1\}^{k_j+1}} (y_{s_R}-\theta-b_i)^2 f(\theta,s_R)d\theta = -h(k_j)-(b_j-b_i)^2,$$

where $h(k_j) = \mathbb{E}\left[Var(\theta|s_R)\right] = \frac{\alpha\beta}{(\alpha+\beta+1+k_j)(\alpha+\beta+1)(\alpha+\beta)}$.

Proof. See Appendix 1.9. □

Assume now that an extra signal is revealed to player j by some player $l \in N$. Then j bases his decision on $k_j + 2$ signals (s_R, s_l). By Lemma 1, the ex-ante expected input from j into i's utility becomes $\mathbb{E}\left[\mathrm{Var}(\theta|s_R, s_l)\right] - (b_j - b_i)^2$. As a result, the ex-ante expected additional benefit that accrues to player i from an additional truthful link directed to j is a change in the expected residual variance of the state:

$$h(k_j) - h(k_j + 1) = \mathbb{E}\left[\mathrm{Var}(\theta|s_R)\right] - \mathbb{E}\left[\mathrm{Var}(\theta|s_R, s_l)\right]. \tag{1.2}$$

Note first, that the extra benefit is positive for all values of k_j, because every additional signal improves the information available to player j, and hence reduces the expected residual variance of θ. Second, the additional benefit of player i given by (1.2) doesn't depend on the preference divergence, $b_i - b_j$. Intuitively, player j chooses an action equal to the expected state θ plus his preference bias b_j, and the only way in which extra signal impacts this action is through the precision of the expected value of θ. Thus, each player $i \in N$, including player j himself, enjoys the same ex-ante expected extra benefit of (1.2) from player j having an extra truthful link directed to him.[13] Finally, because $h(k_j)$ is a decreasing and convex function of k_j, the additional benefit (1.2) is a decreasing function of the number of signals that player j already gets, k_j. Intuitively, the better the information of player j is, the lower is the marginal impact of an additional signal in improving the residual variance, implying a lower ex-ante expected gain in the payoff.

These results are summarized in the following lemma:

Lemma 2. *Fix a truthful network g, and consider player j with the in-degree of $k_j = k_j(g)$. The ex-ante expected benefit of any player $i \in N$ from j receiving an additional signal is $h(k_j) - h(k_j + 1)$. It does not depend on the preference divergence, is strictly positive for any k_j, and is decreasing in k_j.*

Communication through soft links. While Lemma 2 ensures that the truthful communication via a soft link is always desirable from the ex-ante perspective, there is a credibility issue at the

[13] Clearly, the result that player j obtains the same extra benefit as other players is due to the specific form of the utility function, namely, that the terms corresponding to different agents' actions enter the sum with the same coefficients.

interim stage. To study the incentive of truthful reporting via a soft link, consider the case in which agent i has a soft link to agent j. Suppose that j gets to know k_j signals: 1 signal player j gets himself and $(k_j - 1)$ signals he infers from other players' messages, excluding i. Denote a vector of these k_j signals as s_R. Assume that player j believes i's message, i.e., j puts probability 1 on that $s_i = m_{ij}$. Let y_{s_R,s_i} be j's action if he has information s_R and i sends him the true signal $m_{ij} = s_i$; and $y_{s_R,1-s_i}$ be j's action if he has information s_R and i lies by reporting $m_{ij} = 1 - s_i$. Agent i will report truthfully to agent j if and only if this generates a greater interim expected payoff for him:

$$-\int_0^1 \sum_{s_R\in\{0,1\}^{k_j}} (y_{s_R,s_i} - \theta - b_i)^2 f(\theta, s_R|s_i)d\theta \geq -\int_0^1 \sum_{s_R\in\{0,1\}^{k_j}} (y_{s_R,1-s_i} - \theta - b_i)^2 f(\theta, s_R|s_i)d\theta.$$

As is shown in the proof of Theorem 1, this incentive compatibility constraint of truth-telling can be rewritten as

$$|b_j - b_i| \leq \frac{1}{2(\alpha + \beta + k_j + 1)}, \tag{1.3}$$

which means that player i can report truthfully to player j via a soft link as long as player j doesn't get to know too many signals, given their preference divergence. The incentive compatibility constraint (1.3) is exactly the *congestion effect* of [Galeotti *et al.*, 2011] extended to the case of a more general prior of Beta distribution.

To see the intuition behind this, assume that $b_j > b_i$ meaning that the ideal action for player j is greater than the ideal action for player i. The effect of an additional signal on j's action decreases with the number of signals that j gets to know, k_j. Thus, for sufficiently high k_j, the effect of i's message on j's action is so small that i would prefer to lie when $s_i = 1$ and report $m_{ij} = 0$ in order to shift j's action closer towards i's preferred one. On the other hand, when j's in-degree is quite low, the effect of an additional signal on j's action is quite large, in which case misreporting when $s_i = 1$ may shift j's action downward too much, making it undesirable.

Incentives to form/delete hard links. Unlike soft-link communication, information transmission via a costly hard link is not affected by credibility concerns. The decisions regarding the existence of particular hard links are made before the private signals are realized. Thus, in order to study the incentives to form (maintain/not destroy) a hard link ij, I consider the net expected value of ij from the ex-ante perspective using the prior distribution of θ. Fix a truthful network g and assume that k_j other players apart from i report truthfully to player j, i.e., $k_j = |N_j^{-1}(g))/\{i\}|$. By Lemma 2, the ex-ante expected additional benefit that accrues to player i and player j from having a hard link ij is solely in reducing the residual uncertainty regarding the state of the world, $h(k_j) - h(k_j + 1)$.

Consider a general case of the cost structure, where C is distributed between the sender and the receiver with the shares $\alpha \in [0, 1]$ and $1 - \alpha$, respectively. Both players, i and j, would like a hard link ij to be a part of the communication network if and only if the cost paid by each of them is lower than the expected benefit from having the hard link:

$$\max\{\alpha, 1 - \alpha\}C \leq h(k_j) - h(k_j + 1). \tag{1.4}$$

Given the properties of the additional benefit outlined in Lemma 2, a few things can be noted. First, while the willingness to truthfully communicate the signal via a soft link decreases with the preference divergence, the incentive to form (or maintain) a hard link is independent of the preference biases. Second, similar to the condition of credible soft-link communication, the incentive to form a hard link represents a congestion effect. Indeed, because the benefit of the hard link decreases with k_j, the players will be reluctant to form a hard link when k_j is sufficiently high. Finally, the incentive condition to form (or maintain) a hard link to player j (1.4) is the same across every two settings characterized by (C_1, α_1) and (C_2, α_2) such that $\max\{\alpha_1, 1 - \alpha_1\}C_1 = \max\{\alpha_2, 1 - \alpha_2\}C_2$. This also means that the two settings share the same set of equilibria. In particular, the equilibria are the same across the following 3 cases: (i) only the sender bears the cost of the hard link C, (ii) only the receiver bears the cost C, and (iii) the cost of the hard link $2C$ is split equally between the sender and the receiver.

Equilibrium characterization. To fix ideas, I assume that whenever a player is indifferent, the choice is made in favor of forming (or maintaining) a hard link. This assumption along with the incentive conditions (1.3) and (1.4) lead to the following equilibrium characterization:

Theorem 1. *Consider a triple $\{g, (\mu^g, y^g)\}$ and assume that each element of y^g satisfies the optimality condition (1.1). Then $\{g, (\mu^g, y^g)\}$ constitutes an equilibrium if and only if the communication network g is truthful and satisfies the following conditions: for any player j with an in-degree*

$k_j = k_j(g)$ and any player i,

$$\begin{aligned} g_{ij} = s \quad & \textit{only if} \;\; |b_j - b_i| \leq \frac{1}{2(\alpha + \beta + k_j + 1)}, \\ g_{ij} = h \quad & \textit{only if} \;\; \max\{\alpha, 1-\alpha\}C \leq h(k_j - 1) - h(k_j), \\ g_{ij} = 0 \quad & \textit{only if} \;\; \max\{\alpha, 1-\alpha\}C > h(k_j) - h(k_j + 1). \end{aligned}$$

Proof. See Appendix 1.9. □

Intuitively, for player j with an in-degree k_j to get truthful messages through the soft links from players $N_j^{-1,s}(g) = \{i \in N : g_{ij} = s\}$ in equilibrium, the incentive compatibility constraint (1.3) must be satisfied for every $i \in N_j^{-1,s}(g)$. Further, players $N_j^{-1,h}(g) = \{i \in N : g_{ij} = h\}$ reveal their signals to j in a verifiable way whenever no player objects to any existing hard link, i.e., the incentive condition (1.4) is satisfied. Finally, for players $N_j^{-1,0}(g) = \{i \in N : g_{ij} = 0\}$ not to report truthfully to j, it must be the case that there is no player $i \in N_j^{-1,0}(g)$, such that i and j prefer to create a new hard link ij.

More can be noted about the equilibrium networks for extreme levels of the cost C. On the one hand, a sufficiently high cost C precludes the existence of verifiable information transmission in equilibrium. To find the threshold value of C, above which no hard links can be sustained, recall that the benefit from a hard link is a decreasing function of the player's in-degree. Consequently, the expected additional benefit can not exceed $h(0)-h(1)$. This means, that no equilibrium network can have hard links whenever the cost $C > (h(0) - h(1))/\max\{\alpha, 1-\alpha\}$.[14] On the other hand, if the cost C is sufficiently small, then any equilibrium network is necessarily complete. Indeed, if $C \leq h(n-2) - h(n-1)$, then for any two players, i and j, a hard link ij is preferred to no link independently of $k_j \leq n-2$. Hence, in any equilibrium communication network, the in-degree of each player must be $n-1$, i.e., every individual obtains all the information.

1.4 Informational result

In this section I start with the case in which only soft communication is available and show how the introduction of a verifiable communication technology expands the information accumulated

[14] Clearly, if $C > 2(h(0) - h(1))$ then no equilibrium network can have hard links, independently of α.

by each individual—the first (and the intuitive) result of the chapter.

The setting with only cheap communication can be viewed as a previously considered setup with a prohibitively high hard link cost. In particular, assume that $C = C_0 > (h(0) - h(1))/\max\{\alpha, 1 - \alpha\}$ and consider some equilibrium $\left\{g(C_0), (\mu^{g(C_0)}, y^{g(C_0)})\right\}$. Let the in-degrees of players in the equilibrium network be $k_1 = k_1(g(C_0)), ..., k_n = k_n(g(C_0))$. I will refer to such equilibria with only cheap communication as the *pure cheap talk* or the *pure soft-link* equilibria.

Now suppose that the cost of a hard link is decreased, so that hard links can be a part of some equilibrium network. If for every player $j \in N$, $C_1 \geq \frac{h(k_j)-h(k_j+1)}{\max\{\alpha,1-\alpha\}}$, then the considered pure cheap talk equilibrium still remains an equilibrium. If, however, $C_1 < \frac{h(k_j)-h(k_j+1)}{\max\{\alpha,1-\alpha\}}$ for some j, then, by Theorem 1, $g(C_0)$ fails to be an equilibrium network. The question of interest is whether instead there is an equilibrium $\left\{g(C_1), (\mu^{g(C_1)}, y^{g(C_1)})\right\}$, such that the information accumulated by each player is improved, i.e., the in-degrees $k'_1 = k_1(g(C_1)), ..., k'_n = k_n(g(C_1))$ exceed the in-degrees $k_1, ..., k_n$, respectively. The following theorem provides a positive answer: indeed, when communication via hard links becomes feasible, there exists an equilibrium with hard links that is weakly information superior to the pure soft-link equilibrium.

Theorem 2. *Take any cost $C_0 > (h(0) - h(1))/\max\{\alpha, 1 - \alpha\}$ and consider some pure cheap talk equilibrium $\left\{g(C_0), (\mu^{g(C_0)}, y^{g(C_0)})\right\}$. Let the in-degrees in the equilibrium network $g(C_0)$ be $k_j = k_j(g(C_0))$, $j \in N$. Then for any cost $C_1 \leq (h(0) - h(1))/\max\{\alpha, 1 - \alpha\}$ there exists an equilibrium $\left\{g(C_1), (\mu^{g(C_1)}, y^{g(C_1)})\right\}$ in which the players have weakly greater in-degrees:*

$$k'_j = k_j(g(C_1)) \geq k_j \textit{ for any } i \in N.$$

Proof. See Appendix 1.9. □

While the complete proof is presented in Appendix 1.9, here I illustrate the intuition for the result with a particular C_1. Consider a pure soft-link equilibrium that corresponds to the cost C_0 and renumber the players such that their in-degrees in the equilibrium network $g(C_0)$ are increasing in their respective number: $k_1 \leq k_2 \leq ... \leq k_n$. Assume now, that the cost is set to the level of C_1 such that $\max\{\alpha, 1-\alpha\}C_1 \in (h(k_j + 1) - h(k_j + 2), h(k_j) - h(k_j + 1)]$ for some $j \in N$, where $k_j < k_{j+1}$. By Theorem 1, the in-degree of every player must be at least $k_j + 1$, because otherwise, there exists a pair of players who would prefer to form a new hard link. In particular, set $g(C_1)$

such that each player $i = 1, ..., j$ has exactly $k_j + 1$ incoming hard links and no incoming soft links. For other players $j + 1, ..., n$, suppose that the only links directed towards them in $g(C_1)$ are the soft links from the pure cheap talk equilibrium network $g(C_0)$. Clearly, truthful communication through these soft links is still incentive compatible because the in-degrees of the players are the same as in the pure cheap talk equilibrium. Further, no player would like to destroy a hard link directed towards player $i = 1, ..., j$, because $\max\{\alpha, 1-\alpha\}C_1 \leq h(k_j) - h(k_j + 1)$. Also, no two players would want to deviate and create hard links directed towards players $j + 1, ..., n$, because $\max\{\alpha, 1-\alpha\}C_1 > h(k_j + 1) - h(k_j + 2) \geq h(k_i + 1) - h(k_i + 2)$ for any $i = j + 1, ..., n$. Thus, the construction results in an equilibrium network $g(C_1)$ with the in-degrees

$$\underbrace{k'_1 = ... = k'_j = k_j + 1,}_{\text{Hard-link communication}}\ \underbrace{k'_{j+1} = k_{j+1},\ ...\ , k'_n = k_n,}_{\text{Soft-link communication}}$$

that are greater than or equal to the corresponding in-degrees in $g(C_0)$.

1.5 Welfare result

Throughout the section I consider a non-trivial case of a strictly positive hard link cost : $C > 0$.[15] Because there might be multiple equilibria, I focus on those that satisfy the natural refinement of pairwise stability—no two players can profitably deviate by changing the communication pattern between them. As Section 1.4 shows, availability of hard links leads to greater in-degrees which, by Lemma 1, positively affects the total welfare. However, in what follows, I demonstrate that, apart from the positive informational effect, there is a negative crowding out effect: the appearance of costly hard links crowds out costless soft communication which, in turn, harms the total welfare. This section presents the second (and the main) result of the chapter, namely, introducing feasible hard links can be either beneficial and detrimental to the total welfare, depending on the setting and the cost structure.

To gain a better understanding of the welfare implications, I first illustrate the interaction of the two effects in an example with three players. Afterwards, I move to a more general setting and demonstrate that if only one party bears the cost of a hard link, then introducing hard links is

[15] Case $C = 0$ is trivial because each equilibrium network is necessarily complete, and players face zero costs of sustaining it. This implies the highest possible levels of individual and total welfare.

welfare beneficial. Later on, I consider the case in which the cost of a hard link is shared between the parties and derive conditions for the welfare decrease in natural settings of the two communities and the diverse group of people.

1.5.1 Pairwise stable equilibria

There may exist multiple equilibria, and as a natural refinement, I adapt the common notion of pairwise stability from the networks literature (e.g., [Bala and Goyal, 2000], [Goyal, 2007], [Jackson, 2008], [Jackson and Wolinsky, 1996]). In particular, call an equilibrium $\{g, (\mu^g, y^g)\}$ *pairwise stable*, if no pair of players can change the communication pattern between them to improve their ex-ante expected utilities, while satisfying the interim incentive compatibility constraints of truth-telling, holding other strategies fixed. More formally:

Definition. An equilibrium $\{g, (\mu^g, y^g)\}$ is *pairwise stable* if

(i) For any $i, j \in \{1, .., n\}$, $g_{ij} = 0$ only if, holding other strategies fixed, i cannot credibly report to j via a soft link, assuming that j believes i's message, and communication via a hard link is not desired by at least one party.

(ii) For any $i, j \in \{1, .., n\}$, $g_{ij} = h$ only if, holding other strategies fixed, i cannot credibly report to j via a soft link, assuming that j believes i's message.

Characterization of pairwise stable equilibria. In order to prove the existence and study the properties of pairwise stable equilibria, I first define the notion of a *maximal equilibrium* as an equilibrium that generates the maximal vector of in-degrees among all equilibria:

Definition. Equilibrium $\{g, (\mu^g, y^g)\}$ with the in-degrees $k_1 = k_1(g), ..., k_n = k_n(g)$ is *maximal* if for any other equilibrium with the in-degrees $k'_1, ..., k'_n$:

$$k_i \geq k'_i, \; i = 1, ..., n.$$

In turn, the in-degrees $k_1, ..., k_n$ are called *maximal in-degrees.*

The following lemma states that the set of pairwise stable equilibria is non-empty and is a subset of maximal equilibria.

Lemma 3. *There exist a maximal and a pairwise stable equilibrium. Any pairwise stable equilibrium is maximal.*

Proof. See Appendix 1.9. □

Let the individual welfare denote the ex-ante expected individual payoff and the total welfare stand for the sum of all ex-ante expected individual payoffs. Because the pairwise stability incorporates efficient communication in terms of its cost and informativeness, there exists a pairwise stable equilibrium whose total welfare is (weakly) higher than in any other equilibrium.[16] For example, such a pairwise stable equilibrium can be constructed in the following way (I refer to this construction further in the text as well). For each $i \in N$ perform the following procedure: order other players $j \in N/\{i\}$ in the increasing absolute values of their preference divergence from i, $|b_j - b_i|$; let this order be $i_1, ..., i_{n-1}$. Consider the maximal in-degree of player i, k_i. If $k_i = 0$, then nobody can report truthfully to i in equilibrium. If $k_i > 0$, then take the closest player i_1: if truth-telling through a soft link $i_1 i$ is incentive compatible for i_1, given that i gets $k_i - 1$ other truthful messages, then let i_1 report truthfully to i via a soft link. Otherwise, set $g_{i_1 i} = h$. Repeat this procedure for other $k_i - 1$ closest to i players to set the links of particular type with truthful communication through them. In the proof of Lemma 3, I show that this construction leads to a pairwise stable equilibrium, in which each player gets truthful messages through the soft links from the players sufficiently close in their preferences, truthful messages through the hard links from the less close players, and no messages from the more distinct players. This equilibrium generates the greatest total welfare, because, first, the equilibrium construction guarantees the minimal possible number of the costly hard links across all equilibria. And second, the equilibrium network provides the maximal level of individual informativeness which positively influences welfare, because, by Lemma 1, the ex-ante expected individual payoffs increase with the in-degree of each player.

An additional comment regarding the welfare can be made in the pure cheap talk setting: all maximal equilibria generate the same ex-ante expected individual payoffs (and, as a consequence, the same total welfare) that are the greatest across all equilibria.[17] Indeed, provided that the cost

[16] This statement cannot be extended to a per-individual basis because costly hard links can be distributed differently in various maximal and pairwise stable equilibria.

[17] Given this welfare property, the communication network of any maximal equilibrium corresponds to the utility-maximizing equilibrium network studied in [Galeotti *et al.*, 2011].

of a soft link ε is infinitesimal, the ex-ante expected payoff of each individual depends only on the vector of in-degrees; hence, the expected individual payoffs are the same across all maximal equilibria. These results are presented in the lemma below.

Lemma 4. *There exists a pairwise stable equilibrium that generates the greatest total welfare across all equilibria. In the pure cheap talk setting, all maximal equilibria generate the same ex-ante expected individual payoffs that are the greatest across all equilibria.*

Interested readers are referred to Appendix 1.8 for an additional discussion of pairwise stable equilibria.

1.5.2 Example with three players

Let the prior distribution of θ be uniform on the interval $[0, 1]$ and consider three players with the preference biases $b_1 = 0$, $b_2 = b$, $b_3 = 2b$, where $\frac{1}{10} < b \leq \frac{1}{8}$. Such preference structure implies that truth-telling through a soft link ij is incentive compatible for player i if and only if $|b_i - b_j| = b$ and nobody else reports to j truthfully. If hard links are prohibitively costly, $\max\{\alpha, 1-\alpha\}C > h(0) - h(1)$, then there are two pairwise stable pure cheap talk equilibria that generate the following truthful networks (see Figure 1.1, where dashed lines depict soft links):

(i) $g_{21} = g_{23} = g_{12} = s$, $g_{13} = g_{31} = g_{32} = 0$,

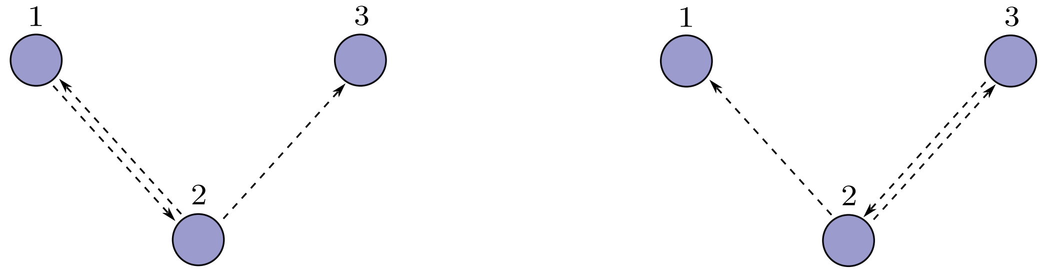
(ii) $g_{21} = g_{23} = g_{32} = s$, $g_{13} = g_{31} = g_{12} = 0$.

Note, that players have the same in-degrees of 1 and the same ex-ante expected payoffs across the pairwise stable equilibria. Consider the equilibrium corresponding to the soft-link truthful network (i), denoted by g^s. Using Lemma 1 and ignoring the infinitesimal costs of the soft links, the ex-ante expected payoff of player i, denoted as $\mathbb{E}u_i(g^s)$, is:

$$\mathbb{E}u_i(g^s) = -\sum_{j=1}^{3}[h(1) + (b_j - b_i)^2] = -3h(1) - B_i, \quad i = 1, 2, 3,$$

where B_i depends only on divergence in preferences, $B_i = \sum_{j=1}^{3}(b_j - b_i)^2$.

Suppose that the cost of a hard link satisfies $\max\{\alpha, 1-\alpha\}C \in (h(1) - h(2), h(0) - h(1)]$, i.e., players i and j prefer to form a hard link ij if the in-degree of player j is 0. Clearly, in this case, the set of the pairwise stable equilibria is the same as in the pure cheap talk setting above.



(i) $g_{21} = g_{23} = g_{12} = s$, $g_{13} = g_{31} = g_{32} = 0$ (ii) $g_{21} = g_{23} = g_{32} = s$, $g_{13} = g_{31} = g_{12} = 0$

Figure 1.1: Communication networks of pairwise stable pure cheap talk equilibria.

Now assume that the cost is decreased further: $\max\{\alpha, 1-\alpha\}C \in (0, h(1) - h(2)]$, i.e., i and j prefer to introduce a hard link ij if the in-degree of player j is $k_j \leq 1$. Consequently, the in-degree of any player in any equilibrium network must be 2. Because player i cannot be credible in cheap talk communication to player j with $k_j = 2$, no soft links can be a part of an equilibrium network. As a result, the only equilibrium (which is also pairwise stable) has a complete communication network consisting of hard links, which I denote by g^h (see Figure 1.2, where solid lines depict hard links).

This analysis shows how soft links are substituted with costly hard links, as the cost decreases.[18] Regarding the welfare effect, there are two forces. On the one hand, crowding out of cheap communication by costly verifiable communication is welfare decreasing. On the other hand, players accumulate more information, which is welfare increasing. Below I show how the resulting impact depends on the cost structure and the cost level.

In the communication network g^h, each person has the in-degree of 2 and supports 4 hard links—2 incoming and 2 outgoing—which implies that each player faces the cost of $2C$ (see Figure 2). The ex-ante expected payoff of agent i then is

$$\mathbb{E}u_i(g^h, C) = -3h(2) - 2C - B_i, \quad i = 1, 2, 3.$$

[18] Consider, for example, the effect of introducing a hard link 31 in the soft-link network g^s. Once the hard link 31 is formed, then player 2 is no longer credible in reporting to player 1 via a soft link. This forces player 2 and player 1 to substitute a soft link 21 with a costly hard link.

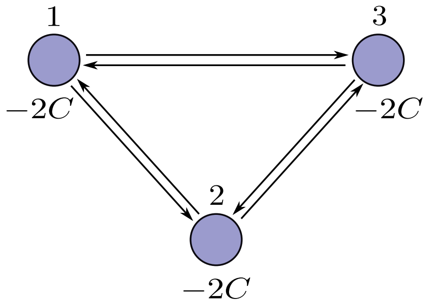


Figure 1.2: Equilibrium network when $\max\{\alpha, 1-\alpha\}C \in (0, h(1)-h(2)]$.

Only one party bears the cost of a hard link. If only one party faces the cost of a hard link, then the difference in the ex-ante expected payoffs corresponding to g^h and g^s is

$$\mathbb{E}u_i(g^h, C) - \mathbb{E}u_i(g^s) = 2\left(\frac{3}{2}(h(1)-h(2)) - C\right) > 0, \quad i = 1, 2, 3,$$

because $C \leq h(1)-h(2)$. This means, that the positive informational effect dominates the negative crowding-out effect, and the pairwise stable equilibrium with the hard links generates higher ex-ante expected individual payoffs (as well as higher total welfare) than the pure cheap talk pairwise stable equilibrium. Clearly, the ex-ante expected individual payoffs (and the total welfare) in a pairwise stable equilibrium weakly increase as C decreases: they are flat when $C > h(1)-h(2)$ (the pure cheap talk equilibrium) and strictly increase when $C \leq h(1)-h(2)$ (the hard-link equilibrium).

The sender and the receiver share the cost of a hard link. When the cost of a hard link is divided between the parties, introducing costly hard links might harm the ex-ante expected individual payoffs. In particular, if $\alpha \in (\frac{1}{3}, \frac{2}{3})$, the difference $\mathbb{E}u_i(g^h, C) - \mathbb{E}u_i(g^s)$ is strictly negative whenever $C \in \left(\frac{3(h(1)-h(2))}{2}, \frac{h(1)-h(2)}{\max\{\alpha,1-\alpha\}}\right]$. In this case, the crowding-out effect dominates the information improvement effect, and the unique hard-link equilibrium generates strictly lower ex-ante expected individual payoffs compared to the pairwise stable soft-link equilibrium. Intuitively, compared to the previous case in which one party internalizes the entire cost of a hard link, the parties now share the cost of a hard link while enjoying the same benefit. This leads to the formation of too many hard links relative to the cost C.[19] If, however, the cost is sufficiently low,

[19] Note that no hard links are present in a pairwise stable equilibrium when only one party bears the cost $C \geq \frac{3(h(1)-h(2))}{2}$.

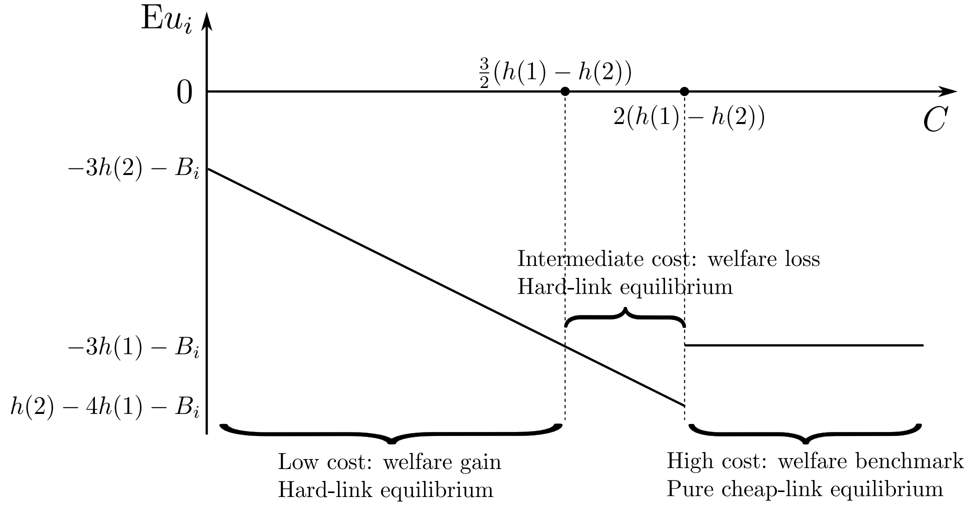


Figure 1.3: Ex-ante expected payoff of player i, $\mathbb{E}u_i$, in a pairwise stable equilibrium when $\alpha = \frac{1}{2}$.

$C \in \left(0, \frac{3(h(1)-h(2))}{2}\right]$, then the hard-link equilibrium generates greater individual ex-ante expected payoffs than any pairwise stable soft-link equilibrium. As a result, the ex-ante expected individual payoffs (and the total welfare) are non-monotonic in the cost C: Figure 1.3 depicts $\mathbb{E}u_i$ in a pairwise stable equilibrium when the cost C is divided equally between the parties, $\alpha = \frac{1}{2}$.

1.5.3 One party bears the cost of a hard link

Consider the setting in which the cost of a hard link C accrues to only one party, either the one sending the message or receiving it. The following theorem states that allowing for an additional means of communication via hard links can only improve the total welfare in a pairwise stable equilibrium.

Theorem 3. *Take any cost $C_0 > h(0) - h(1)$ and consider some pure cheap talk pairwise stable equilibrium with the communication network $g(C_0)$ and the total welfare $W(g(C_0))$. Then for any cost $C_1 \leq h(0) - h(1)$, there exists a pairwise stable equilibrium with the communication network $g(C_1)$ such that the total welfare $W(g(C_1), C_1) \geq W(g(C_0))$.*[20]

Proof. See Appendix 1.9. □

To understand what drives the result, renumber the players such that their in-degrees in $g(C_0)$

[20] Notational comment: the cost enters only the welfare $W(g(C_1), C_1)$ and is omitted from $W(g(C_0))$ to signify that, under the cost of C_0, a pairwise stable equilibrium has no hard links.

are increasing, $k_1 \leq k_2 \leq ... \leq k_n$, and consider a particular case of the cost $C_1 \in (h(k+1) - h(k+2), h(k) - h(k+1)]$ for some k, $k_1 \leq k < k_n$. Clearly, there exists player j such that $k_j < k+1 \leq k_{j+1}$. This means that under the cost of C_1, there are incentives to form hard links towards the first j players, whose in-degrees are less than $k+1$, until their in-degrees become equal to $k+1$. Thus, the set of new maximal in-degrees is:

$$k'_1 = ... = k'_j = k+1,\, k'_{j+1} = k_{j+1},\, ...,\, k'_n = k_n.$$

Note that because of the congestion effect, the new hard links towards players $1, ..., j$ might crowd out some soft links. In the worst case, all soft links directed to players $1, ..., j$ are substituted by the hard links. This corresponds to a maximal equilibrium with the communication network $g(C_1)$, in which all links directed towards players $1, ..., j$ are hard, while all links directed to players $j+1, ..., n$ are soft and the same as in $g(C_0)$. To compare the total welfare in $g(C_0)$ and $g(C_1)$, consider the total cost of the newly introduced hard links and the additional ex-ante expected payoff arising from information improvement. The total cost of the hard links amounts to

$$j(k+1)C_1 \leq j(k+1)(h(k) - h(k+1)).$$

The gain in the total welfare compared to the pure cheap talk case is

$$n \sum_{i=1}^{j} (h(k_i) - h(k+1)) \geq n \cdot j(h(k) - h(k+1)).$$

The lower bound for the welfare gain strictly exceeds the upper bound for the cost, because $n-1 \geq k_n > k$. Intuitively, while $k+1$ hard links are used to increase the in-degree of player $l \in \{1, ..., j\}$ by at least 1, all n players enjoy the additional benefit arising from a greater accuracy of l's action. Hence, for this maximal equilibrium, $W(g(C_1), C_1) > W(g(C_0))$. Next, Lemma 4 insures that there exists a pairwise stable equilibrium that achieves the total welfare of at least $W(g(C_1), C_1)$, which implies the result.

From this example, it becomes apparent that while the total welfare in a pairwise stable equilibrium goes up, individual welfare might go down. In particular, assume that the receiver bears the cost of a hard link and the maximal equilibrium considered above with the communication network

$g(C_1)$ is actually pairwise stable.[21] Then it might be the case that player j would prefer the pure cheap talk equilibrium with the communication network $g(C_0)$ to the pairwise stable equilibrium with $g(C_1)$ because the expected gain in his payoff is dominated by the high cost of maintaining $k+1$ links. This point is emphasized in the following remark.

Remark. The positive welfare result of Theorem 3 does not extend to a per-individual basis. Specifically, a pairwise stable equilibrium with hard links might generate a lower welfare for some player compared to the cheap talk case, if he ends up sustaining too many costly hard links relative to the individual gains from the information improvement.

1.5.4 Two parties bear the cost of a hard link

Assume that the cost of a hard link C is distributed between the sender and the receiver with the shares α and $1-\alpha$, respectively. A previously considered example with 3 players revealed that the welfare might go down when the cost C is decreased, so that hard links appear in a pairwise stable equilibrium. In this section, I follow [Galeotti *et al.*, 2011] and [Hagenbach and Koessler, 2010] in considering the natural cases of the two communities and a diverse group of people, and describe conditions under which the total welfare decreases (or increases) when verifiable information transmission becomes feasible.

The welfare outcome depends on which effect—the positive information improvement or the negative crowding out—dominates. Clearly, if no crowding out of cheap communication occurs, then the total welfare necessarily goes up, because the benefit of an extra hard link enters the payoffs of more than 2 players, which strictly outweighs the total cost of the link, C. A more general version of this intuitive result, which is useful for further analysis, is presented in the following lemma:

Lemma 5. *Consider two cost levels, $C_0 \geq C_1$. For each C_i, fix some equilibrium and consider the corresponding communication network $g(C_i) = g^s(C_i) \cup g^h(C_i)$, where $g^s(C_i)$ is the set of soft links of $g(C_i)$ and $g^h(C_i)$ is the set of hard links of $g(C_i)$. If $g^s(C_0) \subseteq g^s(C_1)$ and $g^h(C_0) \subseteq g^h(C_1)$, then $W(g(C_1), C_1) \geq W(g(C_0), C_0)$, where $W(g(C_i), C_i)$ is the total welfare corresponding to the*

[21] The examples of such pairwise stable equilibria are presented in the two communities setting (with $n_1 = 1$) of the next subsection.

equilibrium network $g(C_i)$ and the cost C_i. Moreover, the same welfare implications hold on a per-individual basis.

Proof. See Appendix 1.9. □

Excessive formation of hard links. The result that the total welfare can decrease when verifiable information transmission becomes feasible hinges on the excessive formation of hard links (and hence, extensive crowding out of soft links) if both parties bear the cost. When $\alpha \in (0, 1)$, the individuals fail to account for the total cost of the hard link, which results in stronger incentives to form hard links than in the case in which one party faces the full cost.[22] For instance, assume that the cost C is such that hard links don't appear when $\alpha \in \{0, 1\}$, but necessarily emerge for some $\alpha \in (0, 1)$. Then as the examples in this chapter illustrate, having the cost of a hard link shared between the parties can be both beneficial and detrimental to the total welfare compared to the case in which one party bears the full cost.

Two communities. The set of players consists of two groups, N_1 and N_2, with sizes n_1 and n_2, respectively, where $1 \leq n_1 < n_2$ and the total number of players is $n = n_1 + n_2$. Each member of group N_1 has a preference bias normalized to 0, while each member of N_2 has a bias b.

In any pairwise stable equilibrium, there is complete communication via cheap talk inside each group, because the incentive compatibility constraint of truth-telling is always satisfied for the players who share the same preferences. In addition, all players in the same group receive the same number of truthful messages, because exactly the same people can report truthfully to them through links of the same type. I introduce notation similar to [Galeotti *et al.*, 2011]: denote by k_i the maximal in-degree of an arbitrary player in group N_i. Further, $k_i = k_{ii} + k_{ij}$, where $k_{ii} = n_i - 1$ reflects the level of *intra-group communication*—the number of truthful messages that a player from N_i receives from the members of the same group, and k_{ij} stands for the level of *cross-group communication*—the number of truthful messages that a player from N_i receives from the members of the opposite community N_j.

Consider the pure cheap talk case that corresponds to a prohibitively high cost $C_0 > \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$. As mentioned above, in the communication network of any pairwise stable equilibrium the level

[22] Indeed, the incentives to form hard links are determined by the maximum individual cost of a hard link $\max\{\alpha, 1-\alpha\}C$ that is lower than the total cost C.

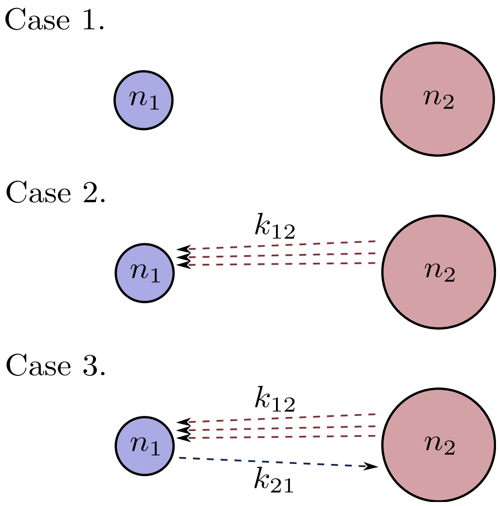


Figure 1.4: Communication networks of pairwise stable pure cheap talk equilibria.

of intra-group communication is $k_{ii} = n_i - 1$. Regarding cross-group communication, note that if members of a smaller group N_1 report truthfully to some members of N_2, then by the congestion effect, the pairwise stability implies that members of a larger group N_2 report truthfully to some members of N_1. Thus, depending on the parameters, cross-group communication can take one of the following three forms (see Figure 1.4):

1. No cross-group communication, i.e., $k_{21} = k_{12} = 0$.

2. Communication from group N_2 to group N_1, i.e., $k_{12} > 0,\ k_{21} = 0$.

3. Cross-group communication, i.e., $k_{12} > 0,\ k_{21} > 0$.

Assume now, that the cost is reduced to $C_1 > 0$, so that hard links become feasible. While complete soft-link intra-group communication is still a part of the truthful network of any pairwise stable equilibrium, cross-group communication can change due to the emergence of cross-group hard links. Because there are only two types of players, pairwise stability requires that members of the same group receive identical numbers of truthful cheap talk messages and verifiable information messages across different equilibrium networks. This, in turn, means that the total welfare is the same across all pairwise stable equilibria. Whether feasibility of hard links results in a welfare gain or a welfare loss depends on how the informational benefit from hard links compares with the loss from crowding out soft links.

In particular, it can be easily seen that introducing feasible hard links in case 1 necessarily improves the welfare because no soft links are crowded out in a pairwise stable equilibrium. As the proof of Theorem 4 demonstrates, the same positive welfare result holds in case 3, even though introducing hard links crowds out some soft cross-group communication. The reason is that the feasibility of hard links in case 3 increases the in-degrees of all players, which outweighs the cost of hard links and the crowding out effect. The negative welfare result arises in case 2 for some parameters, the reason being that the feasibility of hard links increases the in-degrees of only group N_1 members, which is dominated by the negative crowding out effect. The following theorem describes the necessary and sufficient conditions for the welfare decrease.

Theorem 4. *Take any cost* $C_0 > \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$ *and consider some pure cheap talk pairwise stable equilibrium with the communication network* $g(C_0)$ *and the total welfare* $W(g(C_0))$*. Introduce the possibility to form hard links with the cost* C_1 *and consider some pairwise stable equilibrium with the communication network* $g(C_1)$ *and the total welfare* $W(g(C_1), C_1)$*. There exists a non-degenerate set* $(\underline{C}_1, \overline{C}_1)$ *of* C_1 *such that* $W(g(C_1), C_1) < W(g(C_0))$ *if and only if the preferential difference* b *satisfies*

$$b \in \left(\frac{1}{2(\alpha+\beta+n_1+k+1)}, \frac{1}{2(\alpha+\beta+n_1+k)} \right],$$

for some k*, where* $\max\{\alpha, 1-\alpha\}n - 1 < k \leq n_2 - n_1 - 1$*. Otherwise, for all* C_1*:* $W(g(C_1), C_1) \geq W(g(C_0))$*.*

Proof. See Appendix 1.9. □

To gain intuition for why introducing feasible hard links can lead to a welfare decrease in case 2, note that for the negative welfare result to occur, it must be that the two communities are sufficiently unbalanced in their size: $n_1 < \frac{n_2}{3}$. In addition, the in-degree of each N_1 member in a pure cheap talk equilibrium must be sufficiently high: $k_1 > \max\{\alpha, 1-\alpha\}n-1 \geq \frac{n}{2}-1$. Assume that hard links become feasible with the cost C_1 being such that the in-degrees of players in group N_1 increase by just 1, while no hard links from N_1 to N_2 appear. Since in the pure cheap talk equilibrium members of N_1 already receive relatively many signals, increasing their informativeness by 1 signal leads to a moderate additional individual benefit. Given that the size of community N_1 is relatively small, $n_1 < \frac{n}{4}$, this sums up to a moderate increase in the total welfare. At the same time, all cross-group soft links are crowded out and substituted by hard links, which amounts to a considerable cost, given that the level of cross-group communication was $k_{12} > \max\{\alpha, 1-\alpha\}n - 1 \geq \frac{n}{2} - 1$. As a result, the net welfare effect is negative.

The structure of pairwise stable equilibria allows to make several observations regarding communication patterns. In particular, individuals with similar preferences can easily communicate with each other via cheap talk, which leads to complete soft intra-group communication that is robust to introducing feasible hard links. On the contrary, soft-link cross-group communication is less intensive, with the information flow being greater towards the smaller group, and is vulnerable to the appearance of verifiable communication (soft cross-group communication can be easily crowded out with costly verifiable information transmission). This allows one to expect the mode of communication between individuals with different characteristics to be more substantial and proof-oriented than between similar individuals.

Diverse group of people. In this setting, players $1, 2, ..., n$ have equidistant biases that satisfy $b_1 = 0$ and $b_{i+1} - b_i = b > 0$, $i = 1, ..., n-1$. Assume that for some $C > 0$ the maximal in-degrees are $k_1, ..., k_n$, and construct a pairwise stable equilibrium that generates the greatest welfare in a way described in Lemma 4. In the communication network of this equilibrium, player i gets truthful messages through soft links from players with sufficiently close preferences, truthful messages via hard links from more distinct players, and no messages from those who are further away in their preferences. In case of prohibitively costly hard links, $\max\{\alpha, 1-\alpha\}C_0 > h(0) - h(1)$, truthful reporting to player i boils down to the k_i closest players revealing their information to i truthfully via soft links.

As was demonstrated in the example of three players, for some parameters, introducing feasible hard links necessarily harms the total welfare in the pairwise stable equilibrium. The following theorem states that this is no longer the case for bigger groups: if $n \geq 4$, then for any difference in preferences b and any cost $C_1 < \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$, there is a pairwise stable equilibrium that generates a greater total welfare than in the pure cheap talk case.

Theorem 5. *Let $C_0 > \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$ and consider some pure cheap talk pairwise stable equilibrium with the communication network $g(C_0)$ and the total welfare $W(g(C_0))$. Introduce feasible hard links with the cost of $C_1 \leq \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$ and consider a pairwise stable equilibrium with the communication network $g(C_1)$ that generates the greatest total welfare $W(g(C_1), C_1)$. There exists a non-degenerate* *set* $(\underline{C}_1, \overline{C}_1)$ *of* C_1 *such that* $W(g(C_1), C_1) < W(g(C_0))$ *if and only if* $n = 3$, $b \in \left(\frac{1}{10}, \frac{1}{8}\right]$ *and* $\alpha \in (\frac{1}{3}, \frac{2}{3})$. *Otherwise, for all* C_1*:* $W(g(C_1), C_1) \geq W(g(C_0))$.

Proof. See Appendix 1.9. □

1.6 Extensions

1.6.1 Endogenous costs of hard links

One of the possible extensions of the model is to allow the parties to negotiate the way they split the cost of a hard link C each time they create a link (e.g., how to share the traveling cost). Surprisingly, such endogeneity of share α, although implying pairwise efficiency, does not imply aggregate efficiency, and introducing hard links can still lead to lower total welfare.

To see this, consider players i and j such that player i is not credible in reporting to player j (with the in-degree k_j) via cheap talk. The aggregate benefit of players i and j from creating a hard link ij is $2(h(k_j) - h(k_j + 1))$, while the total cost of a hard link is C. Thus, as long as the benefit exceeds the cost, $2(h(k_j) - h(k_j + 1)) > C$, the players can split the cost of a hard link so that they both strictly benefit from its creation. In particular, the players can always share C equally between each other and enjoy the additional benefit of $h(k_j) - h(k_j + 1) - C/2$ each.

In general, the players can find a way to profitably split the cost C and introduce the hard link ij if and only if the hard link ij is desired by the players when the cost C is split equally between the parties, $\alpha = 1/2$. This observation implies that communication patterns (and hence, the total welfare) of the pairwise stable equilibria are the same across the two settings: (i) the share α is endogenous, and (ii) $\alpha = 1/2$. Because in the latter case of $\alpha = 1/2$ introducing hard links can deteriorate the total welfare, it can do so in the case of endogenous α as well.

1.6.2 Heterogeneous costs of hard links

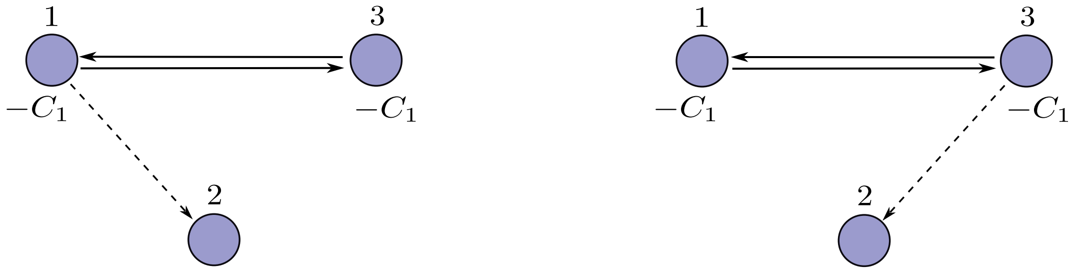
While in some situations a natural assumption about the hard links is that the cost value is similar across players, it might be not quite adequate in others. Thus, a straightforward extension of the model would be to allow the cost of a hard link to depend on the players' identities. As one example of cost heterogeneity, consider managers of an international corporation who might be located in different cities or countries, and who need to communicate with each other before making individual decisions for their respective divisions. A hard link might correspond to a personal meeting, while a soft link corresponds to a phone call or an email. In this case, it is relatively easy for two people from the same location to meet, while a personal meeting of two people from different countries entails additional time and spending.

In this section, I highlight some insights stemming from the cost heterogeneity assumption in an extreme setting, where, for some pairs of agents, it is easy to communicate in verifiable way, while others face a considerable cost. More formally, the cost depends on the pair of agents, i and j, and is the same independent of the link direction: the cost of a hard link ij is equal to the cost of a hard link ji, $C_{ij} = C_{ji} = C$. The cost can be either prohibitively high, $C_0 > \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$, or feasibly low, $C_1 \leq \frac{h(0)-h(1)}{\max\{\alpha,1-\alpha\}}$. For simplicity, I maintain the assumption that the cost structure that is specified by the share $\alpha \in [0,1]$ is the same across all pairs of agents. In some sense, the setting with heterogeneous costs is in between the two cases of homogeneous cost values C_0 and C_1. In particular, switching from the case of homogeneous cost C_0 to heterogeneous costs can be viewed as lowering the cost to the level of C_1 for *some* pairs; while switching to homogeneous cost C_1 corresponds to lowering the cost values to C_1 for *all* pairs.

Consider a particular case where C_1 is sufficiently small: $C_1 \leq \frac{h(n-2)-h(n-1)}{\max\{\alpha,1-\alpha\}}$. Clearly, in any equilibrium, agents in pairs with the cost C_1 must communicate truthfully with each other. This *localization* of communication might change the truthful communication network compared to the pure cheap talk case, while not necessarily leading to an information improvement. Indeed, consider the previously studied example of three players, where the prior distribution of θ is uniform $[0,1]$ and the preference biases are $b_1 = 0$, $b_2 = b$, $b_3 = 2b$, such that $\frac{1}{10} < b \leq \frac{1}{8}$. Consider the case of prohibitively costly hard links—homogeneous cost C_0. A considerable difference in preferences prevents players 1 and 3 from truthful communication with each other via cheap talk. The two pairwise stable pure cheap talk equilibria have the in-degrees $k_1 = k_2 = k_3 = 1$ and the communication networks described in Figure 1.1.

Assume now that the cost of a hard link for players 1 and 3 is decreased to the level of C_1, implying the following heterogenous costs setting: $C_{12} = C_{23} = C_0$, $C_{13} = C_1$. In this case, players 1 and 3 must communicate with each other via hard links, thus, the only two pairwise stable equilibria have the following communication networks (see Figure 1.5):

(i) $g_{12} = s$, $g_{13} = g_{31} = h$, $g_{21} = g_{23} = g_{32} = 0$.



(i) $g_{12} = s$, $g_{13} = g_{31} = h$, $g_{21} = g_{23} = g_{32} = 0$

(ii) $g_{32} = s$, $g_{13} = g_{31} = h$, $g_{21} = g_{23} = g_{12} = 0$

Figure 1.5: Communication networks of pairwise stable equilibria: heterogenous costs.

(ii) $g_{32} = s$, $g_{13} = g_{31} = h$, $g_{21} = g_{23} = g_{12} = 0$.

Note that the in-degrees of the players are the same as in the pure cheap talk case, $k_i' = 1$, $i = 1, 2, 3$, while some links are hard.

The fact that introducing heterogeneous costs creates localization of communication that might fail to generate an informational gain implies that the ex-ante expected total welfare can decrease even when only one party bears the cost of a hard link. This contrasts with the case of homogeneous cost, in which introducing feasible hard links necessarily leads to a welfare improvement when only one party faces the cost. Thus, heterogeneous costs can result in a welfare lower than in both cases of homogeneous cost, C_0 and C_1.

1.7 Conclusion

In this chapter I study the communication patterns and the welfare outcomes when agents are allowed to choose whether to communicate via cheap talk or through a verifiable information channel. Cheap talk is performed through soft links, while verifiable information transmission occurs via costly hard links. The way the cost of a hard link is shared between the parties—the cost structure—is determined by the respective burdens of providing hard evidence and decoding the underlying private information. In the main part of the chapter, the costs of incoming and outgoing hard links are the same for all agents.

While the truthful revelation of a signal is always beneficial to both parties from the ex-ante perspective, there is a credibility issue in cheap talk communication at the interim stage. In particular, I find the same congestion effect as in [Galeotti *et al.*, 2011], i.e., the willingness of player i to report truthfully to player j decreases with the preference divergence and with the number of truthful messages reported to player j—the in-degree of player j, k_j. In the case of verifiable information transmission, the incentive to form a hard link ij does not depend on the preference divergence; rather, it depends only on the in-degree of player j and is strictly decreasing in k_j. The straightforward equilibrium implication for the network structure is that hard links must be directed towards players who receive relatively small numbers of truthful messages via soft links.

The first—and intuitive—result states that a positive informational effect arises from introducing a feasible verifiable communication channel, independently of the cost structure. That is, if a pure cheap talk equilibrium fails to exist when hard links become feasible, then there exists an equilibrium with hard links, in which every player accumulates a weakly greater number of signals.

The second—and the main—result concerns the welfare implications of introducing feasible verifiable communication. The appearance of hard links in the pure cheap talk setting has two effects on the expected total welfare: a positive effect stems from the information improvement and a negative effect arises from crowding out soft communication with costly verifiable communication. The crowding out effect—that is at the heart of this study—arises because newly formed hard links increase players' in-degrees, which by the congestion result, destroys the credibility of communication through some soft links. As a result, for the players to transfer their signals truthfully, those soft links should be replaced by costly hard links.

I show that the positive informational effect always dominates the negative crowding out effect when only one party bears the cost of a hard link. This positive welfare result no longer holds when the cost of a hard link is shared between the players. I illustrate this point by means of two examples. The first example is a two communities setting, in which introducing feasible hard links can wipe out all cross-group soft communication, resulting in a welfare decrease compared with the pure cheap-talk case. The second example is the case of a diverse group of people with equidistant biases, in which adding the possibility to form hard links can decrease the expected total welfare when the number of players is 3.

In one extension, I allow the parties to negotiate how to split the cost of a hard link between them. I show that such endogeneity of the cost shares does not necessarily lead to aggregate efficiency; introducing hard links can still decrease the total welfare. In another extension, I allow the costs of hard links to differ across the pairs of players. When the cost difference is substantial, I show that the availability of hard links is likely to result in the localization of communication with respect to the low cost of hard links, with the information accumulated by every player being the same. As a result, expected total welfare can decrease even when only one party bears the cost.

For further analysis, it would be interesting to study the communication pattern under a different timing: the decision to form the links is made *after* the signals' realization. In this setting, the mere fact that player i wants to form a specific link to player j can reveal some information about i's signal. Another modification would be to consider dynamic communication, e.g., soft communication happens first, then players form a verifiable communication network and reveal their signals accordingly. This setting allows agents to strategically follow up soft communication with costly verifiable information transmission.

1.8 Appendix: Pairwise stable equilibria

An equilibrium $\{g, (\mu^g, y^g)\}$ is *pairwise stable*, if no pair of players can change the communication pattern between them to improve their ex-ante expected utilities, while satisfying the interim incentive compatibility constraints of truth-telling, holding other strategies fixed. To better understand how the communication pattern can be improved, consider three possible cases of information transmission from player i to some player j with in-degree k_j in g:

1. Player i reports truthfully to player j through a soft link: $g_{ij} = s$. There is no way to improve communication: making communication uninformative will result in the ex-ante expected loss of $h(k_j - 1) - h(k_j)$, while switching to verifiable communication will involve a cost.

2. Player i reports truthfully to player j through a hard link: $g_{ij} = h$. Because the hard link ij is a part of the equilibrium, it is preferred to no link by both players, and deleting this link will not improve the ex-ante expected payoffs. It may, however, be possible to reach an improvement by substituting the costly hard link ij with a soft link ij and inducing truthful communication through it. This option can be realized if player i can credibly communicate given that j believes i's message. Otherwise, it is not possible to change the communication pattern in a profitable direction.

3. Player i does not report informatively to player j, $g_{ij} = 0$. Here it may be possible to improve by creating a soft link ij with truthful communication, if it is interim incentive compatible. If not, then the second-best option is creating a hard link ij. If the latter option is undesired by at least one player, then there exists no possibility to improve.

Given these alternatives, pairwise stability can be formally defined as follows:

Definition. An equilibrium $\{g, (\mu^g, y^g)\}$ is *pairwise stable* if

(i) For any $i, j \in \{1, .., n\}$, $g_{ij} = 0$ only if, holding other strategies fixed, i cannot credibly report to j via a soft link, assuming that j believes i's message, and communication via a hard link is not desired by at least one party.

(ii) For any $i, j \in \{1, .., n\}$, $g_{ij} = h$ only if, holding other strategies fixed, i cannot credibly report to j via a soft link, assuming that j believes i's message.

Symmetry. An immediate property of a pairwise stable equilibrium is that any two players i and j with the same preference biases, $b_i = b_j$, must be treated in symmetric way, i.e., they communicate truthfully with each other via cheap talk and receive the same number of truthful messages from other players. Moreover, if some other player l truthfully reports to both, i and j, then it must be the case that l uses the same type of information transmission channel.

Maximality and pairwise stability. There might be multiple maximal equilibria generating the same vector of maximal in-degrees, among which some maximal equilibria might not be pairwise stable.

Example. Here I present an example of a maximal equilibrium which is not pairwise stable. Let the prior distribution of θ be uniform on $[0, 1]$. Consider 3 players with the preference biases $b_1 = b_2 = 0$, $b_3 \in (\frac{1}{10}, \frac{1}{8}]$. Assume that hard links are prohibitively costly. Then there are several maximal pure cheap talk equilibria that generate the in-degrees $k_1 = k_2 = k_3 = 1$. The examples of such communication networks are: (1) $g_{12} = g_{21} = g_{13} = s$, $g_{23} = g_{31} = g_{32} = 0$, and (2) $g_{12} = g_{23} = g_{31} = s$, $g_{13} = g_{32} = g_{21} = 0$. It is easy to see that the first communication network corresponds to a pairwise stable equilibrium. In contrast, the second communication network can not correspond to a pairwise stable equilibrium, because players 1 and 2, who agree in their preferences, would deviate and induce truthful communication through the soft link 21.

Verifiable information and maximal in-degrees. The informational result, coupled with the fact that the incentive to form a hard link decreases with the player's in-degree, ensure that allowing for verifiable information transmission results in weakly greater, and more evenly distributed, maximal in-degrees.

Strong stability. As with pairwise stability, one can adapt from the networks literature the notion of *strong stability* to this framework: coordinated change of communication pattern in a group of agents cannot strictly improve the welfare of some members, while weakly improving the welfare of others. Then a pairwise stable equilibrium generates the greatest total welfare if and only if it is strongly stable.[23]

[23] This statement is non-trivial, because there exist pairwise stable equilibria that do not generate the greatest total welfare. For example, consider 3 players with the preference biases $b_1 = 0$, $b_2 = b$, $b_3 = 2b$, where $b \in (\frac{1}{10}, \frac{1}{8}]$,

Proof. The "only if" part of the statement is straightforward. Concerning the "if" part, note that if a strongly stable equilibrium—which is necessarily maximal—didn't generate the greatest total welfare, then there must be a hard link that can be severed and a soft link (with the truthful reporting through it) that can be introduced. This contradicts the property of strong stability. □

1.9 Appendix: Proofs

Proof of Lemma 1. Because the chosen action y_{s_R} is given by $y_{s_R} = b_j + \mathbb{E}(\theta|s_R)$ and $f(\theta, s_R) = f(\theta|s_R)P(s_R)$, the expected input from player j into i's payoff becomes

$$\begin{aligned}
h(b_i, b_j, k_j) &= -\int_0^1 \sum_{s_R \in \{0,1\}^{k_j+1}} (b_j + \mathbb{E}(\theta|s_R) - \theta - b_i)^2 f(\theta, s_R) d\theta \\
&= -\sum_{s_R \in \{0,1\}^{k_j+1}} \int_0^1 \left((\mathbb{E}(\theta|s_R) - \theta)^2 + 2(b_j - b_i)(\mathbb{E}(\theta|s_R) - \theta) + (b_j - b_i)^2 \right) f(\theta|s_R) P(s_R) d\theta \\
&= -\sum_{s_R \in \{0,1\}^{k_j+1}} \left[\int_0^1 (\mathbb{E}(\theta|s_R) - \theta)^2 f(\theta|s_R) d\theta \right] P(s_R) - (b_j - b_i)^2 \\
&= -\mathbb{E}\left[\mathrm{Var}(\theta|s_R)\right] - (b_j - b_i)^2.
\end{aligned}$$

In what follows, I show that $\mathbb{E}\left[\mathrm{Var}(\theta|s_R)\right]$ is exactly $h(k_j)$. Let l denote the number of signals 1 in s_R, then

$$\begin{aligned}
\mathrm{Var}(\theta|s_R) &= \mathbb{E}(\theta^2|s_R) - (\mathbb{E}(\theta|s_R))^2 \\
&= \int_0^1 \theta^2 f(\theta|l, k) d\theta - \left(\frac{\alpha + l}{\alpha + \beta + k_j + 1} \right)^2 \\
&= \frac{B(\alpha + l + 2, \beta + k_j + 1 - l)}{B(\alpha + l, \beta + k_j + 1 - l)} - \left(\frac{\alpha + l}{\alpha + \beta + k_j + 1} \right)^2 \\
&= \frac{(\alpha + l)(\beta + k_j + 1 - l)}{(\alpha + \beta + k_j + 1)^2 (\alpha + \beta + k_j + 2)}.
\end{aligned}$$

and assume that θ is uniform on $[0, 1]$ (the same setup as in Subsection 1.5.2). If the cost of a hard link C satisfies $\max\{\alpha, 1 - \alpha\}C \in (h(1) - h(2), h(0) - h(1)]$, then the greatest total welfare is generated, for example, in a pairwise stable equilibrium with the following communication network: $g_{21} = g_{23} = g_{12} = s$, $g_{13} = g_{31} = g_{32} = 0$. However, there exists a pairwise stable equilibrium with the communication network $g_{21} = g_{12} = s$, $g_{13} = h$, $g_{23} = g_{31} = g_{32} = 0$ that achieves a strictly lower total welfare.

Using this, $\mathbb{E}\left[\mathrm{Var}(\theta|s_R)\right]$ becomes

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}(\theta|s_R)\right] &= \sum_{s_R\in\{0,1\}^{k_j+1}} \frac{(\alpha+l)(\beta+k_j+1-l)}{(\alpha+\beta+k_j+1)^2(\alpha+\beta+k_j+2)}P(s_R)\\ &= \frac{1}{(\alpha+\beta+k_j+1)^2(\alpha+\beta+k_j+2)}A,\end{aligned}$$

where

$$A=\alpha(\beta+k_j+1)+(\beta+k_j+1-\alpha)\sum_{s_R\in\{0,1\}^{k_j+1}} lP(s_R)-\sum_{s_R\in\{0,1\}^{k_j+1}} l^2P(s_R).$$

Here $\sum_{s_R\in\{0,1\}^{k_j+1}} lP(s_R)=(k_j+1)\mathbb{E}(s_1)$, because l is the sum of signals in s_R and all signals are identically distributed. The unconditional expectation of each signal is

$$\mathbb{E}(s_1)=P(s_1=1)=\int_0^1 \theta f(\theta)d\theta=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\alpha}{\alpha+\beta},$$

hence $\sum_{s_R\in\{0,1\}^{(k_j+1)}} lP(s_R)=(k_j+1)\frac{\alpha}{\alpha+\beta}$

Note that signals s_j are not unconditionally independent: indeed, higher θ will mean higher signals on average. For example, if 9 signals out of 10 are equal to 1, then the probability that the 10-th signal is also 1 is higher compared to the case where the first 9 signals were 0s. However, the signals are *conditionally* independent binary variables with $P(s_j=1|\theta)=\theta$, given θ. To use this fact, I rewrite $\sum_{s_R\in\{0,1\}^{(k_j+1)}} l^2P(s_R)$ using the law of iterated expectations in the following way:

$$\sum_{s_R\in\{0,1\}^{(k_j+1)}} l^2P(s_R)=\int_0^1\left(\sum_{s_R\in\{0,1\}^{(k_j+1)}} l^2P(s_R|\theta)\right)f(\theta)d\theta.$$

Since l is equal to the sum of signals in s_R, signals s_j are identically distributed and independent conditionally on θ, the term inside the integral can be rewritten as

$$\begin{aligned}\sum_{s_R\in\{0,1\}^{(k_j+1)}} l^2P(s_R|\theta) &= \mathbb{E}(l^2|\theta)=\mathrm{Var}(l|\theta)+(\mathbb{E}(l|\theta))^2\\ &= (k_j+1)\mathrm{Var}(s_1|\theta)+(k_j+1)^2(\mathbb{E}(s_1|\theta))^2\\ &= (k_j+1)\theta(1-\theta)+(k_j+1)^2\theta^2.\end{aligned}$$

Taking the integral:

$$\begin{aligned}\sum_{s_R\in\{0,1\}^{(k_j+1)}} l^2 P(s_R) &= \int_0^1 \left((k_j+1)\theta(1-\theta)+(k_j+1)^2\theta^2\right) f(\theta)d\theta \\ &= \frac{(k_j+1)}{B(\alpha,\beta)}B(\alpha+1,\beta+1)+\frac{(k_j+1)^2}{B(\alpha,\beta)}B(\alpha+2,\beta) \\ &= \frac{(k_j+1)}{(\alpha+\beta+1)(\alpha+\beta)}[\alpha\beta+(k_j+1)\alpha(\alpha+1)].\end{aligned}$$

Now $\mathbb{E}\left[\mathrm{Var}(\theta|s_R)\right]$ can be written as

$$\begin{aligned}&\frac{A}{(\alpha+\beta+k_j+1)^2(\alpha+\beta+k_j+2)} \\ =\ &\frac{\alpha(\beta+(k_j+1))+(\beta+(k_j+1)-\alpha)(k_j+1)\frac{\alpha}{\alpha+\beta}-\frac{(k_j+1)}{(\alpha+\beta+1)(\alpha+\beta)}[\alpha\beta+(k_j+1)\alpha(\alpha+1)]}{(\alpha+\beta+k_j+1)^2(\alpha+\beta+k_j+2)},\end{aligned}$$

which after several algebraic transformations boils down to

$$\frac{\alpha\beta}{(\alpha+\beta+k_j+1)(\alpha+\beta+1)(\alpha+\beta)}=h(k_j)$$

Finally, the total impact $h(b_i,b_j,k_j)$ becomes

$$h(b_i,b_j,k_j)=-\frac{\alpha\beta}{(\alpha+\beta+k_j+1)(\alpha+\beta+1)(\alpha+\beta)}-(b_j-b_i)^2=-h(k_j)-(b_j-b_i)^2,$$

which proves the result. **QED.**

Proof of Theorem 1. The necessary conditions for $g_{ij}=h$ and $g_{ij}=0$ follow directly from the incentive condition to form/maintain a hard link (1.4). To derive the necessary condition for $g_{ij}=s$—the incentive compatibility constraint of truthful reporting through a soft link ij—consider player j and let s_R be the set of k_j signals that player j gets to know apart from player i. Specifically, k_j-1 signals from his other communication neighbors $N_j^{-1}(g)/\{i\}$ and his own private signal s_j. Assuming that player j believes that i reports truthfully, let y_{s_R,s_i} be j's action if he has information s_R and i sends him the true signal, $m_{ij}=s_i$; $y_{s_R,1-s_i}$ be j's action if he has information s_R and i misreports, $m_{ij}=1-s_i$. Player i reports truthfully his signal s_i to j if and only if it generates a greater interim expected payoff to i compared to misreporting:

$$-\int_0^1 \sum_{s_R\in\{0,1\}^{k_j}} \left[(y_{s_R,s_i}-\theta-b_i)^2-(y_{s_R,1-s_i}-\theta-b_i)^2\right] f(\theta,s_R|s_i)d\theta \geq 0,$$

which can be rewritten as

$$-\int_0^1 \sum_{s_R\in\{0,1\}^{k_j}} \left[(y_{s_R,s_i}-y_{s_R,1-s_i})(y_{s_R,s_i}+y_{s_R,1-s_i}-2\theta-2b_i)\right] f(\theta,s_R|s_i)d\theta \geq 0.$$

Recalling that $y_{s_R,s_i}=b_j+\mathbb{E}(\theta|s_R,s_i)$, the condition becomes

$$\begin{aligned}-\int_0^1 \sum_{s_R\in\{0,1\}^{k_j}} &[(\mathbb{E}(\theta|s_R,s_i)-\mathbb{E}(\theta|s_R,1-s_i))\\ &\times(\mathbb{E}(\theta|s_R,s_i)+\mathbb{E}(\theta|s_R,1-s_i)+2b_j-2\theta-2b_i)]\, f(\theta,s_R|s_i)d\theta \geq 0.\end{aligned}$$

Note that

$$f(\theta,s_R|s_i)=\frac{f(\theta,s_R,s_i)}{P(s_i)}=\frac{f(\theta,s_R,s_i)}{P(s_R,s_i)}\frac{P(s_R,s_i)}{P(s_i)}=f(\theta|s_R,s_i)P(s_R|s_i).$$

Let $\Delta=\mathbb{E}(\theta|s_R,s_i)-\mathbb{E}(\theta|s_R,1-s_i)$ and change the sum and the integral signs

$$-\sum_{s_R\in\{0,1\}^{k_j}} \int_0^1 \left[\Delta(\mathbb{E}(\theta|s_R,s_i)+\mathbb{E}(\theta|s_R,1-s_i)+2b_j-2\theta-2b_i)\right] f(\theta|s_R,s_i)P(s_R|s_i)d\theta \geq 0.$$

Because $\mathbb{E}(\theta|s_R,s_i)$ and $\mathbb{E}(\theta|s_R,1-s_i)$ are independent of θ,

$$-\sum_{s_R\in\{0,1\}^{k_j}} \left[\Delta(\mathbb{E}(\theta|s_R,s_i)+\mathbb{E}(\theta|s_R,1-s_i)+2b_j-2\int_0^1 \theta f(\theta|s_R,s_i)d\theta-2b_i)\right] P(s_R|s_i) \geq 0;$$

$$-\sum_{s_R\in\{0,1\}^{k_j}} \left[\Delta(\mathbb{E}(\theta|s_R,s_i)+\mathbb{E}(\theta|s_R,1-s_i)+2b_j-2\mathbb{E}(\theta|s_R,s_i)-2b_i)\right] P(s_R|s_i) \geq 0;$$

$$-\sum_{s_R\in\{0,1\}^{k_j}} \left[\Delta(-\Delta+2b_j-2b_i)\right] P(s_R|s_i) \geq 0.$$

If there are l signals 1 in s_R, then

$$\mathbb{E}(\theta|s_R, s_i) = \mathbb{E}(\theta|l + s_i, k_j + 1) = \frac{\alpha + l + s_i}{\alpha + \beta + k_j + 1},$$

$$\mathbb{E}(\theta|s_R, 1 - s_i) = \mathbb{E}(\theta|l + 1 - s_i, k_j + 1) = \frac{\alpha + l + 1 - s_i}{\alpha + \beta + k_j + 1}.$$

Thus,

$$\Delta = \frac{\alpha + l + s_i}{\alpha + \beta + k_j + 1} - \frac{\alpha + l + 1 - s_i}{\alpha + \beta + k_j + 1} = \frac{2s_i - 1}{\alpha + \beta + k_j + 1},$$

which is independent on l, and hence on s_R. Using that $\sum_{s_R \in \{0,1\}^{k_j}} P(s_R|s_i) = 1$, the incentive condition becomes

$$-\frac{2s_i - 1}{\alpha + \beta + k_j + 1}\left(-\frac{2s_i - 1}{\alpha + \beta + k_j + 1} + 2(b_j - b_i)\right) \geq 0.$$

If $s_i = 1$, then player i is willing to communicate his signal if and only if

$$-\frac{1}{\alpha + \beta + k_j + 1}\left(-\frac{1}{\alpha + \beta + k_j + 1} + 2(b_j - b_i)\right) \geq 0;$$

$$b_j - b_i \leq \frac{1}{2(\alpha + \beta + k_j + 1)}.$$

If $s = 0$ then truth-telling is incentive compatible if and only if

$$-\frac{-1}{\alpha + \beta + k_j + 1}\left(-\frac{-1}{\alpha + \beta + k_j + 1} + 2(b_j - b_i)\right) \geq 0;$$

$$(b_j - b_i) \geq -\frac{1}{2(\alpha + \beta + k_j + 1)}.$$

As a result,

$$|b_j - b_i| \leq \frac{1}{2(\alpha + \beta + k_j + 1)},$$

which completes the proof of Theorem 1. **QED.**

Proof of Theorem 2. Consider a pure soft-link equilibrium $\left\{g(C_0), (\mu^{g(C_0)}, y^{g(C_0)})\right\}$. Renumber the players such that their in-degrees in the communication network $g(C_0)$ are increasing in their respective number: $k_1 \leq k_2 \leq ... \leq k_n$.

If $k_1 = n-1$, then the communication network $g(C_0)$ is complete. When the cost drops to the level of C_1, there exists an equilibrium $\left\{g(C_1), (\mu^{g(C_1)}, y^{g(C_1)})\right\} = \left\{g(C_0), (\mu^{g(C_0)}, y^{g(C_0)})\right\}$ that generates the same complete truthful network. Indeed, given a complete soft-link network $g(C_1)$, the strategy profile $(\mu^{g(C_0)}, y^{g(C_0)})$ forms a PBE. Moreover, no two players would like to substitute a soft link with a costly hard link, because that will result in ex-ante expected utility decrease. Consequently, in this case $k'_1 = k_1, ..., k'_n = k_n$.

Assume now that $k_1 < n-1$, i.e., the equilibrium network $g(C_0)$ is not complete and there are players who don't have all the signals reported to them. Consider three cases:

Case 1. Sufficiently large cost: $\max\{\alpha, 1-\alpha\}C_1 > (h(k_1)-h(k_1+1))$. In this case, there exists an equilibrium with the same communication network $g(C_0)$. Indeed, $\left\{g(C_0), (\mu^{g(C_0)}, y^{g(C_0)})\right\}$ remains an equilibrium when the cost is C_1, because the pair $(\mu^{g(C_0)}, y^{g(C_0)})$ forms a PBE given $g(C_0)$ and no two players would want to add a hard link or substitute a soft link with a hard link, because the cost is too high. Thus, it is possible to define $\left\{g(C_1), (\mu^{g(C_1)}, y^{g(C_1)})\right\} = \left\{g(C_0), (\mu^{g(C_0)}, y^{g(C_0)})\right\}$, hence, in this case $k'_j = k_j$ for all $j \in N$.

Case 2. Intermediate cost: $\max\{\alpha, 1-\alpha\}C_1 \in (h(k+1)-h(k+2), h(k)-h(k+1)]$ for some k, $k_1 \leq k \leq n-2$. Then by Theorem 1, it must be the case that in the equilibrium network every player has at least $(k+1)$ links directed to him. In particular, this means that player 1 with k_1 and several other players with in-degrees $< k+1$ will have a strict improvement in their information sets.

Consider the two subcases: (i) If $k+1 \geq k_n$, then each player gets at least $k+1 \geq k_n$ signals, which proves the result. (ii) If $k+1 < k_n$, then there is player j such that $k_j < k+1 \leq k_{j+1}$. Theorem 1 implies that $k'_l \geq k+1 > k_l$, $l = 1, ..., j$, in any equilibrium network $g(C_1)$. Concerning other players, there exists an equilibrium network with $k'_l = k_l$ for $l = j+1, ..., n$. To see this, suppose that the only links in $g(C_1)$ directed towards players $l = j+1, ..., n$ are the soft links from the pure cheap talk equilibrium network $g(C_0)$. Clearly, truthful communication along these links is still incentive compatible. Further, because $\max\{\alpha, 1-\alpha\}C_1 > h(k+1)-h(k+2)$, then no two players want to deviate and create hard links directed towards players $l = j+1, ..., n$. Thus, in the considered equilibrium network the in-degrees of players $l = 1, ..., j$ are strictly greater than in the pure cheap talk equilibrium, while the in-degrees of the other players are the same, which confirms the statement of the theorem.

Case 3. Sufficiently low cost: $\max\{\alpha, 1-\alpha\}C_1 \in [0, h(k_{n-1}) - h(k_n)]$. By Theorem 1, in any equilibrium network $g(C_1)$ every player has exactly $n-1$ links directed to him, which immediately implies that $n-1 = k'_j \geq k_j$ for all $j \in N$. **QED.**

Proof of Lemma 3. I split the proof into three steps:

Step 1: Existence of a maximal equilibrium. Because the number of players and strategies is finite, the number of the pure strategy equilibria is also finite. Thus, there exists a well-defined set of numbers, $k_1, ..., k_n$, where k_i is the highest in-degree of player i that can appear in some equilibrium: for any equilibrium network g', $k_i \geq k'_i = k_i(g')$. Note that the in-degrees k_i and k_j, $i \neq j$, in principle, might be achieved in different equilibrium networks. To prove an existence of a maximal equilibrium, I need to show that $k_1, ..., k_n$ might be achieved in the same equilibrium, i.e., that there exists an equilibrium network g such that $k_i = k_i(g)$ for all i. In order to do this, I construct the equilibrium in the following way: for each $i \in N$ consider an equilibrium where k_i is achieved and let those (and only those) players who report to i truthfully in that equilibrium to report truthfully to i through the same links (soft or hard) in the constructed equilibrium. Recall that the incentives to form the hard links depend only on the receiver's in-degree, while the incentives of truthful communication through the soft links depend also on the players' biases. Hence, it is still incentive compatible for those players to report truthfully to i through the respective soft links and to sustain the corresponding hard links. Thus, this is, indeed, an equilibrium, and, by construction, it is maximal.

Step 2: Maximality of a pairwise stable equilibrium. Consider some pairwise stable equilibrium and assume that it is not maximal. Then there exists player i whose in-degree in the equilibrium network is lower than his maximal in-degree. Fix some maximal equilibrium; then it must be the case that there is some agent j who reports truthfully to i (through either a soft or a hard link) in this maximal equilibrium, but not in the pairwise stable equilibrium. But then it is profitable for i and j to deviate and induce truthful communication from j to i, which contradicts the pairwise stability. Hence, every pairwise stable equilibrium must be maximal.

Step 3: Existence of a pairwise stable equilibrium. I illustrate this statement by constructing one of (possibly multiple) pairwise stable equilibria. For each $i \in N$ perform the following procedure: order other players $j \in N/\{i\}$ in the increasing absolute values of their preference divergence from i, $|b_j - b_i|$; let this order be $i_1, ..., i_{n-1}$. Consider the maximal in-degree of player i, k_i. If $k_i = 0$, then nobody can report truthfully to i in equilibrium. If $k_i > 0$, then take the closest player i_1. If truth-telling through a soft link $i_1 i$ is incentive compatible for i_1, given that i gets $k_i - 1$ other truthful messages, then let i_1 report truthfully to i via a soft link. Otherwise, set $g_{i_1 i} = h$. Repeat this procedure for other $k_i - 1$ closest to i players to set the links of particular type with truthful communication through them. Note, that for each of these closest k_i players, truthful communication through the corresponding link is desired, because truth-telling through a soft link is easier to sustain for closer preferences and the expected benefit of a hard link does not depend on a preference divergence. Since k_i is the maximal possible in-degree of player i, players $i_{k_i+1}, ..., i_{n-1}$ cannot communicate to i truthfully via either channel.

The described procedure leads to an equilibrium where each player gets truthful messages through the soft links from the players sufficiently close in their preferences, truthful messages through the hard links from the less close agents, and gets no messages from the more distinct players. To see that this equilibrium satisfies pairwise stability, first, note that by construction no two players i and j such that $g_{ji} = h$, could deviate to truthful communication through a soft link. Second, no two players i and j such that $g_{ji} = 0$, could implement truthful communication through either a soft or a hard link ji. Indeed, $g_{ji} = 0$ means that there are k_i other players closer to player i in their preferences than player j. If it were possible for i and j to deviate and implement truthful communication of j's signal through a soft link, given that k_i other players report truthfully to i, then truthful communication through soft links from each of the closest k_i players is also incentive compatible (closer biases relax the incentive condition of the truth-telling). But then it means that the in-degree of $k_i + 1$ can be sustained in the equilibrium network, which contradicts that k_i is maximal. On the other hand, if it were desirable for i and j to deviate and create a hard link ji, then truthful communication through the hard links from each of the closest k_i players must be desirable for them compared to no communication as well. Again, this means that the in-degree of $k_i + 1$ can be sustained in equilibrium, which contradicts that k_i is maximal. As a result, the constructed equilibrium satisfies pairwise stability. **QED.**

Proof of Theorem 3. Reorder the players in the pure cheap talk pairwise stable equilibrium correspondingly to their in-degrees in $g(C_0)$: $k_1 \leq k_2 \leq ... \leq k_n$. Because any pairwise stable equilibrium is maximal, the in-degrees $k_1, ..., k_n$ are maximal when the cost is C_0.

If $k_1 = n - 1$, then the truthful network $g(C_0)$ is complete, hence $g(C_1)$ must be also complete and consist of only soft links. Evidently, the total welfare is the same across the two cases: $W(g(C_1), C_1) = W(g(C_0))$.

Assume now that $k_1 < n - 1$ and consider two cases with regard to the cost C_1:

Case 1. Sufficiently large cost: $C_1 > h(k_1) - h(k_1 + 1)$. In this case the in-degrees $k_1, ..., k_n$ are still maximal and the pure cheap-talk pairwise stable equilibrium still remains a pairwise stable equilibrium. Thus, $g(C_1) = g(C_0)$ and $W(g(C_1), C_1) = W(g(C_0))$.

Case 2. Moderate cost: $C_1 \in (h(k+1) - h(k+2), h(k) - h(k+1)]$ for some $k,\ k \geq k_1$. Then by Theorem 1, it must be the case that in any equilibrium network every player has at least $\min\{k+1, n-1\}$ links directed to him.

If $k < k_n$, then there is player j such that $k_j < k+1 \leq k_{j+1}$. Then the set of maximal in-degrees becomes:

$$k'_1 = ... = k'_j = k+1,\ k'_{j+1} = k_{j+1},\ ...,\ k'_n = k_n.$$

There exists a maximal equilibrium with the following communication network: all links directed towards players $1, ..., j$ are hard, all links directed towards players $j+1, ..., n$ are soft and the same as in $g(C_0)$. The upper bound for the total cost of hard links is:

$$j(k+1)C_1 \leq j(k+1)(h(k) - h(k+1)).$$

The gain in the total welfare compared to the pure cheap talk case is

$$n \sum_{i=1}^{j} (h(k_i) - h(k+1)) \geq n \cdot j(h(k) - h(k+1)).$$

The lower bound for the welfare gain strictly exceeds the upper bound for the cost, because $n - 1 \geq k_n > k$, meaning that the constructed maximal equilibrium generates the welfare greater than $W(g(C0))$. Lemma 4 ensures that there is a pairwise stable equilibrium with the communication network $g(C_1)$ generating the total welfare greater than in the constructed maximal equilibrium. Hence, for this pairwise stable equilibrium, $W(g(C_1), C_1) > W(g(C_0))$.

If $k \geq k_n$, then the set of maximal in-degrees becomes $k'_1 = ... = k'_n = \min\{k+1, n-1\}$.

Analyze two possibilities:

(a) In case where $k_n < n-1$, consider a maximal equilibrium in which all links are hard. The upper bound for the total cost of the hard links is

$$n \min\{k+1, n-1\} C_1 \leq n \min\{k+1, n-1\}(h(k) - h(k+1)).$$

The gain in the total welfare compared to the pure cheap talk case is

$$n \sum_{i=1}^{n} (h(k_i) - h(k+1)) \geq n \cdot n(h(k) - h(k+1)).$$

The lower bound for the welfare gain strictly exceeds the upper bound for the cost, because $n > \min\{k+1, n-1\}$. The considered maximal equilibrium has the greatest level of the total cost and any other pairwise stable equilibrium generates at least the same total welfare. Hence, for any pairwise stable equilibrium with the communication network $g(C_1)$: $W(g(C_1) < C_1) > W(g(C_0))$.

(b) In case where $k_n = n-1$, there exists j such that

$$k_1 \leq k_2 \leq ... \leq k_j < n-1 = k_{j+1} = ... = k_n.$$

The set of maximal in-degrees becomes $k'_1 = ... = k'_n = n-1$. Consider a maximal equilibrium in which players $1, ..., j$ receive signals through only hard links, while players $j+1, ..., n$ receive messages through only soft links which are part of $g(C_0)$. The upper bound for the total cost of hard links is

$$j(n-1)C_1 \leq j(n-1)(h(k) - h(k+1)),$$

while the gain in the total welfare compared to the pure cheap talk case is

$$n \sum_{i=1}^{j} (h(k_i) - h(n-1)) \geq n \cdot j(h(n-2) - h(n-1)).$$

Since $k \geq k_n = n-1$, $h(k) - h(k+1) < h(n-2) - h(n-1)$ and the welfare gain outweighs the cost. Thus, $W(g(C_1), C_1) > W(g(C_0))$.

QED.

Proof of Lemma 5. Since $g^x(C_0) \subseteq g^x(C_1)$ for $x = s, h$, the in-degrees in $g(C_1)$ are larger: $k_i(g(C_1)) \geq k_i(g(C_0))$ for all $i \in N$. To see that $W(g(C_1), C_1) \geq W(g(C_0), C_0)$, perform the following procedure:

First, fix the truthful communication network to be $g(C_0)$ and set the cost level at C_1. Then the total welfare $W(g(C_0), C_1)$ is weakly greater than $W(g(C_0), C_0)$.

Second, add links from $g(C_1)/g(C_0)$ one by one and trace the welfare changes. In what follows I show that at each step of adding a link, the ex-ante expected payoff of every individual (and hence, the total welfare) increases. Consider some step at which the in-degrees of the players are $k_1, ..., k_n$, where $k_i(g(C_0)) \leq k_i \leq k_i(g(C_1))$ for all $i \in N$ and $k_j < k_j(g(C_1))$ for at least one player j. If a soft link ij from the remaining soft links $g^s(C_1)/g^s(C_0)$ is added, then the welfare of every player goes up by $h(k_j) - h(k_j + 1) > 0$. If a hard link ij from the remaining hard links $g^h(C_1)/g^h(C_0)$ is added, then the expected payoff of every player $l \neq i, j$ goes up by $h(k_j) - h(k_j + 1) > 0$, while the payoffs of players i and j increase by at least

$$h(k_j) - h(k_j + 1) - \max\{\alpha, 1 - \alpha\}C_1 \geq 0,$$

because by Theorem 1

$$\max\{\alpha, 1 - \alpha\}C_1 \leq h(k_j(g(C_1)) - 1) - h(k_j(g(C_1))) \leq h(k_j) - h(k_j + 1).$$

QED.

Proof of Theorem 4. Consider the setting in which only soft links are available. Condition $n_2 > n_1$ and the congestion effect ensure that cross-group truthful communication in any pairwise stable equilibrium might be one of the following 3 kinds:

1. No cross-group communication, i.e. $k_{21} = k_{12} = 0$.

2. Communication from group N_2 to group N_1, i.e. $k_{12} > 0,\ k_{21} = 0$.

3. Cross-group communication, i.e. $k_{12} > 0,\ k_{21} > 0$.

In what follows, I examine each case separately and show that introducing hard links is welfare increasing in cases 1 and 3, and is welfare decreasing for some parameters in case 2.

Case 1 corresponds to

$$b > \max\left(\frac{1}{2(\alpha+\beta+n_1)}, \frac{1}{2(\alpha+\beta+n_2)}\right) = \frac{1}{2(\alpha+\beta+n_1)}.$$

If costly hard links become feasible, then intra-group soft-link communication remains unchanged. Regarding cross-group communication, if the cost C_1 is such that $\max\{\alpha, 1-\alpha\}C_1 > h(n_1-1) - h(n_1)$, then cross-group communication remains empty and $W(g(C_1), C_1) = W(g(C_0))$. If $\max\{\alpha, 1-\alpha\}C_1 \leq h(n_1-1) - h(n_1)$, then some cross-group communication via hard links appears. Because the set of links in this equilibrium includes the set of links from the cheap talk case (no crowding out occurs), by Lemma 5, the equilibrium with hard links generates a greater total (and individual) welfare, $W(g(C_1), C_1) \geq W(g(C_0))$.

Case 2 applies when the preference bias b satisfies

$$b \in \left(\frac{1}{2(\alpha+\beta+n_1+k_{12}+1)}, \frac{1}{2(\alpha+\beta+n_1+k_{12})}\right], \tag{1.5}$$

where $n_1 + k_{12} \leq n_2$. Consider separately two possibilities: $n_1 + k_{12} = n_2$ and $n_1 + k_{12} < n_2$.

First, consider $n_1 + k_{12} = n_2$, which means that $k_1 = k_2 = n_2 - 1$. The total welfare in the pure cheap talk equilibrium is

$$W(g(C_0)) = -n(n_1 h(k_1) + n_2 h(k_2)) - B = -n^2 h(n_2 - 1) - B,$$

where the term B depends only on preference bias b, $B = 2\sum_{i\in N_1}\sum_{j\in N_2} b^2 = 2n_1 n_2 b^2$.

If the cost C_1 is such that $\max\{\alpha, 1-\alpha\}C_1 > h(n_1-1+k_{12}) - h(n_1+k_{12})$, then the maximal in-degrees remain the same and each pairwise stable equilibrium involves only soft-link communication, meaning that $W(g(C_1), C_1) = W(g(C_0))$. Now let the cost be $\widehat{C}$ such that $\max\{\alpha, 1-\alpha\}\widehat{C} = h(n_1 - 1 + k_{12}) - h(n_1 + k_{12}) = h(n_2-1) - h(n_2)$, which means that in a new pairwise stable equilibrium $k'_{21} = 1$, $k'_{12} = k_{12} + 1$, and hence, $k'_1 = k'_2 = n_2$.[24] Given the preference divergence, members of the opposite communities can not report to each other truthfully via soft links, thus,

[24] Recall the assumption that whenever the player is indifferent, the choice is made in favor of a hard link creation.

all cross-group soft links are substituted out by costly hard links leading to the following welfare:

$$\begin{aligned} W(g(\widehat{C}), \widehat{C}) &= -n(n_1 h(k_1') + n_2 h(k_2')) - B - \frac{1}{\max\{\alpha, 1-\alpha\}}(h(n_2-1) - h(n_2))[n_1 k_{12}' + n_2] \\ &= -n^2 h(n_2) - B - \frac{1}{\max\{\alpha, 1-\alpha\}}(h(n_2-1) - h(n_2))[n_1(k_{12}+1) + n_2]. \end{aligned}$$

Thus, the difference between the levels of total welfare is

$$W(g(C_0)) - W(g(\widehat{C}), \widehat{C}) = (h(n_2-1) - h(n_2))\left[-n^2 + \frac{1}{\max\{\alpha, 1-\alpha\}}(n_1(k_{12}+1) + n_2)\right] \leq 0,$$

because

$$\begin{aligned} -n^2 + \frac{1}{\max\{\alpha, 1-\alpha\}}(n_1(k_{12}+1) + n_2) &\leq -n^2 + 2(n_1(k_{12}+1) + n_2) \\ &= -(n_1+n_2)^2 + 2n_1(n_2 - n_1 + 1) + 2n_2 \\ &= -3n_1^2 - n_2^2 + 2n_1 + 2n_2 \leq 0, \end{aligned}$$

given that $1 \leq n_1 < n_2$. For any lower cost level $C_1 < \frac{h(n_2-1)-h(n_2)}{\max\{\alpha,1-\alpha\}} = \widehat{C}$, cross-group communication is also carried out through only hard links. Compared to the considered case of $\widehat{C}$, no soft links are severed and w.l.o.g. it can be assumed that all hard links are retained and some new hard links are added. Then Lemma 5 implies that the total welfare increases: $W(t(C_1), C_1) \geq W(g(\widehat{C}), \widehat{C}) \geq W(g(C_0))$. As a result, any pairwise stable equilibrium with hard links generates a greater total welfare than the pure cheap talk equilibrium.

Second, consider $n_1 + k_{12} + 1 \leq n_2$. In this case $k_1 = n_1 - 1 + k_{12}$, $k_2 = n_2 - 1$ and the total welfare in the pure cheap talk case is

$$W(g(C_0)) = -n(n_1 h(n_1 - 1 + k_{12}) + n_2 h(n_2 - 1)) - B.$$

If the cost $C_1 > \frac{1}{\max\{\alpha,1-\alpha\}}(h(n_1 - 1 + k_{12}) - h(n_1 + k_{12}))$, then the maximal in-degrees remain the same and the pairwise stable equilibria are pure soft-link, hence, $W(g(C_1), C_1) = W(g(C_0))$. Let the cost be $\widehat{C} = \frac{1}{\max\{\alpha,1-\alpha\}}(h(n_1 - 1 + k_{12}) - h(n_1 + k_{12}))$. Then cross-group communication becomes $k_{12}' = k_{12} + 1$, while k_{21}' remains 0; all cross-group links are hard. The new maximal in-degrees are $k'_1 = n_1 + k_{12}$ and $k'_2 = k_2 = n_2 - 1$, implying the following total welfare

$$\begin{aligned} W(g(\widehat{C}), \widehat{C}) = \; & - \quad n(n_1 h(n_1 + k_{12}) + n_2 h(n_2 - 1)) - B \\ & - \quad \frac{1}{\max\{\alpha, 1-\alpha\}}(h(n_1 - 1 + k_{12}) - h(n_1 + k_{12}))[n_1(k_{12} + 1)]. \end{aligned}$$

The difference in the levels of the welfare is then

$$W(g(C_0)) - W(g(\widehat{C}), \widehat{C}) = n_1(h(n_1 - 1 + k_{12}) - h(n_1 + k_{12})) \left[-n + \frac{1}{\max\{\alpha, 1-\alpha\}}(k_{12} + 1) \right],$$

which is strictly greater than 0 if and only if $k_{12} > \max\{\alpha, 1-\alpha\}n - 1$. Since the inequality is strict, there exists $\underline{C}_1 < \widehat{C}$ such that the welfare difference remains strictly positive for $C_1 \in (\underline{C}_1, \overline{C}_1)$, where $\overline{C}_1 = \widehat{C}$. Thus, given the preference divergence (1.5), condition $\max\{\alpha, 1-\alpha\}n - 1 < k_{12} \leq n_2 - n_1 - 1$ is sufficient for existence of the cost that leads to a lower welfare. The necessity follows from the fact that for any $C_1 < \widehat{C}$ Lemma 5 implies that $W(g(C_1), C_1) \geq W(g(\widehat{C}), \widehat{C})$.

Finally, **Case 3** corresponds to a sufficiently small preference divergence:

$$b \leq \frac{1}{2(\alpha + \beta + n_2 + 1)}.$$

Non-zero cross-group communication in both directions implies that the maximal in-degrees are the same for members of both communities, $k_1 = k_2$. If the truthful cheap talk network $g(C_0)$ is complete ($k_1 = k_2 = n - 1$), which corresponds to $b \leq \frac{1}{2(\alpha+\beta+n)}$, then introducing feasible hard links does not alter this pairwise stable equilibrium and $W(g(C_1), C_1) = W(g(C_0))$.

Case where $g(C_0)$ is not complete ($k_1 = k_2 < n - 1$) corresponds to the preference divergence

$$b \in \left(\frac{1}{2(\alpha + \beta + n_2 + k_{21} + 1)}, \frac{1}{2(\alpha + \beta + n_2 + k_{21})} \right],$$

where $k_{21} < n_1$. The total welfare of the pure cheap talk pairwise stable equilibrium is

$$W(g(C_0)) = -n^2 h(k_1).$$

If the cost $C_1 > \frac{1}{\max\{\alpha, 1-\alpha\}}(h(k_1) - h(k_1 + 1))$, then the maximal in-degrees are the same as in the cheap talk case and any pairwise stable equilibrium involves only soft communication, meaning that $W(g(C_1), C_1) = W(g(C_0))$. Consider $\widehat{C} = \frac{1}{\max\{\alpha, 1-\alpha\}}(h(k_1) - h(k_1+1))$. The new maximal in-degrees become $k_1' = k_2' = k_1 + 1$ and all cross-group communication in any pairwise stable equilibrium is performed via hard links, such that each member of community N_1 gets $k_{12}+1$ hard links from members of N_2 and, similarly, each member of N_2 gets $k_{21}+1$ hard links from members of N_1. The total welfare is

$$W(g(\widehat{C}), \widehat{C}) = -n^2 h(k_1+1) - \frac{1}{\max\{\alpha, 1-\alpha\}}(h(k_1) - h(k_1+1))(n_1(k_{12}+1) + n_2(k_{21}+1)).$$

The difference between the levels of the welfare is

$$\begin{aligned} & W(g(C_0)) - W(g(\widehat{C}), \widehat{C}) \\ = \quad & (h(k_1) - h(k_1+1)) \left[-n^2 + \frac{1}{\max\{\alpha, 1-\alpha\}}(n_1(k_{12}+1) + n_2(k_{21}+1)) \right] \leq 0, \end{aligned}$$

because

$$\begin{aligned} & -n^2 + \frac{1}{\max\{\alpha, 1-\alpha\}}(n_1(k_{12}+1) + n_2(k_{21}+1)) \leq -n^2 + 2n_1(k_{12}+1) + 2n_2(k_{21}+1) \\ = \quad & -n^2 + 2n_1(k_1 + 2 - n_1) + 2n_2(k_1 + 2 - n_2) = -n^2 + 2(k_1+2)n - 2n_1^2 - 2n_2^2 \\ \leq \quad & -n^2 + 2(k_1+2)n - n^2 = -2n(n - (k_1+2)) \leq 0, \end{aligned}$$

given that $k_1 < n - 1$. For any lower level of the cost $C_1 < \widehat{C}$ cross-group communication is performed through only hard links as well. Since, compared to the case of $\widehat{C}$, no cheap talk links are severed and w.l.o.g. it can be assumed that all hard links are retained and some new hard links are added, Lemma 5 implies that $W(g(C_1), C_1) \geq W(g(\widehat{C}), \widehat{C}) \geq W(g(C_0))$. **QED.**

Proof of Theorem 5. The case of $n = 3$ was already analyzed in the main body of the chapter; it remains to show the positive welfare result for $n \geq 4$.

Consider prohibitively costly hard links, $\max\{\alpha, 1-\alpha\}C_0 > h(0) - h(1)$, and construct a pairwise stable equilibrium that generates the greatest welfare in a way described in Lemma 4. In the corresponding truthful network, player i gets truthful messages from k_i closest players. This type of cheap talk equilibrium corresponds to the utility-maximizing equilibrium considered in [Galeotti *et al.*, 2011]. Following [Galeotti *et al.*, 2011], the communication network $g(C_0)$ can be formally described as follows: if $b > \frac{1}{2(\alpha+\beta+2)}$ then $g(C_0)$ is empty, otherwise, let $V(b) = \max\{V \in \{1, ..., n\} : b \leq \frac{1}{2V(2V-1+1+\alpha+\beta)}\}$, then

1. For every $j \in \{V(b)+1, ..., n-V(b)\}$, $g_{ij} = s$ if $|i-j| < V(b)$ and $g_{ij} = 0$ if $|i-j| > V(b)$; if $b > \frac{1}{2V(b)(2V(b)+1+\alpha+\beta)}$, then $g_{ij} = s$ for one and only one player i such that $|i-j| = V(b)$; if $b \leq \frac{1}{2V(b)(2V(b)+1+\alpha+\beta)}$, then $g_{ij} = s$ for both players i such that $|i-j| = V(b)$.

2. For all players $j \in \{1, ..., V(b)\} \cup \{n-V(b)+1, ..., n\}$, $g_{ij} = s$ if and only if $|i-j| \leq M(j,b)$, where $M(j,b) = \max\{M \in \{1, ..., n\} : b \leq \frac{1}{2M(\min\{j-1, n-j\}+M+1+\alpha+\beta)}\}$.

In the pure cheap talk equilibrium, the set of maximal in-degrees $K = \{k_1, ..., k_n\}$ is the following: $k_j = 0$, $j \in N$, if $b > \frac{1}{2(\alpha+\beta+2)}$; otherwise, for every $i \in \{V(b)+1, ..., n-V(b)\}$,

$$k_i = \begin{cases} 2V(b) - 1, \text{ if } b > \frac{1}{2V(b)(2V(b)+1+\alpha+\beta)} \\ 2V(b), \text{ if } b \leq \frac{1}{2V(b)(2V(b)+1+\alpha+\beta)}, \end{cases}$$

and for each $j \in \{1, ..., V(b)\} \cup \{n-V(b)+1, ..., n\}$,

$$k_j = \min\{j-1, n-j\} + M(j,b).$$

Given the set of maximal in-degrees K, define

$$\begin{aligned} k_{(n)} &= \max\{k_i \in K\}, \\ k_{(j)} &= \max\{k_i \in K/\{k_{(n)}, ..., k_{(j-1)}\}\}, \end{aligned}$$

i.e., $k_{(1)} \leq ... \leq k_{(n)}$ is a reordering of K in the increasing order. It can be easily seen, that players with moderate preferences, $\{V(b)+1, ..., n-V(b)\}$, have the highest in-degree $k_{(n)}$. The in-degrees of other players decrease as their preference biases get closer to the extremes, such that player 1 (with bias 0) and player n (with bias $(n-1)b$) have the same in-degree of $k_{(1)}$. Since $M(j,b) \in \{M(j+1,b), M(j+1,b)+1\}$ for $j \in \{1, ..., V(b)\}$ (similarly, $M(j+1,b) \in \{M(j,b), M(j,b)+1\}$ for $j \in \{n-V(b)+1, ..., n\}$) and $M(V(b),b) = M(n-V(b),b) = V(b)$, the structure of a pairwise stable equilibrium ensures that for every $i = 1, ..., n-1$ the difference $k_{(i+1)} - k_{(i)}$ is either 0 or 1.

Now introduce hard links with the cost $C_1 \leq \frac{h(0)-h(1)}{\max\{\alpha, 1-\alpha\}}$. If the communication network of the pairwise stable pure cheap talk equilibrium is empty ($b > \frac{1}{2(1+1+\alpha+\beta)}$), then there is no crowding out when hard links become available, and Lemma 5 implies that the welfare increases. Consider now the case where the pure cheap talk communication network is not empty, i.e., $k_{(n)} > 0$, and study three possibilities for C_1 separately

1. $\max\{\alpha, 1-\alpha\}C_1 \in (h(k_{(1)}) - h(k_{(1)}+1), h(0) - h(1)]$.
2. $\max\{\alpha, 1-\alpha\}C_1 \in (h(k+1) - h(k+2), h(k) - h(k+1)]$, for some $k_{(1)} \leq k < k_{(n)}$.
3. $\max\{\alpha, 1-\alpha\}C_1 \in (0, h(k_{(n)}) - h(k_{(n)}+1)]$.

Case 1. For the cost higher than $\frac{h(k_{(1)})-h(k_{(1)}+1)}{\max\{\alpha,1-\alpha\}}$, the maximal in-degrees remain the same and the pure cheap talk equilibrium remains pairwise stable, meaning that w.l.o.g. $g(C_1) = g(C_0)$ and $W(g(C_1), C_1) = W(g(C_0))$.

Case 2. Because the difference $k_{(i+1)} - k_{(i)}$ is either 0 or 1, there must exist i such that $k_{(i)} = k$ and $k_{(i+1)} = k+1$. The structure of the pure cheap talk equilibrium implies that i is an even number less or equal to $2V(b)$. For any C_1 that satisfies $\max\{\alpha, 1-\alpha\}C_1 \in (h(k+1)-h(k+2), h(k)-h(k+1)]$, the maximal in-degrees are

$$k'_j = \begin{cases} k+1, & \text{if } k_j \leq k, \\ k_j, & \text{if } k_j > k. \end{cases}$$

Consider a pairwise stable equilibrium that generates the greatest welfare. It must be the case that in the corresponding communication network $g(C_1)$ players $\{j : k_j > k\}$ receive messages via only soft links, in particular, assume that they receive truthful messages from the same players as in the pure cheap talk equilibrium. Other players $\{j : k_j \leq k = k_{(i)}\} = \{1, ..., \frac{i}{2}\} \cup \{n - \frac{i}{2}, ..., n\}$ have the new in-degrees equal to $k+1$ and can receive truthful messages through both, soft and hard links. If $j \in \{1, ..., \frac{i}{2}\}$, then the number of soft links directed to player j in $g(C_1)$ is at least $j - 1 + V(b)$. This implies that the upper bound on the number of costly hard links directed to j is $k + 1 - (j - 1 + V(b))$. Similarly, if $j \in \{n - \frac{i}{2}, ..., n\}$, then the number of hard links directed to player j is less or equal than $k + 1 - (n - j + V(b))$. Thus, the upper bound for the total cost of hard links is

$$\begin{aligned} & 2 \cdot \frac{h(k)-h(k+1)}{\max\{\alpha,1-\alpha\}} \sum_{j=1}^{\frac{i}{2}} [k+1-(j-1+V(\beta))] \\ = \quad & i\frac{h(k)-h(k+1)}{\max\{\alpha,1-\alpha\}} \left[k+2-V(\beta)-\frac{i+2}{4}\right]. \end{aligned}$$

The additional welfare is

$$2n\sum_{j=1}^{\frac{i}{2}} [h(k_j)-h(k+1)] \geq n \cdot i(h(k)-h(k+1)).$$

The upper bound for the cost is less than the lower bound for the additional benefit, because

$$n \geq 2\left[k+2-V(\beta)-\frac{i+2}{4}\right] \geq \frac{1}{\max\{\alpha,1-\alpha\}}\left[k+2-V(\beta)-\frac{i+2}{4}\right],$$

where the first inequality is satisfied due to $\frac{n}{2}+V(\beta) \geq k_{(n)} \geq k+1$ and $1-\frac{i+2}{4} \leq 0$. As a result, $W(g(C_1), C_1) \geq W(g(C_0))$.

Case 3. If $\max\{\alpha, 1-\alpha\}C_1 \in (0, h(k_{(n)}) - h(k_{(n)}+1)]$, then there is $k \geq k_{(n)}$ such that $\max\{\alpha, 1-\alpha\}C_1 \in (h(k+1)-h(k+2), h(k)-h(k+1)]$. There are several possibilities to consider:

1. If $k_{(n)} = n-1$, then $g(C_0)$ is either complete, or not. In case it is complete, the pure cheap talk equilibrium remains pairwise stable once hard links become feasible, thus, $W(g(C_1), C_1) = W(g(C_0))$. If $g(C_0)$ is incomplete, then there exists i: $1 \leq i < N$ such that

   $$k_{(1)} \leq ... \leq k_{(i)} < n-1 = k_{(i+1)} = ... = k_{(n)}.$$

   Consider a pairwise stable equilibrium that generates the greatest welfare when hard links are available. Condition $\max\{\alpha, 1-\alpha\}C_1 \leq h(n-1)-h(n)$ insures that in the corresponding communication network $g(C_1)$ each player $j \in N$ has the in-degree $k'_j = n-1$. Note that for the greater cost $C_2 = \frac{h(n-2)-h(n-1)}{\max\{\alpha,1-\alpha\}}$, a pairwise stable equilibrium that generates the greatest welfare has the same communication network, $g(C_2) = g(C_1)$. Since C_2 satisfies the conditions of case 2, then $W(g(C_1), C_1) > W(g(C_2), C_2) \geq W(g(C_0))$.

2. Consider $k_{(n)} \leq n-2$. Note that if the positive welfare result $W(g(C_1), C_1) \geq W(g(C_0))$

holds for $k = n - 2$, then by Lemma 5 it also holds for any $k > n - 2$. Thus, for the rest of the proof assume that $k \leq n - 2$. The in-degrees in the communication network of any equilibrium when hard links are available become $k'_j = k + 1$, $j \in N$. The total cost of hard links in the corresponding $g(C_1)$ does not exceed

$$n(k+1)\frac{h(k) - h(k+1)}{\max\{\alpha, 1-\alpha\}},$$

while the gain in the welfare compared to the cheap talk case is at least

$$n \cdot n(h(k) - h(k+1)).$$

The lower bound for the gain is greater than the upper bound for the cost if $k \leq n \max\{\alpha, 1-\alpha\} - 1$, meaning that for such C_1 the welfare in *any* equilibrium with hard links exceeds $W(g(C_0))$, in particular, $W(g(C_1), C_1) \geq W(g(C_0))$.

Consider now $n - 2 \geq k > n \max\{\alpha, 1-\alpha\} - 1$ and analyze two possibilities:

(a) If $k > k_{(n)}$, then the lower bound for the additional expected total benefit is

$$n^2(h(k-1) - h(k+1)) = n^2(h(k-1) - h(k) + h(k) - h(k+1)) \geq 2n^2(h(k) - h(k+1)),$$

which exceeds the upper bound for the cost $n(k+1)\frac{h(k)-h(k+1)}{\max\{\alpha,1-\alpha\}}$, because $\frac{k+1}{\max\{\alpha,1-\alpha\}} \leq 2(k+1) < 2n$. This means, that *any* equilibrium with hard links outperforms the cheap talk equilibrium in terms of welfare, hence, $W(g(C_1), C_1) \geq W(g(C_0))$.

(b) If $k = k_{(n)}$, then in any equilibrium with hard links each player j has the in-degree of $k'_j = k + 1 > n \max\{\alpha, 1-\alpha\} \geq \frac{n}{2}$, $j \in N$. Take a pairwise stable equilibrium that generates the greatest welfare, and consider how many links in the corresponding communication network $g(C_1)$ can be soft. Player i with the bias b_i such that $|b_i - b_j| = lb$, will report truthfully via cheap talk to player j if

$$b \leq \frac{1}{2l(k + 2 + \alpha + \beta)}.$$

Note that this inequality is satisfied for $l = V(b) - 1$, because by the definition of $V(b)$

and the fact that $k = k_{(n)} \leq 2V(b)$,

$$\frac{1}{2l(k+2+\alpha+\beta)} > \frac{1}{2V(b)(2V(b)+\alpha+\beta)} \geq b.$$

Thus, assuming $l = V(b) - 1$, each player $j \in \{1, ..., l\}$ gets truthful messages through soft links from $j - 1 + l$ players $1, ..., j-1, j+1, ..., j+l$. Similarly, each player $j \in \{n-l, ..., n\}$ gets at least $n - j + l$ truthful messages through soft links. Finally, each player $j \in \{l+1, ..., n-l-1\}$ gets at least $2l$ truthful messages via cheap talk. Thus, the number of hard links in $g(C_1)$ is bounded from above by

$$n(k+1) - 2\sum_{j=1}^{l}(j-1+l) - 2l(n-2l) = n(k+1) + l^2 + l - 2ln,$$

meaning that the total cost does not exceed

$$\frac{h(k) - h(k+1)}{\max\{\alpha, 1-\alpha\}}(n(k+1) + l^2 + l - 2ln).$$

Since $k = k_{(n)} \leq 2V(b)$ and $l = V(b) - 1$, then $k + 1 \leq 2(l+1)$ and the upper bound for the cost becomes

$$\frac{h(k) - h(k+1)}{\max\{\alpha, 1-\alpha\}}(2n(l+1) + l^2 + l - 2ln) = \frac{h(k) - h(k+1)}{\max\{\alpha, 1-\alpha\}}(2n + l^2 + l).$$

Note that $\frac{1}{\max\{\alpha, 1-\alpha\}} \leq 2$ and $l \leq \frac{n}{2} - 1$, because $2l < 2V(b) - 1 \leq k \leq n - 2$, which allows to write the upper bound for the cost as:

$$2(h(k) - h(k+1))\left(2n + \left(\frac{n}{2} - 1\right)^2 + \frac{n}{2} - 1\right) = (h(k) - h(k+1))\left(\frac{n^2}{2} + 3n\right)$$

The additional welfare is at least

$$n^2(h(k) - h(k+1)).$$

The lower bound for the extra welfare is greater than the upper bound for the cost if $n \geq 6$. Thus, $W(g(C_1), C_1) \geq W(g(C_0))$ for $n \geq 6$.

It remains to show that $W(g(C_1), C_1) \geq W(g(C_0))$ for $n = 4$ and $n = 5$ as well, assuming that

$$n - 2 \geq k = k_{(n)} > n \max\{\alpha, 1-\alpha\} - 1 \geq \frac{n}{2} - 1$$

and

$$\max\{\alpha, 1-\alpha\} C_1 \in (h(k+1) - h(k+2), h(k) - h(k+1)].$$

Consider, first, $n = 5$ and $k = 2, 3$. Depending on b, case $k = k_{(n)} = 2$ corresponds to pairwise stable pure cheap talk equilibria with the following in-degrees (see Figure 1.6):

(i) If $\frac{1}{4(3+\alpha+\beta)} < b \leq \frac{1}{2(3+\alpha+\beta)}$, then $k_1 = k_5 = 1$, $k_2 = k_3 = k_4 = 2$.

(ii) If $\frac{1}{4(4+\alpha+\beta)} < b \leq \frac{1}{4(3+\alpha+\beta)}$, then $k_1 = k_2 = k_3 = k_4 = k_5 = 2$.

When hard links become available with $\max\{\alpha, 1-\alpha\} C_1 \in (h(3) - h(4), h(2) - h(3)]$, the maximal in-degrees become $k_1' = ... = k_5' = 3$. In case (i), the upper bound for the total cost of hard links is

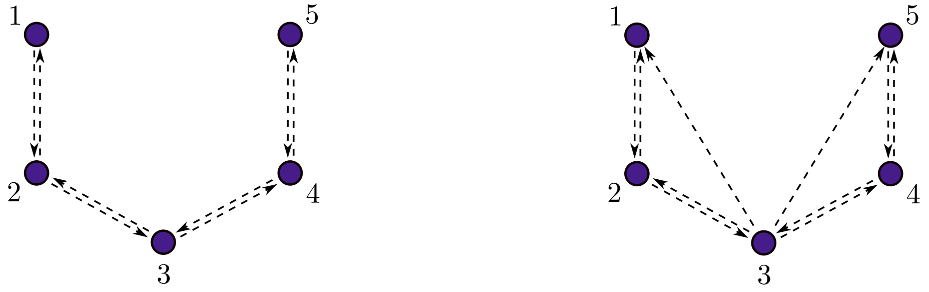
$$5 \cdot 3 \cdot 2(h(2) - h(3)),$$

which is lower than the gain in the total welfare

$$5 \cdot [3(h(2) - h(3)) + 2(h(1) - h(3))],$$

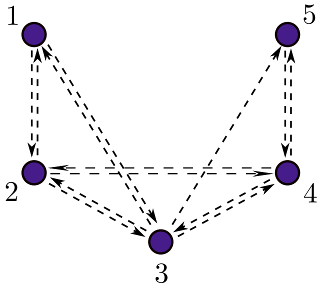
because $h(2) - h(3) < 2(h(1) - h(2))$. In case (ii), soft links between the players with adjacent biases can be a part of the communication network of a pairwise stable equilibrium. Thus, the total cost of hard links in $g(C_1)$ does not exceed $7 \cdot 2(h(2) - h(3))$, which, in turn, is lower than the additional total welfare, $5 \cdot 5(h(2) - h(3))$.

Case $k = k_{(n)} = 3$ corresponds to $\frac{1}{4(5+\alpha+\beta)} < b \leq \frac{1}{4(4+\alpha+\beta)}$ and maximal in-degrees $k_1 = k_5 = 2$, $k_2 = k_3 = k_4 = 3$ in the pure cheap talk setting (see Figure 1.7 for an example of $g(C_0)$). When hard links become available with $\max\{\alpha, 1-\alpha\} C_1 \in (h(4) - h(5), h(3) - h(4)]$, the maximal in-degrees become $k_1' = ... = k_5' = 4$. Note that soft links between the players with adjacent biases can be a part of the communication network of a pairwise stable equilibrium, implying that the total cost of hard links in $g(C_1)$ has an upper bound of $12 \cdot 2(h(3) - h(4))$, which is lower than the gain in the total welfare, $5 \cdot [3(h(3) - h(4)) + 2(h(2) - h(4))]$. Hence, $W(g(C_1), C_1) \geq W(g(C_0))$, when $n = 5$.



(i) $k_1 = k_5 = 1, k_2 = k_3 = k_4 = 2$ (ii) $k_1 = k_2 = k_3 = k_4 = k_5 = 2$

Figure 1.6: Communication networks of pairwise stable pure cheap talk equilibria when $n = 5$ and $k = 2$.



$k_1 = k_5 = 2,\ k_2 = k_3 = k_4 = 3$

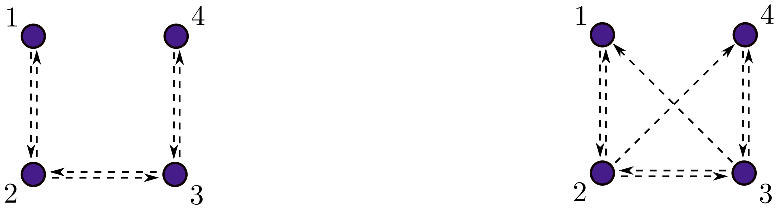
Figure 1.7: Communication network of pairwise stable pure cheap talk equilibrium when $n = 5$ and $k = 3$.

Consider now $n = 4$ and $k = 2$. Depending on b, $g(C_0)$ can have two different structures with the following in-degrees (see Figure 1.8):

(i) If $\frac{1}{4(3+\alpha+\beta)} < b \le \frac{1}{2(3+\alpha+\beta)}$, then $k_1 = k_4 = 1,\ k_2 = k_3 = 2$.

(ii) If $\frac{1}{4(4+\alpha+\beta)} < b \le \frac{1}{4(3+\alpha+\beta)}$, then $k_1 = k_2 = k_3 = k_4 = 2$.

When hard links become available with $\max\{\alpha, 1-\alpha\}C_1 \in (h(3) - h(4), h(2) - h(3)]$, the maximal in-degrees become $k'_1 = ... = k'_4 = 3$. In case (i), the total cost of hard links has an upper bound of



(i) $k_1 = k_4 = 1, k_2 = k_3 = 2$ (ii) $k_1 = k_2 = k_3 = k_4 = 2$

Figure 1.8: Communication network of pure cheap talk equilibrium when $n = 4$ and $k = 2$.

$3 \cdot 4 \cdot 2(h(2) - h(3))$, which is lower than the gain in the total welfare $4 \cdot [2(h(2) - h(3)) + 2(h(1) - h(3))]$, because $h(2) - h(3) < h(1) - h(2)$. In case (ii), soft links between the players with adjacent biases can be a part of the communication network of a pairwise stable equilibrium. Thus, the total cost of hard links in $g(C_1)$ is below $6 \cdot 2(h(2) - h(3))$, which, in turn, is lower than the additional total welfare, $4 \cdot 4(h(2) - h(3))$. As a result, $W(g(C_1), C_1) \geq W(g(C_0))$, when $n = 4$. **QED.**

Chapter 2

Strategic communication in networks: Preferences versus opinions

Uliana Loginova

Abstract

This chapter studies situations in which every agent needs to take an action that matches his preferences given the state of the world and that action affects the payoffs of all other agents. Before deciding upon the action, agents can choose to whom they communicate their private information about the state by means of cheap talk. I find that the information transmission pattern crucially depends on the nature of the disagreement. If the agents diverge in their preferences, then information transmission exhibits a *negative externality effect*: greater information obtained by some agent discourages further information accumulation by harming the credibility of other agents. In contrast, information transmission displays a *positive externality effect* when the agents have divergent opinions: greater information obtained by some agent encourages further information accumulation by improving the credibility of other agents. I show that, under conflicting preferences, the information is likely to flow from a larger community to a smaller one; while, under conflicting opinions, the information flow is reversed and a smaller community is more likely to report to a larger one. Finally, I demonstrate that replicating the set of agents curbs cross-type communication when the agents diverge in their preferences; and boosts cross-type communication when the agents disagree in their opinions.

2.1 Introduction

In many real-life situations, people need to decide on the most appropriate action to take in the face of an uncertain environment, actions may influence other participants as well. For example, managers of a corporation must choose marketing strategies, the effectiveness of which depends on tomorrow's market condition. Managers are mainly responsible for their respective divisions and might differ in their goals and in their perceptions of the uncertain economic environment. Nevertheless, the divisions are parts of one corporation, and the strategy chosen by some manager has an effect on other divisions' payoffs. Another example is a group of politicians, each responsible for choosing a policy in his respective jurisdiction; the chosen policies generate externalities across districts. Politicians might have different preferences over the possible strategy courses and different judgements of an uncertain future political environment that reflect their own political views as well as the views of their constituencies. In this chapter, I formalize the interdependence of individuals' payoffs by assuming that each agent faces a loss when his action and the actions of other agents differ from his own "ideal action"—an action that matches the individual's preferences given the state of the world.

Oftentimes, individuals base their decisions not only on their private information, but also on the information obtained from other agents during preliminary discussions. However, even when individuals get to know the same information about the state of the world, they might still disagree on the proper actions, either because they have *conflicting preferences* or because they have *conflicting opinions*. Under conflicting preferences, the individuals agree on the fundamentals of the uncertain economic environment but simply like different courses of action. E.g., some managers advocate rather aggressive expansion strategies because they have a taste for empire building. Under conflicting opinions, the individuals would prefer the same actions had they known the exact state of the world; however, under incomplete information, they diverge in their opinions about the uncertain environment and, thus, deem different actions optimal. E.g., some managers prefer a more aggressive expansion strategy, because they hold a more optimistic belief about the corporation's prospects in tomorrow's market.[1]

[1] On possible reasons for the existence and persistence of different priors see, e.g., [Aumann, 1976], [Tversky and Kahneman, 1974], [Acemoglu *et al.*, 2007], [Sethi and Yildiz, 2009]. Also see [Morris, 1995] for a discussion of the assumption of different priors in the economics literature.

This chapter analyzes how information transmission patterns depend on the nature of disagreement. The framework I study is closely related to [Galeotti *et al.*, 2011], [Loginova, 2012b] and [Hagenbach and Koessler, 2010]. In the model, there are n players with preference biases $b_1, ..., b_n$, who face an uncertain economic environment summarized by θ. Individuals might hold different beliefs about the economic environment: agent i's prior of θ is characterized by Beta distribution with parameters (α_i, β_i). The state of the world relevant to the decision-making process, S, is composed of D different aspects, $S = \sum_{d=1}^{D} s_d$. Conditional on the underlying economic environment θ, the aspects are independent and identically distributed, with each s_d taking a value of 1 with probability θ and 0 with complementary probability $1 - \theta$. The total number of aspects is greater than the number of agents, $D \geq n$, and each agent i is privately informed of the respective aspect s_i.

Individuals can simultaneously communicate their private information to each other according to the communication network, which is set prior to the aspects' realization. The network is described by a directed graph, such that player i can send cheap talk messages to other agents that he has links to. After the communication takes place, individuals simultaneously choose actions $y_1, ..., y_n$ that influence each other's payoffs, such that every individual i obtains $-\sum_{j=1}^{n}(y_j - S - b_i)^2$. That is, individual i's payoff depends on how close his action, as well as the actions of other agents, are to individual i's ideal action, $\theta + b_i$. Under conflicting preferences, individuals have different biases $b_1, ..., b_n$, but share the common prior belief, $(\alpha_i, \beta_i) = (\alpha, \beta)\ \forall i$. Under conflicting opinions, individuals diverge in their prior beliefs about θ, but have fully aligned preferences, $b_i = 0\ \forall i$.

Assume that player i has a link to player j and, apart from i's message, player j gets to know k_j aspects from other agents. Under either type of disagreement between the agents, the two parties, i and j, would ex-ante prefer truthful revelation of s_i to agent j. However, the credibility of communication and the way it is influenced by k_j crucially depends on the nature of disagreement between the individuals. When the two parties diverge in their preferences, I find a *negative externality effect* of information transmission: greater information available to player j—higher k_j—*harms the credibility* of player i.[2] Intuitively, the magnitude of the effect from an additional signal on j's action decreases with the informativeness of agent j. Thus, for sufficiently high k_j, the expected effect might become so small that i would prefer to lie in order to shift j's action closer

[2] This corresponds to the congestion effect of [Galeotti *et al.*, 2011] and [Loginova, 2012b].

towards i's preferred one.

In contrast, when the parties disagree in their opinions, the communication pattern exhibits *positive externality effect*: greater number of aspects obtained by agent j—higher k_j—*improves the credibility* of agent i. Intuitively, as k_j increases, two things happen. First, agent i expects the ex-post belief of agent j to become more congruent. Indeed, agent i deems other signals revealed to player j distributed according to i's belief, hence, he expects player j to be persuaded and adjust his ex-post belief in the "right" direction (from i's point of view). Second, the magnitude of the effect of an additional signal on j's action decreases with k_j. However, the additive nature of the state S in aspects s_d insures that the rate of decrease is sufficiently low, so that the effect of i's message on j's action remains significant enough to prevent player i from misreporting to player j (given that i expects j to become more congruent).

As one application of the main findings, I study communication patterns between two communities composed of players who agree in their preferences (or priors). Similarly to [Galeotti *et al.*, 2011] and [Loginova, 2012b], I find that under conflicting preferences, the information is more likely to flow from a larger community to a smaller one. Intuitively, members of the large community accumulate many aspects from their like-minded colleagues, which, by the negative externality effect, curbs truthful reporting from the members of the small community. Under conflicting opinions, however, members of the small community are more likely to report truthfully to members of the large community. The reason being that greater information accumulated inside the large community, by the positive externality effect, improves the credibility of the small community members.

As another application, I consider replicas of a diverse group of people. I find that, under conflicting preferences, increasing the number of individuals of the same type curbs cross-type communication and can lead to a segregation of communication according to the players' types. Indeed, more intense communication between the same-type individuals, by the negative externality effect, can deteriorate the credibility of the cross-type communication. As a result, adding new players who are privately informed about other aspects does not necessarily improve the aggregation of information. Under conflicting opinions, on the contrary, replicating the set of players, by the positive externality effect, boosts cross-type communication, which necessarily leads to a more information aggregation.

The chapter is organized as follows. Section 2.2 develops the model and describes the types of conflict. Section 2.3 studies the credibility of communication and characterizes equilibrium communication patterns for each type of conflict. Applications of the main results to the particular cases of two communities and replicas of a diverse group of people are considered in Section 2.4. Section 2.5 concludes the chapter. Finally, Appendix 2.6 provides additional details on characterization of the pairwise stable equilibria, and Appendix 2.7 contains all the proofs.

2.1.1 Related literature

This chapter contributes to the strand of literature on strategic information transmission, which builds on the classic model of cheap talk communication by [Crawford and Sobel, 1982] and [Green and Stokey, 2007]. This literature is very rich, and here I mention only the few of most related studies.

The setting of this chapter builds on [Galeotti *et al.*, 2011] and [Loginova, 2012b]: it features the same type of payoff interdependence and departs from their framework by considering the additive state of the world and allowing for opinion conflict. Naturally, because the marginal value of learning an additional signal decreases with the informativeness of the individual, I obtain similar results under conflicting preferences, namely, the negative externality result that replicates the congestion effect of [Galeotti *et al.*, 2011] and [Loginova, 2012b]. Another closely related paper is [Hagenbach and Koessler, 2010], who focus on preference conflict and study strategic cheap talk communication in networks. Importantly, their model differs in two ways: First, the state of the world is a sum of aspects that are sampled from some *known* distributions. Hence, knowing one aspect does not help in predicting some other unknown aspect (while in my model, in which all aspects come from the same distribution, knowing some aspect allows one to make an inference about an unknown one). This assumption implies a constant marginal value of every additional signal; hence, the negative externality result is absent from their model. Second, they consider a different type of coordination, i.e., each player not only wants others' actions to be closer to his action (as is the case in my model), but also also wants to match his action with the actions of others.

Other related papers with cheap talk communication under preference conflict, include settings with multiple senders (e.g. [Morgan and Stocken, 2008]) and multiple receivers (e.g. [Caillaud and Tirole, 2007]). While in this chapter I focus on private communication, there are studies that compare private with public information transmission, e.g. [Farrell and Gibbons, 1989], [Goltsman and Pavlov, 2011] and [Galeotti *et al.*, 2011].

That a difference in the outcomes under conflicting preferences and under conflicting opinions can be crucial is documented in other papers as well. [Che and Kartik, 2009] show that difference of opinion between a decision maker and an advisor can increase the advisor's incentive to exert effort to persuade the decision maker; at the same time, differences in preferences cannot induce such a persuasion motive. [Hirsch, 2011], [Van den Steen, 2006] and [Van den Steen, 2009] also analyze this mechanism. [Hirsch, 2011] illustrates how open disagreement in opinions between a principal and an agent creates a persuasion-based rationale for deference: the principal can allow the agent to implement the agent's preferred policy in the first period in order to let the agent learn from his own mistakes and increase his effort on the principal's preferred policy in the second period. [Van den Steen, 2009] shows that a principal will rely on persuasion—costly alteration of the agent's beliefs—for projects that need high agent's effort. Relatedly, [Van den Steen, 2006] shows that a principal may delegate decision rights to the agent in order to increase his efforts, which is crucial for the project's success. [Loginova and Persson, 2012] demonstrate that a benevolent authority's decision to constrain or inform a population of agents crucially depends on whether the authority deems it a matter of preference or opinion. In the former case, the authority gives truthful advice and safeguards liberty; in the latter, the authority constrains liberty, believing that she acts in the population's interest.

Characterization of communication patterns in cases of two communities and replicas of a diverse group of people contributes to the literature on organizational design with cheap talk (e.g. [Alonso *et al.*, 2008], [Rantakari, 2008]), verifiable information (e.g. [Bolton and Dewatripont, 1994], [Radner, 1992], [Radner, 1993], [Sah and Stiglitz, 1986], [Van Zandt and Radner, 2001]) noisy hard talk, where the information is transmitted perfectly with some probability less than 1 (e.g. [Dessein and Santos, 2006]), and combined soft and hard information (e.g. [Dessein, 2007]).

Also related are studies that focus on questions of coordination and adaptation with verifiable information transmission in communication networks. In particular, [Chwe, 2000] studies a collective action problem with preliminary communication regarding participation activity in a deterministic exogenous network. [Calvó-Armengol and de Martí, 2007] and [Calvó-Armengol and de Martí, 2009] analyze how the communication pattern affects individual behavior and aggregate welfare in a setting in which the agents not only want to coordinate their actions, but also adapt to an unknown state of the world. Instead, [Calvó-Armengol *et al.*, 2011] consider local uncertainty regarding the state, and study information transmission patterns that arise when the agents are allowed to alter the communication precision.

Finally, the analysis of the communication patterns arising in equilibria contributes to the literature of strategic network formation, which includes [Bala and Goyal, 2000], [Goyal, 2007], [Jackson, 2008], [Jackson and Wolinsky, 1996].

2.2 Model

Let the set of players be $N = \{1, ..., n\}$ with $n \geq 2$, where each player i has the preference bias b_i. The underlying *economic environment* is summarized by θ that is unknown to the players. Each player i's prior of θ is characterized by Beta distribution with parameters (α_i, β_i) and density of

$$f_i(\theta) = \frac{1}{B(\alpha_i, \beta_i)} \theta^{\alpha_i - 1}(1-\theta)^{\beta_i - 1}.$$

The preference profile $\{b_1, ..., b_n\}$ and the players' priors are publicly known.

There are D different *aspects* $s_1, ..., s_D$ that determine the *state of the world* as $S = \sum_{d=1}^{D} s_d$. Conditional on the underlying economic environment θ, the aspects $\{s_d\}_{d=1}^{D}$ are independent and identically distributed, and $s_d = 1$ with probability θ and $s_d = 0$ with complementary probability $1 - \theta$. The total number of aspects is greater than the number of players, $D \geq n$, and each player i is privately informed of the aspect s_i (i.e., player i receives the private signal s_i). Thus, all players cumulatively get to know the first n of D aspects. Note that if $D = n$, then the players jointly hold all the relevant information regarding the state of the world.

The communication network is set prior to the aspects' realization and is described by a directed graph $g \in \{0, 1\}^{n \times n}$, where player i communicates a cheap talk message about his signal s_i to player j if and only if $g_{ij} = 1$.[3] It is assumed that the communication links are cheap to sustain: each involved party, either sending the message or receiving it, bears just an infinitesimal cost $\varepsilon > 0$.

[3] The assumption that the communication network is set up before realization of the signals can, for example, be justified by the necessity of forming a communication schedule beforehand.

While the players are aware of each other's existence and each other's preferences and priors, the communication network g is not commonly known. Rather, each player i knows only the structure of his respective neighborhood: the set of players to whom player i has links, $N_i(g) = \{j \in N : g_{ij} = 1\}$, and the set of players who have links directed to player i, $N_i^{-1}(g) = \{j \in N : g_{ji} = 1\}$.

Communication. After the signals are realized, players communicate their privately observed aspects of the state of the world. Each player i sends private message $m_{ij}^g \in \{0,1\}$ to every player j that he has a link to in the communication network g. It is assumed that communication takes the form of cheap talk, and that messages are sent simultaneously and are observed only by the sending and the receiving parties. A *communication strategy* of player i with the private signal s_i defines a vector

$$\mu_i^g(s_i) = \{\mu_{ij}^g(s_i)\}_{j \in N_i(g)} \in \{0,1\}^{|N_i(g)|}.$$

A communication strategy profile is denoted by $\mu^g = \{\mu_1^g, \ldots, \mu_n^g\}$. The messages actually sent by player i are denoted by vector $\widehat{m}_i^g$; the profile of all sent messages is $\widehat{m}^g = \{\widehat{m}_1^g, \ldots, \widehat{m}_n^g\}$.[4]

Decision making. After the communication has taken place, each player i chooses an action $y_i^g \in \mathbb{R}$. The *action strategy* of player i is a function of his information set that consists of his own signal, s_i, and the messages he gets from $N_i^{-1}(g)$, $\widehat{m}^g_{N_i^{-1}(g),i}$:

$$y_i^g : \{0,1\} \times \{0,1\}^{|N_i^{-1}(g)|} \to \mathbb{R}.$$

Let $y^g = \{y_1^g, ..., y_n^g\}$ denote the action strategy profile. Conditional on the state of the world S, if the chosen action profile is $\hat{y}^g = \{\hat{y}_1^g, ..., \hat{y}_n^g\}$ then the realized payoff (utility) of player i is

$$u_i(\hat{y}^g | S) = -\sum_{j=1}^{n} (\hat{y}_j^g - S - b_i)^2.$$

The payoff of player i increases as his own action and the actions of other players get closer to player i's ideal action, $S + b_i$.

[4] The superscript g for communication strategies highlights the dependence on the network structure. For the sake of simplicity, I use the same superscript g for all players; note, however, that every player i conditions his communication strategy only on the structure of his respective neighborhood.

Time notation. Introduce the following time notation in order to distinguish between periods with different scopes of information available to the agents: *"ex-ante"* to denote the stage prior to when the signals are realized, *"interim"* for the time period after the signals' realization but prior to communication, and finally, *"ex-post"* to represent the period after communication has occurred but before the actions are taken.

2.2.1 Nature of disagreement

I develop the analysis under two different assumptions about the nature of disagreement between the players:

Conflicting preferences. Under conflicting preferences, at least some players diverge in their preference biases, i.e., $b_i \neq b_j$ for some i and j. However, all agents share the common prior belief about the underlying economic environment θ that has Beta distribution with parameters (α, β), i.e., $(\alpha_i, \beta_i) = (\alpha, \beta)$ for every $i \in N$.

Conflicting opinions. Under conflicting opinions, the players have fully aligned preferences, i.e., $b_i = 0$ for all $i \in N$; however, at least some players hold different prior beliefs about θ, i.e., $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$ for some i and j. In the main body of the chapter, I assume that the sum of parameters of Beta distribution is the same across all players: $\alpha_i + \beta_i = \gamma$, $i \in N$.[5]

2.2.2 Solution concept

To solve the model, I use the concept of pure strategies Perfect Bayesian Equilibrium (PBE). The restriction to pure strategies simplifies the analysis and implies that cheap talk communication can be either *truthful* (the message reflects the signal perfectly), or *uninformative* (for any signal s_i player i sends the same message, either 0 or 1). In the case of uninformative communication, I assume that an off-equilibrium-path message is ignored by the receiving party. This simplification

[5] Condition $\alpha_i + \beta_i = \gamma$, $i \in N$, simplifies derivations and presentation of the results. Importantly, it ensures that the monotone likelihood ratio for prior distributions holds, which, in particular, means that the players interpret the signals and update their beliefs consistently with each other. Note, however, that the main results hold in a general case as well, even though the intuition behind them in cases in which the monotone likelihood property fails is somewhat more subtle.

implies that the equilibrium beliefs are defined as follows: any message received in truthful communication induces perfect knowledge about the underlying signal, while any message received in uninformative communication leaves the prior belief about the underlying signal unchanged.

For any given communication network g, it is natural to determine the communication and action strategy profile $(\mu^g, y^g) = (\{\mu_i^g\}_{i\in N}, \{y_i^g\}_{i\in N})$ by using the standard PBE solution concept. Note, however, conditional on a particular choice of g and equilibrium (μ^g, y^g), some links might be ex-ante undesired by at least one involved party. In particular, all links with uninformative communication through them are ex-ante unprofitable to both parties.[6] To account for this, I define an equilibrium as a triple $\{g, (\mu^g, y^g)\}$ such that (1) the pair (μ^g, y^g) forms a PBE given the communication network g, and (2) no player would prefer to break some incoming or outgoing link at the ex-ante stage.

To formally state the equilibrium definition, let $\mathbb{E}u_l\left(g, \mu^g, y^g\right)$ be the ex-ante expected utility of agent l that takes into account all link costs that accrue to agent l. Let $g(g_{ij} = 0)$ be the communication network with the same set of links as in g, except that there is no link from i to j; let $\mu^{g(g_{ij}=0)}$ be the profile of communication strategies that coincides with μ^g everywhere, except that now there is no communication from i to j; and let $y^{g(g_{ij}=0)}$ be the same action profile as y^g for all players but j, while player j's action is now optimally defined conditional on the lower number of truthful messages. Below is the formal definition,in which, under conflicting preferences, all expectations are evaluated using the common prior, while under conflicting opinions, each player uses his own prior when choosing communication and action strategies.

Definition. *Equilibrium* $\{g, (\mu^g, y^g)\}$ consists of a communication network g and a strategy profile $(\mu^g, y^g) = (\{\mu_i^g\}_{i\in N}, \{y_i^g\}_{i\in N})$, such that the following properties hold:

(1) The pair (μ^g, y^g) forms a PBE given the communication network g.

(i) Given the action strategies profile y^g, every agent $i \in N$ for any $s_i \in \{0, 1\}$ and every player j to whom i has a link, $j \in N_i(g)$, chooses a message $\mu_{ij}^g(s_i) \in \{0, 1\}$ in order to maximize his interim expected utility.

[6] Note that no player will object to a link with truthful communication. Indeed, as is shown in Lemmas 7 and 8, destroying the link with truthful communication will strictly harm the ex-ante expected payoffs of both parties involved in the link (provided that the cost ε is infinitesimal).

(ii) Every player $i \in N$, for any private signal, $s_i \in \{0,1\}$, and any set of received messages, $\widehat{m}^g_{N_i^{-1}(g),i} \in \{0,1\}^{|N_i^{-1}(g)|}$, chooses an action $y_i^g\left(s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right)$ to maximize his ex-post expected utility.

(iii) The beliefs are consistent with the communication strategies.

(2) For any i and j such that $g_{ij} = 1$:

$$\mathbb{E}u_l\left(g, \mu^g, y^g\right) \geq \mathbb{E}u_l\left(g(g_{ij}=0), \mu^{g(g_{ij}=0)}, y^{g(g_{ij}=0)}\right), \quad \text{for } l = i, j.$$

Remark. In any equilibrium $\{g, (\mu^g, y^g)\}$, communication network g is *truthful*, i.e., all links of g represent truthful revelation of private signals. Different equilibria correspond to different communication networks.

2.3 Analysis

2.3.1 Choice of action

To derive the optimal choice of action, fix the communication network g and consider agent i who learned his private signal s_i and received the messages $\widehat{m}^g_{N_i^{-1}(g),i}$ from his neighbors $N_i^{-1}(g)$. Player i then chooses an action $y_i^g(s_i, \widehat{m}^g_{N_i^{-1}(g),i})$ to maximize his ex-post expected payoff,

$$\mathbb{E}_i\left(-\sum_{j=1}^{n}(y_j^g - S - b_i)^2 \,\middle|\, s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right),$$

where the subscript i in the expectation operator $\mathbb{E}_i$ signifies that player i uses his prior of Beta distribution with parameters (α_i, β_i). This means that the agent optimally picks

$$\begin{aligned} y_i^g(s_i, \widehat{m}_{N_i^{-1}(g),i}) &= \arg\max_{y_i^g}\left\{\mathbb{E}_i\left(-(y_i^g - S - b_i)^2 \,\middle|\, s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right)\right\} \\ &= b_i + \mathbb{E}_i\left(S \,|\, s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right). \end{aligned}$$

Because the communication is assumed to be either truthful or uninformative, the information set $\left(s_i, \widehat{m}^g_{N_i^{-1}(g),i}\right)$ can be equivalently represented as the set of revealed signals. Specifically, assume that player i gets to know k signals summarized in a set s_R. The unknown $D-k$ signals are denoted as a set s_{-R}. Using this notation, i's optimal action can be written as

$$\begin{aligned} y_i^g(s_R) &= b_i + \sum_{s_d \in s_R} s_d + \mathbb{E}_i\left(\sum_{s_d \in s_{-R}} s_d \middle| s_R\right) \\ &= b_i + \sum_{s_d \in s_R} s_d + (D-k)\mathbb{E}_i\left(s \,|\, s_R\right), \end{aligned}$$

where the second equality holds because the aspects are identically distributed. Thus, the optimal action is the sum of the preference bias, the known aspects, and the prediction of the unknown part of the state.

The binary nature of the signal ensures that $\mathbb{E}_i\left(s \,|\, s_R\right) = \mathbb{E}_i\left(\theta \,|\, s_R\right)$. Assume that l signals in s_R are 1. A straightforward algebra implies that player i's posterior is Beta distribution with parameters $(\alpha_i + l, \beta_i + k - l)$:

$$f_i(\theta|l,k) = \frac{1}{B(\alpha_i + l, \beta_i + k - l)}\theta^{\alpha_i + l - 1}(1-\theta)^{\beta_i + k - l - 1}.$$

The expected value of θ then is $\mathbb{E}_i(\theta|l,k) = \frac{\alpha_i + l}{\alpha_i + \beta_i + k} = \frac{\alpha_i + l}{\gamma + k}$ and the optimal action becomes

$$y_i^g(s_R) = b_i + \sum_{s_d \in s_R} s_d + (D-k)\frac{\alpha_i + l}{\gamma + k}. \tag{2.1}$$

Assume now that player i gets to know an additional signal s. The net effect on the optimal action $y_i^g(s_R)$ consists of two parts: the first is transferring the aspect from the predicted to the known part of the state; the second is improving the prediction of the unknown part with the additional information. The following lemma states that for any $j \in N$ the magnitude of the effect from player j's ex-ante perspective is lower the more aspects that are revealed to player i, but remains bounded away from 0.

Lemma 6. *Consider player i who obtains k aspects s_R. Then the magnitude of the effect from an additional signal s on i's action from player j ex-ante perspective $\mathbb{E}_j(y_i(s_R) - y_i(s_R, s))^2$ decreases with k, but remains greater than $\frac{\alpha_j \beta_j}{(\gamma+1)^2}$ for all k and D.*

Proof. See Appendix 2.7. □

Intuitively, the common underlying economic environment insures that the expected magnitude of the impact from an extra signal decreases, while the additive nature of the state of the world guarantees that it remains bounded from 0. Indeed, learning an additional signal improves the information about the underlying economic environment θ by less and less the more signals are already known. At the same time, the magnitude of the effect from an additional signal remains strictly bounded away from zero because the state of the world is additive in aspects that have a non-zero variance for any $\theta \in (0, 1)$. Importantly, it is this sufficiently slow rate at which the impact of an additional signal decreases—and not the particular assumptions behind the nature of the state of the world—that is one of the crucial elements for the main results.[7]

2.3.2 Equilibrium networks

Assume that the strategy profile (μ^g, y^g) is such that the communication network g is truthful: communication through each link of g leads to a perfect signal revelation. Define $k_j(g)$ to be the number of other players who report truthfully to j via either channel, and refer to it as the *in-degree* of player j.

In order to characterize the structure of equilibrium networks, I study how the ex-ante expected benefits from signal revelation and the interim credibility of player i in communicating to player j depend on the number of signals revealed to j by other players. Assume that agent i has a link to agent j and, apart from player i's message, player j gets to know k_j truthful aspects: 1 aspect player j obtains himself and $k_j - 1$ aspects he infers from other players' messages, excluding i. Denote the set of known k_j aspects as $s_R \in \{0, 1\}^{k_j}$ and the set of other aspects, excluding s_i, as $s_{-R} \in \{0, 1\}^{D-k_j-1}$.

I show that under either type of disagreement between the players, the two parties, i and j, would ex-ante prefer truthful revelation of s_i to player j. However, the credibility of communication and the way it is influenced by k_j crucially depend on the nature of disagreement between the players. In what follows, I consider each case of disagreement in turn.

[7] This pattern is different from [Galeotti *et al.*, 2011], in which the rate of decrease is high and not bounded from 0, when the number of revealed signals increases. It is also different from [Hagenbach and Koessler, 2010], in which the magnitude of the impact is constant, because each aspect is independently drawn from its individual distribution.

2.3.2.1 Conflicting preferences

Benefits of signal revelation. First, assume uninformative communication from player i to player j, which implies that player j conditions his action only on s_R and optimally chooses $y_j(s_R) = b_j + \sum_{s_d \in s_R} s_d + (D - k_j)\frac{\alpha+l}{\gamma+k_j}$. Consider the ex-ante expected input from player j into i's utility. Given the quadratic loss utility function, the input is comprised of an expected residual variance of the unknown part of the state S and a square of the preference divergence:

$$-\mathbb{E}\left[\operatorname{Var}\left(\sum_{s_d \in (s_i, s_{-R})} s_d | s_R\right)\right] - (b_j - b_i)^2 = -h(k_j) - (b_j - b_i)^2.$$

Next, if player i truthfully communicates his signal to player j, then j conditions his action on $k_j + 1$ signals (s_R, s_i), and the ex-ante expected input becomes $-h(k_j + 1) - (b_j - b_i)^2$. Thus, the expected benefit of learning an additional signal s_i is in reducing the expected residual variance, $h(k_j) - h(k_j + 1) > 0$ (because of the smaller unknown part of the state S and greater information). Clearly, player j himself—as well as any other player from N—enjoys the same ex-ante benefit of $h(k_j) - h(k_j + 1)$ independently of the bias difference.[8] The following lemma presents the exact expression for this benefit and establishes that it is a decreasing function of k_j.

Lemma 7. *Fix a truthful network g and consider player j with the in-degree $k_j - 1 < n - 1$. If j learns one extra signal s_i, then each player $l \in N$ derives the ex-ante expected benefit of $h(k_j) - h(k_j + 1) > 0$ that decreases with k_j, where*

$$h(k_j) = \frac{\alpha\beta(\gamma + D)(D - k_j)}{\gamma(\gamma + 1)(\gamma + k_j)}.$$

Proof. See Appendix 2.7. □

Negative externality effect. While Lemma 7 states that the truthful reporting of s_i is desired by both parties i and j from an ex-ante perspective, it might not be interim credible. To study the incentive to report truthfully, assume that player j believes i's message, i.e., j assigns the probability 1 that $s_i = m_{ij}$. If i reveals his private signal, $m_{ij} = s_i$, then player j optimally picks

[8] That all players derive the same benefit from player j learning an additional signal is due to the assumption that the inputs from all players are equally weighted in the payoff function.

an action $y_j(s_R, s_i) = b_j + \sum_{s_d \in s_R} s_d + s_i + (D - k_j - 1)\frac{\alpha+l+s_i}{\gamma+k_j+1}$; if i misreports and sends $m_{ij} = 1 - s_i$, then j chooses $y_j(s_R, 1-s_i) = b_j + \sum_{s_d \in s_R} s_d + 1 - s_i + (D - k_j - 1)\frac{\alpha+l+1-s_i}{\gamma+k_j+1}$. Player i reveals his signal whenever it results in a greater interim expected payoff:

$$\sum_{s_R \in \{0,1\}^{k_j}, s_{-R} \in \{0,1\}^{D-k_j-1}} -\left[(y_j(s_R, s_i) - S - b_i)^2 - (y_j(s_R, 1-s_i) - S - b_i)^2\right] P(s_R, s_{-R}|s_i) \geq 0.$$

As I show in Theorem 6, this incentive compatibility constraint can be rewritten as

$$|b_j - b_i| \leq \frac{\gamma + D}{2(\gamma + k_j + 1)}. \tag{2.2}$$

While this constraint (2.2) is always satisfied when $|b_j - b_i| \leq 1/2$, it might fail to hold when the preference divergence is significant, $|b_j - b_i| > 1/2$. In the latter case, player i can be credible so long as player j doesn't get to know too many signals relative to the divergence in their preferences. To see the intuition behind this, recall that, by Lemma 6, the magnitude of the effect from an additional signal on j's action decreases with k_j. Thus, for sufficiently high k_j, the expected effect might become so small that i would prefer to lie in order to shift j's action closer towards i's preferred one. On the other hand, when k_j is quite low, the expected effect of an additional signal is quite big, in which case misreporting can change j's action by too much, making it undesirable.

The incentive compatibility constraint (2.2) corresponds to the congestion effect of [Galeotti *et al.*, 2011] and [Loginova, 2012b]. In this chapter, I refer to this effect as the *negative externality effect* of information transmission—greater information has a negative effect on further information accumulation by discouraging other players to report truthfully.[9] Clearly, in order for player j to receive truthful messages from k_j players in equilibrium, the incentive compatibility constraint (2.2) must be satisfied for each of those players. The following theorem provides the formal equilibrium characterization.

Theorem 6. *Consider a triple $\{g, (\mu^g, y^g)\}$ and assume that each element of y^g satisfies the optimality condition (2.1). Then $\{g, (\mu^g, y^g)\}$ forms an equilibrium if and only if the communication*

[9] Another natural way to think about the negative externality effect is to view the privately observed aspects as *substitutes*. Indeed, if player i reports to j truthfully, then some other player l might not be credible in communicating to player j. If, on the other hand, player i does not report to j, then player l might be able to transmit his private information truthfully. This means that in such a communication process, private signals of i and l act as substitutes.

network g is truthful, and for any player j with the in-degree $k_j = k_j(g)$

$$|b_j - b_i| \leq \frac{\gamma + D}{2(\gamma + k_j + 1)} \text{ for all } i \in N_j^{-1}(g) = \{i \in N : g_{ij} = 1\}.$$

Proof. See Appendix 2.7. □

2.3.2.2 Conflicting opinions

Benefits of signal revelation. Start from an uninformative communication from player i to player j, in which case player j optimally chooses $y_j(s_R) = \sum_{s_d \in s_R} s_d + (D - k_j)\frac{\alpha_i + l}{\gamma + k_j}$. The ex-ante expected input from player j into i's utility (from player i's perspective) is

$$\begin{aligned} & -\mathbb{E}_i\left[\mathbb{E}_j\left(\sum_{s_d \in (s_i, s_{-R})} s_d | s_R\right) - \mathbb{E}_i\left(\sum_{s_d \in (s_i, s_{-R})} s_d | s_R\right)\right]^2 - \mathbb{E}_i\left[\operatorname{Var}_i\left(\sum_{s_d \in (s_i, s_{-R})} s_d | s_R\right)\right] \\ = \quad & -A_{ij}(k_j) - B_i(k_j). \end{aligned}$$

Here the first term depends on the ex-post opinion divergence between players i and j given the information s_R, and the second term is the expected residual variance. The subscripts ij and i for A and B, respectively, signify the priors used in evaluating (parts of) these expressions.

When player j learns an additional signal s_i, both terms are reduced. Indeed, the greater information and the smaller unknown part of the state S imply that the expected residual variance $B_i(k_j)$ decreases. Regarding the first term $A_{ij}(k_j)$, the monotone likelihood ratio guarantees that players are "consistent" with each other in updating their priors. Hence, the term $A_{ij}(k_j)$ decreases, because player i expects the additional signal to "persuade" player j, such that they have a smaller expected divergence in their ex-post opinions. On the whole, player i expects to derive a benefit of $A_{ij}(k_j) + B_i(k_j) - A_{ij}(k_j + 1) - B_i(k_j + 1) > 0$, which is a decreasing function of k_j.

Analogously, some player $l \in N$ enjoys an expected benefit of $A_{lj}(k_j) + B_l(k_j) - A_{lj}(k_j + 1) - B_l(k_j + 1)$. In particular, player j himself expects to get the benefit only from reducing the residual variance (evaluated using player j's prior), $B_j(k_j) - B_j(k_j + 1)$. The exact expressions for the terms are given in the following lemma.

Lemma 8. *Fix a truthful network g and consider player j with the in-degree $k_j - 1 < n - 1$. If j learns one extra signal s_i, then player $l \in N$ from the ex-ante perspective expects to receive a benefit* *of* $A_{lj}(k_j) + B_l(k_j) - A_{lj}(k_j+1) - B_l(k_j+1) > 0$ *that decreases with* k_j*, where*

$$A_{lj}(k_j) = \frac{(\alpha_l - \alpha_j)^2(D-k_j)^2}{(\gamma+k_j)^2} \text{ and } B_l(k_j) = \frac{\alpha_l\beta_l(D+\gamma)(D-k_j)}{\gamma(\gamma+1)(\gamma+k_j)}.$$

Proof. See Appendix 2.7. □

Positive externality effect. To study the credibility of communication, assume that player j believes i's message. If player i reports truthfully, player j chooses $y_j(s_R, s_i) = \sum_{s_d \in s_R} s_d + s_i + (D - k_j - 1)\frac{\alpha_i + l + s_i}{\gamma + k_j + 1}$; if player i lies, then player j picks $y_{s_R, 1-s_i} = \sum_{s_d \in s_R} s_d + 1 - s_i + (D - k_j - 1)\frac{\alpha_i + l + 1 - s_i}{\gamma + k_j + 1}$. Player i will choose truthful reporting if and only if this induces a greater interim expected payoff:

$$\sum_{s_R \in \{0,1\}^{k_j}, s_{-R} \in \{0,1\}^{D-k_j-1}} -\left[(y_j(s_R, s_i) - S)^2 - (y_j(s_R, 1-s_i) - S)^2\right] P_i(s_R, s_{-R}|s_i) \geq 0.$$

Note, that player j uses his prior in determining his actions $y_j(s_R, s_i)$ and $y_j(s_R, 1-s_i)$, while player i uses his own prior in probability assessments $P_i(s_R, s_{-R}|s_i)$ when evaluating the expected payoffs.

Clearly, if $k_j = n - 1 = D - 1$, then reporting the last remaining signal s_i truthfully is incentive compatible because it eliminates any uncertainty about the state of the world. As a result, player j is able to match his action to the state S—an ideal outcome for each player $i \in N$. Consider now $k_j \leq n - 1 < D - 1$ and the baseline case where $\alpha_i + \beta_i = \gamma$, $i \in N$. The incentive compatibility constraint of truth-telling takes a simple form of:[10]

$$|\alpha_j - \alpha_i| \leq \frac{\gamma + D}{2(D - k_j - 1)}, \quad k_j < D - 1. \tag{2.3}$$

Contrary to the conflicting preferences case, under conflicting opinions, greater k_j relaxes the incentive compatibility constraint, so that it becomes easier for player i to report to player j truthfully the more signals player j gets to know from other players. The intuition behind this result is the following. As the number of signals that j gets to know increases, two things happen. First, player i expects the ex-post belief of player j to become more congruent (hence, the optimal action of player j to become closer to i's preferred one). Indeed, because player i considers other

[10] The derivation of this condition is contained in the proof of Theorem 7 presented in Appendix 2.7.

signals revealed to player j to be distributed according to i's interim belief, he expects player j to be "persuaded" and to adjust his ex-post belief in the "right" direction from i's point of view. Second, as Lemma 6 demonstrates, the effect of an additional signal on j's action decreases with k_j. However, the rate of decrease is sufficiently low, so that the effect of i's message on j's action remains significant enough to prevent player i from misreporting to player j, given that i expects j to become more congruent.

Thus, player i might be non-credible when player j doesn't have much information because player i is concerned that if player j has an extreme opinion, he will misuse the reported signal. On the contrary, player i can become credible when player j's information improves, because player i expects player j to be persuaded and to become able to properly take into account i's signal when choosing his action. I refer to this effect as the *positive externality effect* of information transmission, because greater information has a positive effect on further information accumulation by encouraging other players to report truthfully as well.[11] In any equilibrium network, for player j to have the in-degree of $k_j(g) > 0$, it must be the case that the incentive compatibility constraint (2.3) is satisfied for every player i reporting truthfully to j. The formal equilibrium characterization is provided on the theorem below.

Theorem 7. *Consider a triple $\{g, (\mu^g, y^g)\}$ and assume that the individual action strategies satisfy the optimality condition (2.1). Then $\{g, (\mu^g, y^g)\}$ forms an equilibrium if and only if the communication network g is truthful, and for any player j with the in-degree $k_j = k_j(g) < D - 1$*

$$|\alpha_j - \alpha_i| \leq \frac{\gamma + D}{2(D - k_j - 1)} \text{ for all } i \in N_j^{-1}(g) = \{i \in N : g_{ij} = 1\}.$$

Proof. See Appendix 2.7. □

It is worthwhile to mention that the positive externality effect is not an artifact of a particular assumption that $\alpha_i + \beta_i = \gamma$ for all $i \in N$. Rather, the positive externality effect is present in a general case as well. The corresponding incentive compatibility constraint has a more complicated form, which is presented and derived in the proof of Theorem 7 (see Appendix 2.7).

[11] In case of positive externality, it is natural to view the private aspects as *complements*. Indeed, if player i reports to j truthfully, then other player l might be credible in truthful reporting to j as well. On the other hand, if player i does not report to player j, then it might not be incentive compatible for l to transmit his private information truthfully. Thus, private signals of i and l act as complements.

Remark. In a particular case when the players collectively hold all the information about the state of the world, i.e., $n = D$, reporting the last missing signal truthfully is incentive compatible because it eliminates any uncertainty and disagreement. Thus, there always exists an equilibrium in which each player gets to know all aspects of the state, i.e., the communication network is complete.

2.4 Applications

In this section I study the implications of the nature of disagreement on communication patterns in the particular cases of two communities and replicas of a diverse group of people. In the two communities setting, I show that under conflicting preferences, the larger group tends to communicate to the smaller group; while under conflicting opinions, the opposite is more likely to arise. In case of replicas, increasing the number of players of the same type localizes communication under conflicting preferences; and intensifies cross-type communication under conflicting opinions.

2.4.1 Pairwise stable equilibria

The previous analysis shows that, under both types of conflict, each player i ex-ante prefers that he and any other player $j \in N$ learn more aspects of the state, but the respective communication may fail to be credible from the interim perspective. There may be multiple equilibria that induce different truthful networks, and so, as a natural refinement, I adapt the notion of pairwise stability commonly used in the networks literature (e.g., [Bala and Goyal, 2000], [Goyal, 2007], [Jackson, 2008], [Jackson and Wolinsky, 1996]). Specifically, call an equilibrium $\{g, (\mu^g, y^g)\}$ *pairwise stable*, if no two players can improve the communication pattern between them, while satisfying the interim truth-telling incentives and holding other strategies fixed.

Definition. An equilibrium $\{g, (\mu^g, y^g)\}$ is *pairwise stable* if for any $i, j \in N$, $g_{ij} = 0$ only if, holding other strategies fixed, i cannot be credible in reporting to j, assuming that j believes i's message.

To prove the existence and determine the characteristics of pairwise stable equilibria, define the *maximal equilibrium* as an equilibrium in which the truthful network has the maximal in-degrees across all equilibrium networks:[12]

[12] The concepts of pairwise stable and maximal equilibria are the same as considered in [Loginova, 2012b].

Definition. Equilibrium $\{g, (\mu^g, y^g)\}$ with the in-degrees $k_1 = k_1(g), ..., k_n = k_n(g)$ is *maximal* if for any other equilibrium network with in-degrees $k'_1, ..., k'_n$:

$$k_i \geq k'_i, \ i = 1, ..., n.$$

In turn, the in-degrees $k_1, ..., k_n$ are called *maximal in-degrees.*

The following lemma states that the set of pairwise stable equilibria is non-empty and, under conflicting preferences, is a subset of maximal equilibria. Under conflicting opinions, the maximal equilibrium is unique and is necessarily pairwise stable.

Lemma 9. *There exist a maximal and a pairwise stable equilibrium. Under conflicting preferences, any pairwise stable equilibrium is maximal. Under conflicting opinions, the unique maximal equilibrium is pairwise stable.*

Proof. See Appendix 2.7. □

Since each player benefits when any other player receives more signals, the ex-ante expected payoff of every player is the largest in the maximal equilibria. Clearly, the players are indifferent between the maximal equilibria, because the ex-ante expected payoffs depend only on the vector of in-degrees and not on the particular network structure (see Lemmas 7 and 8). Assume that the maximal in-degrees are $k_1, ..., k_n$. Then, the maximal equilibrium that is pairwise stable can, for example, be constructed in the following way. Let every player i receive truthful messages from k_i players who are the closest to i in their preference divergence $|b_j - b_i|$ (or opinions divergence $|\alpha_j - \alpha_i|$). Because closer preferences (or opinions) relax the truth-telling incentives, these players are credible. At the same time, due to the maximality of k_i, other players cannot report truthfully to player i.

Lemma 9 implies that, under conflicting preferences, any pairwise stable equilibrium generates the same set of in-degrees. These in-degrees are necessarily maximal and, by the negative externality effect, are neither too low nor too large. On the contrary, under conflicting opinions, players can have different in-degrees in different pairwise stable equilibria. Moreover, by the positive externality effect, these in-degrees are either sufficiently low or quite large (it is easy to be credible to a player with a high in-degree, and it is difficult to be truthful to a low in-degree player).[13]

Interested readers are referred to Appendix 2.6 for an additional discussion of the pairwise stable equilibria.

2.4.2 Two communities

A set of players consists of two homogenous communities (or groups), N_1 and N_2, with sizes n_1 and n_2, respectively, where $1 \leq n_1 < n_2$ and the total number of players is $n = n_1 + n_2$. Under conflicting preferences, group N_1 members have preference biases normalized to 0, while group N_2 members have biases of b. Under conflicting opinions, all individuals in N_1 agree on the same prior with parameters (α_1, β_1); all players in N_2 have the prior with (α_2, β_2), and I still maintain the assumption that $\alpha_i + \beta_i = \gamma,\ i = 1, 2$.

Conflicting preferences. Clearly, in any pairwise stable equilibrium, there is complete communication inside each group. In addition, members of the same group have equal in-degrees, because exactly the same players can report truthfully to them. Introduce notation similar to [Galeotti *et al.*, 2011] and [Loginova, 2012b]: denote by k_i the in-degree of an arbitrary player in group N_i. Further, $k_i = k_{ii} + k_{ij}$, where $k_{ii} = n_i - 1$ reflects the level of *intra-group communication*—the number of signals revealed by N_i members, and k_{ij} stands for the level of *cross-group communication*—the number of truthful messages received from members of the opposite community N_j.

By the negative externality effect, if members of a smaller group N_1 report truthfully to some members of N_2, then the pairwise stability implies that members of a larger group N_2 reveal their signals to some members of N_1. Thus, depending on the parameters, cross-group communication can take one of the following 3 forms (see cases 1-3 in Figure 2.1):

1. No cross-group communication, i.e., $k_{21} = k_{12} = 0$, whenever $\frac{\gamma+D}{2(\gamma+n_1+1)} < b$.

2. Communication from group N_2 to group N_1, i.e., $k_{12} > 0$ and $k_{21} = 0$, whenever $\frac{\gamma+D}{2(\gamma+n_2+1)} < b \leq \frac{\gamma+D}{2(\gamma+n_1+1)}$.

[13] In a particular case in which the players cumulatively possess all the relevant information, $D = n$, under conflicting opinions, the unique maximal equilibrium has a complete communication network. However, aside from the maximal equilibrium, there may also exist pairwise stable equilibria that fail to aggregate all of the information (some players in the communication networks have quite low in-degrees).

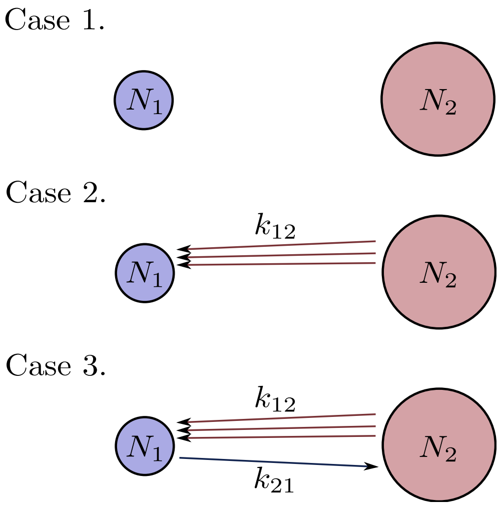


Figure 2.1: Communication networks of pairwise stable equilibria under conflicting preferences.

3. Cross-group communication, i.e., $k_{12} > 0$ and $k_{21} > 0$, whenever $b \leq \frac{\gamma+D}{2(\gamma+n_2+1)}$. In particular, when $b \leq \frac{\gamma+D}{2(\gamma+n)}$ the network is complete, i.e., $k_{12} = n_2$ and $k_{21} = n_1$.

Conflicting opinions. In any pairwise stable equilibrium, communication inside each group is necessarily complete, because members of the same community agree in their opinions. Regarding cross-group communication, note that, by the positive externality effect, if one player from N_j reports truthfully to a particular player in the opposite community N_i, then so do all of the other players in N_j. These observations immediately imply that the in-degree of a player in N_i can be either $n_i - 1$ or $n - 1$.

To simplify the characterization, I focus only on symmetric pairwise stable equilibria, i.e., equilibria in which players with the same priors have the same in-degrees. This implies that all players in group N_i have the same in-degree $k_i = k_{ii} + k_{ij}$, where k_{ij} is the level of cross-group communication and $k_{ii} = n_i - 1$ is the level of intra-group communication. Assuming that $D > n$, cross-group communication can take the following 4 forms (see cases 1-4 in Figure 2.2):[14]

1. No cross-group communication, i.e., $k_{21} = k_{12} = 0$, if $\frac{\gamma+D}{2(D-n_2-1)} < |\alpha_1 - \alpha_2|$.

2. Communication from group N_2 to group N_1, i.e., $k_{12} = n_2$ and $k_{21} = 0$, if $\frac{\gamma+D}{2(D-n_2-1)} < |\alpha_1 - \alpha_2| \leq \frac{\gamma+D}{2(D-n)}$.

[14] The case of $D = n$ differs in that the pairwise stable equilibrium with the complete communication network exists for all parameters.

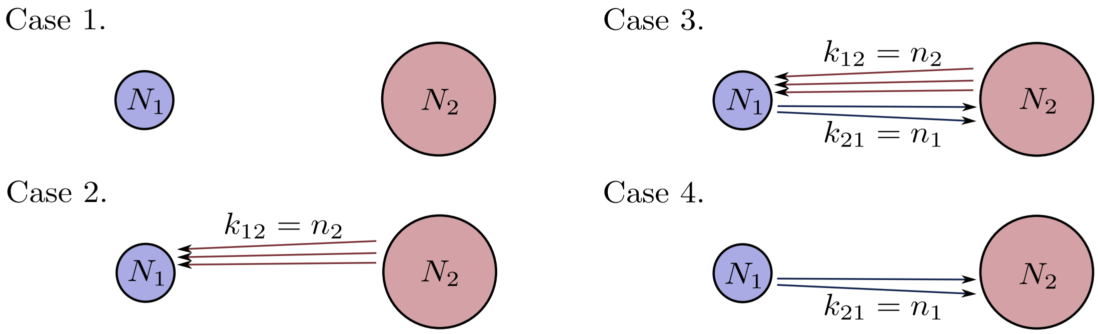


Figure 2.2: Communication networks of symmetric pairwise stable equilibria under conflicting opinions.

3. Cross-group communication (complete truthful network), i.e., $k_{12} = n_2$ and $k_{21} = n_1$, if $|\alpha_1 - \alpha_2| \leq \frac{\gamma+D}{2(D-n)}$.

4. Communication from group N_1 to group N_2, i.e., $k_{12} = 0$ and $k_{21} = n_1$, if $\frac{\gamma+D}{2(D-n_1-1)} < |\alpha_1 - \alpha_2| \leq \frac{\gamma+D}{2(D-n)}$.

Discussion. Comparing the communication patterns under conflicting preferences and under conflicting opinions, two things should be noted. First, under conflicting preferences, the preference divergence uniquely pins down the vector of in-degrees in a pairwise stable equilibrium. Under conflicting opinions, for some parameters, multiple in-degrees can be realized in different pairwise stable equilibria. Specifically, a member of N_i can receive truthful messages from either only N_i members (low in-degree of $n_i - 1$), or from all players (high in-degree of $n - 1$).

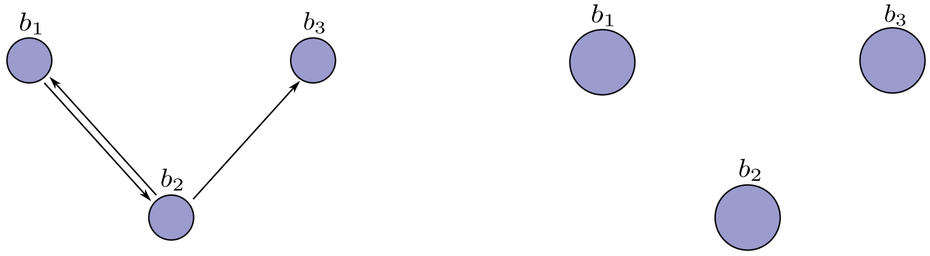
Second, under conflicting preferences, the information is more likely to flow from a larger group to a smaller one: either only members of a large group report to members of a small group, or members from opposite groups report to each other (with more information flow from the large group to the small one). Under conflicting opinions, the opposite communication pattern appears: for some parameters, there exists a pairwise stable equilibrium in which only members of a small group report to members of a large group, and not the other way around.

2.4.3 Replicas

Consider a set of $m \geq 2$ players and let the total number of aspects be D such that $Rm \leq D < (R+1)m$ for some $R \geq 1$. Under conflicting preferences, players are characterized by the preference biases $b_1, ..., b_m$ with the minimum difference of at least b: $\min_{i,j} |b_i - b_j| = b > 0$. Under conflicting opinions, players' prior beliefs are described by $(\alpha_1, \beta_1), ..., (\alpha_m, \beta_m)$ where $\min_{i,j} |\alpha_i - \alpha_j| = \alpha > 0$ and $\alpha_i + \beta_i = \gamma$ for all $i = 1, ..., m$. Call by *r-replica* the setting where there are rm players, r of each preference type b_i (respectively, belief type (α_i, β_i)), $i = 1, ..., m$. Assume that $r \leq R$, the total number of aspects in r-replica remains D and each player becomes privately informed of the respective aspect, so collectively all players in r-replica learn rm aspects. In what follows I focus on pairwise stable equilibria.

Conflicting preferences. Consider some r_1-replica, $1 \leq r_1 < R$. In any pairwise stable equilibrium, players with the same preference bias communicate truthfully with each other (meaning that the in-degree of each player is at least $r_1 - 1$). Now consider r_2-replica, where $r_1 < r_2 \leq R$. Clearly, in r_2-replica, intra-type communication is more intense than in r_1-replica: each player necessarily gets to know $r_2 - 1 > r_1 - 1$ aspects from players of the same type. By the negative externality effect, this threatens the credibility of communication between players of different types. The fact that some player j gains $r_2 - r_1$ new same-type communication links in r_2-replica might crowd out up to $r_2 - r_1$ cross-type communication links directed to j. As a result, in r_2-replica players have (weakly) greater in-degrees caused by more intra-type communication and (weakly) less cross-type communication.

Interestingly, regardless of the fact that in a higher-order r_2-replica the players collectively get to know more aspects than in r_1-replica, the individual awareness might fail to improve—the players' in-degrees in r_2-replica might remain the same as in r_1-replica. The following example illustrates this point. Let the total number of aspects $D = 6$ and the prior distribution of θ be uniform on the interval $[0, 1]$ that corresponds to Beta distribution with parameters $(1, 1)$ and $\gamma = 2$. Consider $m = 3$ preference types $b_1 = 0$, $b_2 = b$, $b_3 = 2b$, where $\frac{4}{5} < b \leq 1$. Such preference structure implies that truthful reporting to a player with the bias b_j is incentive compatible for a player with the bias $b_i \neq b_j$ if and only if $|b_i - b_j| = b$ and nobody else reports truthfully. Thus, in 1-replica, each player receives exactly one truthful message from a player with an adjacent bias (see Figure 2.3 for a possible communication network). In 2-replica, the players collectively hold all of the relevant information about the state of the world, but fail to aggregate it: each player communicates truthfully only with other same-type player; cross-type communication is necessarily empty (see Figure 2.3).



1-replica, some cross-type communication

2-replica, complete intra- and empty cross-type communication

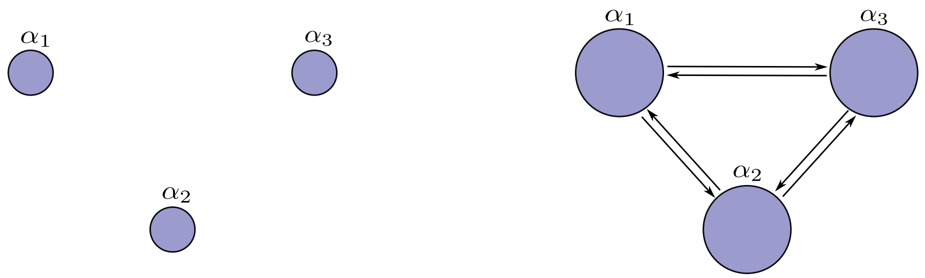
Figure 2.3: Communication networks in 1- and 2-replicas, conflicting preferences. Each circle corresponds to a group of players of particular type.

This example also demonstrates how increasing the number of same-type players can wipe out all cross-type communication. More generally, whenever $b > \frac{\gamma+D}{2(\gamma+r+1)} \geq \frac{1}{2}$, $D \geq rm$, then in any pairwise stable equilibrium of r-replica, there is a complete segregation of communication with respect to the types of preferences.

Conflicting opinions. Consider r_1- and r_2-replicas with $1 \leq r_1 < r_2 \leq R$. Because the intra-type communication is necessarily complete, players in r_2-replica receive $r_2 - r_1$ more truthful messages from the same-type players. The positive externality effect implies that some different-type players that could not be credible in r_1-replica might find truthful communication incentive compatible in r_2-replica. Thus, players have strictly greater maximal in-degrees in r_2-replica than in r_1-replica, due to (strictly) more intensive intra-type communication and (weakly) more intensive cross-type communication.

As an illustration of how players' awareness improves in a higher-order replica, consider the following example. Assume that the total number of aspects $D = 10$ and let $m = 3$ prior belief types be such that $\gamma = 2$, $\alpha_1 = 1$, $\alpha_2 = 1 + \alpha$ and $\alpha_3 = 1 + 2\alpha$, where $\frac{6}{7} < \alpha \leq 1$. Condition $\frac{6}{7} < \alpha$ insures that, in 1-replica, the only pairwise stable equilibrium has an empty communication network. At the same time, condition $\alpha \leq 1$ implies that, in 3-replica, the unique pairwise stable equilibrium has a complete communication network—within and across different types of players (see Figure 2.4).

This example shows how increasing the number of same-type players can induce complete available information aggregation. More generally, a soon as $\max_{i,j} |\alpha_i - \alpha_j| \leq \frac{\gamma+D}{2(\gamma+rm)}$, $D \geq rm$,



1-replica, empty cross-type communication

3-replica, complete intra- and cross-type communication

Figure 2.4: Communication networks in 1- and 2-replicas, conflicting opinions. Each circle corresponds to a group of players of particular type.

then the maximal pairwise stable equilibrium of r-replica has a complete communication network.

Discussion. Does introduction of new players who are privately informed about other aspects help agents to accumulate greater information about the state of the world? As the above analysis illustrates, the answer crucially depends on the nature of disagreement. Under conflicting preferences, replicating the set of players curbs cross-type communication and can lead to a segregation of communication according to the players' types. As a result, it does not necessarily improve the aggregation of the available information. Under conflicting opinions, on the contrary, adding new informed players boosts cross-type communication, which necessarily leads to greater information aggregation.

2.5 Conclusion

In this chapter, I characterize communication patterns when individuals diverge in their preferences or in their opinions. While, for any type of conflict, truthful communication is always beneficial from the ex-ante perspective, there is a credibility issue at the interim stage. In particular, I find that under conflicting preferences, information transmission exhibits a negative externality effect: the incentive of individual i to report truthfully to individual j deteriorates with the number of aspects reported to j. Intuitively, as the information of agent j expands, the impact of an additional aspect lessens; hence, agent i might find it optimal to misreport in order to shift j's action closer to i's preferred action. In contrast, under conflicting priors, information transmission displays a positive externality effect: the credibility of agent i improves with the number of aspects reported to agent j. Intuitively, as agent j becomes more informed, agent i expects j's belief to become more congruent with agent i's, which improves agent i's incentive to report truthfully.

As one application, I consider the case of two communities and show that, under conflicting preferences, the information is likely to flow from a larger community to a smaller one; while, under conflicting opinions, a smaller community is more likely to report to a larger one. As another application, I study replicas of a diverse group of people and demonstrate that, under preference conflict, replicating the set of agents curbs cross-type communication and can lead to a segregation of communication according to the agent' types. As a result, adding agents with new private information does not necessarily improve the aggregation of information. In contrast, under opinion conflict, replicating the set of individuals, boosts cross-type communication, which necessarily leads to greater information aggregation.

2.6 Appendix: Pairwise stable equilibria

An equilibrium $\{g, (\mu^g, y^g)\}$ is *pairwise stable*, if no pair of players can improve the communication pattern between them and increase their ex-ante expected utilities, while satisfying the truth-telling constraints and keeping other strategies fixed. To better understand how the communication pattern can be improved, consider the possible cases of information transmission from player i to some player j with in-degree k_j in g. If player i reports truthfully to player j, $g_{ij} = 1$, then there is no way to improve. If player i does not report informatively to player j, $g_{ij} = 0$, then it is possible to increase the ex-ante expected payoffs by inducing the truthful reporting through the link ij, if it is interim incentive compatible. Otherwise, there is no way of improvement. Given these alternatives, pairwise stability is formally defined as

Definition. An equilibrium $\{g, (\mu^g, y^g)\}$ is *pairwise stable* if for any $i, j \in N$, $g_{ij} = 0$ only if, holding other strategies fixed, i cannot be credible in reporting to j, assuming that j believes i's message.

Maximality and pairwise stability. Note, that the sets of maximal and pairwise stable equilibria might not coincide. In particular, under conflicting preferences, there might be maximal equilibria that are not pairwise stable. For example, let the prior distribution of θ be uniform on $[0, 1]$. Consider 3 players with the preference biases $b_1 = b_2 = 0$, $b_3 \in (\frac{2+D}{10}; \frac{2+D}{8}]$. Then there are several maximal equilibria that generate the in-degrees $k_1 = k_2 = k_3 = 1$. The examples of such truthful networks are: (1) $g_{12} = g_{21} = g_{13} = 1$, $g_{23} = g_{31} = g_{32} = 0$, and (2) $g_{12} = g_{23} = g_{31} = 1$, $g_{13} = g_{32} = g_{21} = 0$. It is easy to see that the first communication network corresponds to a pairwise stable equilibrium. In contrast, the second communication network doesn't correspond to a pairwise stable equilibrium, because players 1 and 2, who agree in their preferences, would deviate and induce truthful communication through the soft link 21.[15]

Similarly, under conflicting opinions, there exist pairwise stable equilibria that are not maximal. For such examples, see the equilibria discussed in the two communities setting.

Symmetry. Under conflicting preference, pairwise stability immediately implies that any two players i and j with the same preference biases must be treated in a symmetric way, i.e., they

[15] This example coincides with the one considered in [Loginova, 2012b]

communicate truthfully with each other and receive the same number of truthful messages from other players. Under conflicting opinions, players with the same priors might not be treated symmetrically in a pairwise stable equilibrium; but they are necessarily symmetric in any maximal equilibrium. In particular, they communicate truthfully with each other, get truthful reports from the same set of players and reveal their signals to the same set of players (by the positive externality effect).

2.7 Appendix: Proofs

Proof of Lemma 6. Let l denote the number of signals 1 in s_R and consider the effect on i's action of an additional signal s. After some algebra, the this effect can be expressed as

$$\begin{aligned} y_i(s_R) - y_i(s_R, s) &= -s + (D-k)\mathbb{E}_i(\theta|s_R) - (D-k-1)\mathbb{E}_i(\theta|s_R, s) \\ &= -s + (D-k)\frac{\alpha_i + l}{\gamma + k} - (D-k-1)\frac{\alpha_i + l + s}{\gamma + k + 1} \\ &= \frac{\gamma + D}{\gamma + k + 1}\left[-s + \frac{\alpha_i + l}{\gamma + k}\right]. \end{aligned}$$

Thus, the expected magnitude of the effect from an additional signal is

$$\begin{aligned} \mathbb{E}_j(y_i(s_R) - y_i(s_R, s))^2 &= \frac{(\gamma + D^2)}{(\gamma + k + 1)^2}\mathbb{E}_j\left(s - \mathbb{E}_i(\theta|s_R)\right)^2 \\ &= \frac{(\gamma + D^2)}{(\gamma + k + 1)^2}\left[\mathbb{E}_j\left(s - \mathbb{E}_j(\theta|s_R)\right)^2 + \mathbb{E}_j\left(\mathbb{E}_j(\theta|s_R) - \mathbb{E}_i(\theta|s_R)\right)^2\right]. \end{aligned}$$

In this expression,

$$\begin{aligned} \mathbb{E}_j\left(s - \mathbb{E}_j(\theta|s_R)\right)^2 &= \mathbb{E}_j[s^2] - \mathbb{E}_j[\mathbb{E}_j(\theta|s_R)^2] \\ &= \frac{\alpha_j}{\gamma} - \frac{1}{(\gamma + k)^2}\mathbb{E}_j(\alpha_j + l)^2. \end{aligned}$$

From the proof of Lemma 7 below, $\mathbb{E}_j(\alpha_j + l)^2 = \frac{\alpha_j(\gamma+k)}{\gamma(\gamma+1)}[k(\alpha_j + 1) + \alpha_j(\gamma + 1)]$. This implies that

$$\begin{aligned} \mathbb{E}_j\left(s - \mathbb{E}_j(\theta|s_R)\right)^2 &= \frac{\alpha_j}{\gamma} - \frac{\alpha_j}{(\gamma + k)\gamma(\gamma + 1)}[k(\alpha_j + 1) + \alpha_j(\gamma + 1)] \\ &= \frac{\alpha_j\beta_j(\gamma + k + 1)}{(\gamma + k)\gamma(\gamma + 1)}. \end{aligned}$$

Consider now the term $\mathbb{E}_j\left(\mathbb{E}_j(\theta|s_R) - \mathbb{E}_i(\theta|s_R)\right)^2$:

$$\mathbb{E}_j\left(\mathbb{E}_j(\theta|s_R) - \mathbb{E}_i(\theta|s_R)\right)^2 = \mathbb{E}_j\left(\frac{\alpha_j + l}{\gamma + k} - \frac{\alpha_i + l}{\gamma + k}\right)^2 = \frac{(\alpha_j - \alpha_i)^2}{(\gamma + k)^2}.$$

As a result, the expected magnitude is

$$\mathbb{E}_j(y_i(s_R) - y_i(s_R, s))^2 = \frac{(\gamma + D)^2}{(\gamma + k + 1)^2}\left[\frac{\alpha_j\beta_j(\gamma + k + 1)}{(\gamma + k)\gamma(\gamma + 1)} + \frac{(\alpha_j - \alpha_i)^2}{(\gamma + k)^2}\right].$$

It is straightforward to see that the magnitude of the effect from an additional signal on i's action is a decreasing function of k. For the last aspect, i.e., $k = D - 1$, it boils down to

$$\mathbb{E}_j(y_i(s_R) - y_i(s_R, s))^2 = \frac{\alpha_j\beta_j(\gamma + D)}{(\gamma + D - 1)\gamma(\gamma + 1)} + \frac{(\alpha_j - \alpha_i)^2}{(\gamma + D - 1)^2},$$

which exceeds $\frac{\alpha_j\beta_j}{\gamma(\gamma+1)}$ for any D.

Note that under conflicting preferences, the players agree on the common prior and the expected magnitude is just $\mathbb{E}_j(y_i(s_R) - y_i(s_R, s))^2 = \frac{(\gamma+D)^2}{(\gamma+k+1)^2} \cdot \frac{\alpha_j\beta_j(\gamma+k+1)}{(\gamma+k)\gamma(\gamma+1)}$. In case of conflicting opinions, there is an additional decreasing in k term $\frac{(\gamma+D)^2}{(\gamma+k+1)^2} \cdot \frac{(\alpha_j-\alpha_i)^2}{(\gamma+k)^2}$, which corresponds to the fact that the additional signal in expectation makes the posteriors of players i and j closer to each other. Indeed, because the condition $\alpha_i + \beta_i = \gamma$, for all $i \in N$, implies the monotone likelihood ratio for prior distributions, the players interpret the signals and update their beliefs consistently with each other. **QED.**

Proof of Lemma 7. Consider player j who learns information $s_R \in \{0, 1\}^{k_j}$ ($k_j - 1$ signals coming from other players excluding player i, 1 signal j receives himself) and optimally chooses action $y_j(s_R) = b_j + \sum_{s_d \in s_R} s_d + \mathbb{E}(\sum_{s_d \in (s_i, s_{-R})} s_d | s_R)$. An ex-ante expected input from player j

into i's utility is then

$$\begin{aligned}
& -\sum_{s_R \in \{0,1\}^{k_j}} \sum_{(s_i,s_{-R}) \in \{0,1\}^{D-k_j}} (y_j(s_R) - S - b_i)^2 P(s_R, s_i, s_{-R}) \\
= & -\sum_{s_R} \sum_{(s_i,s_{-R})} \left[b_j - b_i + \mathbb{E}\left(\sum_{s_d \in (s_i,s_{-R})} s_d | s_R \right) - \sum_{s_d \in (s_i,s_{-R})} s_d \right]^2 P(s_R, s_i, s_{-R}) \\
= & -(b_j - b_i)^2 - 2(b_j - b_i) \sum_{s_R} \sum_{(s_i,s_{-R})} \left[\mathbb{E}\left(\sum_{s_d \in (s_i,s_{-R})} s_d | s_R \right) - \sum_{s_d \in (s_i,s_{-R})} s_d \right] P(s_R, s_i, s_{-R}) \\
& -\sum_{s_R} \sum_{(s_i,s_{-R})} \left[\mathbb{E}\left(\sum_{s_d \in (s_i,s_{-R})} s_d | s_R \right) - \sum_{s_d \in (s_i,s_{-R})} s_d \right]^2 P(s_R, s_i, s_{-R}).
\end{aligned}$$

It is easy to see that the second term in this sum is zero. The third term is an expected residual variance $h(k_j) = \mathbb{E}\left[\text{Var}\left(\sum_{s_d \in (s_i,s_{-R})} s_d | s_R \right) \right]$ and can be rewritten as

$$\underbrace{-\sum_{s_R} \left[\mathbb{E}\left(\sum_{s_d \in (s_i,s_{-R})} s_d | s_R \right) \right]^2 P(s_R)}_{A} + \underbrace{\sum_{(s_i,s_{-R})} \left[\sum_{s_d \in (s_i,s_{-R})} s_d \right]^2 P(s_i, s_{-R})}_{B}.$$

Calculate the first term A, denoting l to be the number of 1s in s_R:

$$\begin{aligned}
A &= -\sum_{s_R} \left[(D - k_j) \mathbb{E}(\theta | s_R) \right]^2 P(s_R) = -\frac{(D-k_j)^2}{(\gamma + k_j)^2} \sum_{s_R} (\alpha + l)^2 P(s_R) \\
&= -\frac{(D-k_j)^2}{(\gamma + k_j)^2} \left[\alpha^2 + 2\alpha \sum_{s_R} l P(s_R) + \sum_{s_R} l^2 P(s_R) \right].
\end{aligned}$$

In this expression

$$\sum_{s_R} l P(s_R) = k_j \mathbb{E}(s_1) = \frac{k_j \alpha}{\gamma},$$

$$\sum_{s_R} l^2 P(s_R) = \int_0^1 \left(\sum_{s_R} l^2 P(s_R | \theta) \right) f(\theta) d\theta.$$

Since l is equal to the sum of signals in s_R, signals s_j are identically distributed and independent conditionally on θ, the term inside the integral can be rewritten as

$$\begin{aligned}\sum_{s_R} l^2 P(s_R|\theta) &= \mathbb{E}(l^2|\theta) = \mathrm{Var}(l|\theta) + (\mathbb{E}(l|\theta))^2 \\ &= k_j\theta(1-\theta) + k_j^2\theta^2.\end{aligned}$$

Taking the integral,

$$\begin{aligned}\sum_{s_R} l^2 P(s_R) &= \frac{k_j}{B(\alpha,\beta)}B(\alpha+1,\beta+1) + \frac{k_j^2}{B(\alpha,\beta)}B(\alpha+2,\beta) \\ &= \frac{k_j}{\gamma(\gamma+1)}[\alpha\beta + k_j\alpha(\alpha+1)].\end{aligned}$$

Substituting these to A yields

$$\begin{aligned}A &= -\frac{(D-k_j)^2}{(\gamma+k_j)^2}\left[\alpha^2 + 2\alpha\frac{k_j\alpha}{\gamma} + \frac{k_j(\alpha\beta + k_j\alpha(\alpha+1))}{\gamma(\gamma+1)}\right] \\ &= -\frac{(D-k_j)^2\alpha}{(\gamma+k_j)\gamma(\gamma+1)}\left[\alpha(\gamma+1) + k_j(\alpha+1)\right].\end{aligned}$$

Now consider the second term B, assuming that the number of 1s in (s_i, s_{-R}) is $\tilde{l}$:

$$B = \sum_{(s_i,s_{-R})} \tilde{l}^2 P(s_i, s_{-R}) = \frac{(D-k_j)}{\gamma(\gamma+1)}[\alpha\beta + (D-k_j)\alpha(\alpha+1)].$$

Then after some algebraic transformations, $h(k_j) = A + B$ becomes:

$$h(k_j) = \frac{\alpha\beta(\gamma+D)(D-k_j)}{\gamma(\gamma+1)(\gamma+k_j)}.$$

Finally, since the ex-ante expected input from player j into i's payoff is $-h(k_j) - (b_j - b_i)^2$, then the benefit from improving j's information by one additional signal is $h(k_j) - h(k_j+1)$. Because $h(k_j)$ is a positive, decreasing and convex function of k_j, the benefit exceeds 0 and decreases with k_j. **QED.**

Proof of Theorem 6. Consider the truthful network g and some player j who gets at least one truthful message, i.e., $N_j^{-1}(g) = \{i \in N : g_{ij} = 1\} \neq \emptyset$. It must be incentive compatible for every person $i \in N_j^{-1}(g)$ to report truthfully, given that player j believes their messages. Fix some $i \in N_j^{-1}(g)$ and let s_R be the set of k_j signals that player j gets to know himself and from other players apart from player i; denote $D - k_j - 1$ unknown signals excluding s_i as s_{-R}. If i reports truthfully, j optimally chooses $y_j(s_R, s_i)$; if player i misreports and sends $m_{ij} = 1 - s_i$, j picks the action $y_j(s_R, s_i)$. Player i reports truthfully his signal if and only if it generates a greater interim expected payoff to i compared to misreporting:

$$\sum_{s_R, s_{-R} \in \{0,1\}^{D-1}} - \left[(y_j(s_R, s_i) - S - b_i)^2 - (y_j(s_R, 1 - s_i) - S - b_i)^2 \right] P(s_R, s_{-R} | s_i) \geq 0.$$

This condition can be rewritten as

$$- \sum_{s_R, s_{-R}} \left[(y_j(s_R, s_i) - y_j(s_R, 1 - s_i))(y_j(s_R, s_i) + y_j(s_R, 1 - s_i) - 2S - 2b_i) \right] P(s_R, s_{-R} | s_i) \geq 0.$$

Assume that there are l signals 1 in s_R and recall that the actions $y_j(s_R, s_i)$ and $y_j(s_R, 1 - s_i)$ are given by (2.1), then the condition for truth-telling becomes

$$- \sum_{s_R, s_{-R}} P(s_R, s_{-R} | s_i) \left[2s_i - 1 + (D - k_j - 1) \left(\frac{\alpha + l + s_i}{\alpha + \beta + k_j + 1} - \frac{\alpha + l + 1 - s_i}{\alpha + \beta + k_j + 1} \right) \right] \times$$

$$\times \left[2b_j - 2b_i + 1 - 2s_i + (D - k_j - 1) \left(\frac{\alpha + l + s_i}{\alpha + \beta + k_j + 1} + \frac{\alpha + l + 1 - s_i}{\alpha + \beta + k_j + 1} \right) - 2 \sum_{s_d \in s_{-R}} s_d \right] \geq 0.$$

Using $P(s_R, s_{-R} | s_i) = P(s_{-R} | s_i, s_R) P(s_R | s_i)$, this can be simplified to

$$-(2s_i - 1) \frac{\alpha + \beta + D}{\alpha + \beta + k_j + 1} \times$$

$$\times \sum_{s_R} \left[2(b_j - b_i) + 1 - 2s_i + (D - k_j - 1) \frac{2\alpha + 2l + 1}{\alpha + \beta + k_j + 1} - 2A(s_i, s_R) \right] P(s_R | s_i) \geq 0,$$

where

$$\begin{aligned} A(s_i, s_R) &= \sum_{s_{-R}} \left(\sum_{s_d \in s_{-R}} s_d \right) P(s_{-R}|s_i, s_R) = \mathbb{E}\left(\sum_{s_d \in s_{-R}} s_d | s_i, s_R \right) \\ &= (D - k_j - 1)\mathbb{E}\left(\theta | s_i, s_R\right) = (D - k_j - 1)\frac{\alpha + l + s_i}{\alpha + \beta + k_j + 1}. \end{aligned}$$

After accounting for that and canceling the positive term $\frac{\alpha+\beta+D}{\alpha+\beta+k_j+1}$, the truth-telling condition becomes:

$$-(2s_i - 1)\sum_{s_R}\left[2(b_j - b_i) + 1 - 2s_i + (D - k_j - 1)\frac{1 - 2s_i}{\alpha + \beta + k_j + 1}\right] P(s_R|s_i) \geq 0,$$

$$-(2s_i - 1)\left[2(b_j - b_i) + (1 - 2s_i)\frac{\alpha + \beta + D}{\alpha + \beta + k_j + 1}\right] \geq 0.$$

If $s_i = 1$, the truth-telling condition becomes

$$b_j - b_i \leq \frac{\alpha + \beta + D}{2(\alpha + \beta + k_j + 1)}.$$

If $s_i = 0$, the truth-telling condition becomes

$$b_j - b_i \geq -\frac{\alpha + \beta + D}{2(\alpha + \beta + k_j + 1)}.$$

As a result,

$$|b_j - b_i| \leq \frac{\alpha + \beta + D}{2(\alpha + \beta + k_j + 1)}.$$

Since this condition must hold for every $i \in N_j^{-1}(g)$, this completes the proof of Theorem 6. **QED.**

Proof of Lemma 8. Assume that $\alpha_i + \beta_i = \alpha_j + \beta_j = \gamma$. If player j learns signals $s_R \in \{0,1\}^{k_j}$, he optimally chooses the action $y_j(s_R) = \sum_{s_d \in s_R} s_d + \mathbb{E}_j(\sum_{s_d \in (s_i, s_{-R})} s_d | s_R)$. This implies the following ex-ante expected input into i's utility

$$\begin{aligned}
& -\sum_{s_R\in\{0,1\}^{k_j}} \sum_{(s_i,s_{-R})\in\{0,1\}^{D-k_j}} (y_j(s_R)-S)^2 P_i(s_R,s_i,s_{-R}) \\
= & -\sum_{s_R}\sum_{(s_i,s_{-R})}\left[\mathbb{E}_j\left(\sum_{s_d\in(s_i,s_{-R})} s_d|s_R\right)-\sum_{s_d\in(s_i,s_{-R})} s_d\right]^2 P_i(s_R,s_i,s_{-R}) \\
= & -\sum_{s_R}\left[\mathbb{E}_j\left(\sum_{(s_i,s_{-R})} s_d|s_R\right)-\mathbb{E}_i\left(\sum_{(s_i,s_{-R})} s_d|s_R\right)\right]^2 P_i(s_R) \\
& -\sum_{s_R}\sum_{(s_i,s_{-R})}\left[\mathbb{E}_i\left(\sum_{s_d\in(s_i,s_{-R})} s_d|s_R\right)-\sum_{s_d\in(s_i,s_{-R})} s_d\right]^2 P_i(s_R,s_i,s_{-R}) \\
= & -A_{ij}(k_j)-B_i(k_j).
\end{aligned}$$

The second term of the sum is the expected residual variance, expression for which was derived in the proof of Lemma 7:

$$B_i(k_j)=\mathbb{E}_i\left[\mathrm{Var}_i\left(\sum_{s_d\in s_{-R}} s_d|s_R\right)\right]=\frac{\alpha_i\beta_i(\gamma+D)(D-k_j)}{\gamma(\gamma+1)(\gamma+k_j)}.$$

The first term from the sum can be rewritten as

$$\begin{aligned}
A_{ij}(k_j) &= (D-k_j)^2\sum_{s_R}\left[\mathbb{E}_j\left(\theta|s_R\right)-\mathbb{E}_i\left(\theta|s_R\right)\right]^2 P_i(s_R) \\
&= (D-k_j)^2\sum_{s_R}\left(\frac{\alpha_j+l}{\gamma+k_j}-\frac{\alpha_i+l}{\gamma+k_j}\right)^2 P_i(s_R) \\
&= (D-k_j)^2\sum_{s_R}\left(\frac{\alpha_j+l}{\gamma+k_j}-\frac{\alpha_i+l}{\gamma+k_j}\right)^2 P_i(s_R) \\
&= (D-k_j)^2\frac{(\alpha_i-\alpha_j)^2}{(\gamma+k_j)^2}.
\end{aligned}$$

Because $A_{ij}(k_j)$ and $B_i(k_j)$ are positive, decreasing and convex functions of k_j, the ex-ante expected benefit $A_{ij}(k_j)+B_i(k_j)-A_{ij}(k_j+1)-B_i(k_j+1)$ is positive and decreasing in k_j. **QED.**

Proof of Theorem 7. Consider the truthful network g and some player j who gets at least one truthful message. A necessary condition for player j to have the in-degree $k_j = k_j(g) < D-1$ is that for every person $i \in N_j^{-1}(g)$ it is incentive compatible to report truthfully, given that j believes the message. Fix some $i \in N_j^{-1}(g)$ and, as before, let s_R be the set of k_j signals that player j gets to know and s_{-R} be the set of $D-k_j-1$ unknown signals (excluding s_i). Assuming that player j believes i's message, let $y_j(s_R, s_i)$ and $y_j(s_R, 1-s_i)$ be j's actions when i reports truthfully or lies, respectively. Player i reports his signal s_i to j truthfully if and only if it results in greater ex-interim expected payoff compared to lying:

$$\sum_{s_R \in \{0,1\}^{k_j}, s_{-R} \in \{0,1\}^{D-k_j-1}} - \left[(y_j(s_R, s_i) - S)^2 - (y_j(s_R, 1-s_i) - S)^2\right] P_i(s_R, s_{-R}|s_i) \geq 0,$$

$$-\sum_{s_R, s_{-R}} \left[(y_j(s_R, s_i) - y_j(s_R, 1-s_i))(y_j(s_R, s_i) + y_j(s_R, 1-s_i) - 2S)\right] P_i(s_R, s_{-R}|s_i) \geq 0.$$

Let that the number of signals 1s in s_R be l, and use the expression (2.1) for optimal actions $y_j(s_R, s_i)$ and $y_j(s_R, 1-s_i)$:

$$-\sum_{s_R, s_{-R}} P_i(s_R, s_{-R}|s_i) \left[2s_i - 1 + (D-k_j-1)\left(\frac{\alpha_j + l + s_i}{\alpha_j + \beta_j + k_j + 1} - \frac{\alpha_j + l + 1 - s_i}{\alpha_j + \beta_j + k_j + 1}\right)\right] \times$$

$$\times \left[1 - 2s_i + (D-k_j-1)\left(\frac{\alpha_j + l + s_i}{\alpha_j + \beta_j + k_j + 1} + \frac{\alpha_j + l + 1 - s_i}{\alpha_j + \beta_j + k_j + 1}\right) - 2\sum_{s_d \in s_{-R}} s_d\right] \geq 0.$$

Expressing $P_i(s_R, s_{-R}|s_i) = P_i(s_{-R}|s_i, s_R)P_i(s_R|s_i)$, this can be simplified as

$$-(2s_i - 1)\frac{\alpha_j + \beta_j + D}{\alpha_j + \beta_j + k_j + 1} \times$$

$$\times \sum_{s_R} \left[1 - 2s_i + (D-k_j-1)\frac{2\alpha_j + 2l + 1}{\alpha_j + \beta_j + k_j + 1} - 2A(s_i, s_R)\right] P_i(s_R|s_i) \geq 0,$$

where

$$\begin{aligned} A(s_i, s_R) &= \sum_{s_{-R}} \left(\sum_{s_d \in s_{-R}} s_d \right) P_i(s_{-R}|s_i, s_R) = \mathbb{E}_i \left(\sum_{s_d \in s_{-R}} s_d | s_i, s_R \right) \\ &= (D - k_j - 1) \mathbb{E}_i \left(\theta | s_i, s_R \right) = (D - k_j - 1) \frac{\alpha_i + l + s_i}{\alpha_i + \beta_i + k_j + 1}. \end{aligned}$$

Denote $\alpha_i + \beta_i = \gamma_i$ for all i and cancel positive term $\frac{\gamma_j + D}{\gamma_j + k_j + 1}$:

$$-(2s_i - 1) \sum_{s_R} \left[1 - 2s_i + (D - k_j - 1) \left(\frac{2\alpha_j + 2l + 1}{\gamma_j + k_j + 1} - 2 \frac{\alpha_i + l + s_i}{\gamma_i + k_j + 1} \right) \right] P_i(s_R|s_i) \geq 0,$$

$$-(2s_i - 1) \left[1 - 2s_i + (D - k_j - 1) \left(\frac{2\alpha_j + 2B_i(k_j, s_i) + 1}{\gamma_j + k_j + 1} - 2 \frac{\alpha_i + B_i(k_j, s_i) + s_i}{\gamma_i + k_j + 1} \right) \right] \geq 0,$$

where $B_i(k_j, s_i)$ denotes ex-ante expected number of 1s in a set of k_j signals:

$$B_i(k_j, s_i) = \sum_{s_R} l P_i(s_R|s_i) = k_j \mathbb{E}_i(\theta|s_i) = k_j \frac{\alpha_i + s_i}{\gamma_i + 1}.$$

Substituting this into incentive condition,

$$-(2s_i - 1) \left[1 - 2s_i + (D - k_j - 1) \left(\frac{2\alpha_j + 2k_j \frac{\alpha_i + s_i}{\gamma_i + 1} + 1}{\gamma_j + k_j + 1} - 2 \frac{\alpha_i + k_j \frac{\alpha_i + s_i}{\gamma_i + 1} + s_i}{\gamma_i + k_j + 1} \right) \right] \geq 0,$$

$$-(2s_i - 1) \left[1 - 2s_i + (D - k_j - 1) \left(\frac{2\alpha_j + 2k_j \frac{\alpha_i + s_i}{\gamma_i + 1} + 1}{\gamma_j + k_j + 1} - 2 \frac{\alpha_i + s_i}{\gamma_i + 1} \right) \right] \geq 0,$$

which after some simplification becomes:

$$-(2s_i - 1) \left[1 - 2s_i + (D - k_j - 1) \frac{2\alpha_j(\gamma_i + 1) + \gamma_i + 1 - 2(\alpha_i + s_i)(\gamma_j + 1)}{(\gamma_i + 1)(\gamma_j + k_j + 1)} \right] \geq 0.$$

In case of $s_i = 1$ player i reveals the signal iff

$$-1 + (D - k_j - 1) \frac{2\alpha_j(\gamma_i + 1) + \gamma_i + 1 - 2(\alpha_i + 1)(\gamma_j + 1)}{(\gamma_i + 1)(\gamma_j + k_j + 1)} \leq 0,$$

$$-(\gamma_i + 1)(\gamma_j + k_j + 1) + (D - k_j - 1)(2\alpha_j(\gamma_i + 1) + \gamma_i + 1 - 2(\alpha_i + 1)(\gamma_j + 1)) \leq 0,$$

which after some algebraic transformations boils down to

$$\gamma_i - \gamma_j + \alpha_j(\gamma_i + 1) - \alpha_i(\gamma_j + 1) \leq \frac{(\gamma_i + 1)(\gamma_j + D)}{2(D - k_j - 1)}.$$

Subtracting $\frac{\gamma_i - \gamma_j}{2}$ from both sides,

$$\frac{\gamma_i - \gamma_j}{2} + \alpha_j(\gamma_i + 1) - \alpha_i(\gamma_j + 1) \leq \frac{(\gamma_i + 1)(\gamma_j + D)}{2(D - k_j - 1)} - \frac{\gamma_i - \gamma_j}{2}.$$

Consider now $s = 0$. The truth-telling condition in this case becomes

$$1 + (D - k_j - 1)\frac{2\alpha_j(\gamma_i + 1) + \gamma_i + 1 - 2\alpha_i(\gamma_j + 1)}{(\gamma_i + 1)(\gamma_j + k_j + 1)} \geq 0,$$

which can be simplified to

$$\alpha_j(\gamma_i + 1) - \alpha_i(\gamma_j + 1) \geq -\frac{(\gamma_i + 1)(\gamma_j + D)}{2(D - k_j - 1)}.$$

Adding $\frac{\gamma_i - \gamma_j}{2}$ to both sides,

$$\frac{\gamma_i - \gamma_j}{2} + \alpha_j(\gamma_i + 1) - \alpha_i(\gamma_j + 1) \geq -\frac{(\gamma_i + 1)(\gamma_j + D)}{2(D - k_j - 1)} + \frac{\gamma_i - \gamma_j}{2}.$$

As a result, the truth-telling condition for general signal s_i is

$$\left|\frac{\gamma_i - \gamma_j}{2} + \alpha_j(\gamma_i + 1) - \alpha_i(\gamma_j + 1)\right| \leq \frac{(\gamma_i + 1)(\gamma_j + D)}{2(D - k_j - 1)} - \frac{\gamma_i - \gamma_j}{2}.$$

In a particular case where the sum of prior's parameters is the same across all players, $\alpha_i + \beta_i = \gamma$ for any i. In this case truth-telling condition simplifies to:

$$|\alpha_j - \alpha_i| \leq \frac{\gamma + D}{2(D - k_j - 1)}.$$

QED.

Proof of Lemma 9. The logic of the proof (especially in the case of conflicting preferences) closely follows the one for Lemma 3 of [Loginova, 2012b]. I split the proof into three steps for each type of conflict:

Conflicting preferences, Step 1: Existence of a maximal equilibrium. Because the number of players and strategies is finite, the number of the pure strategy equilibria is also finite. Thus, there exists a well-defined set of numbers, $k_1, ..., k_n$, where k_i is the highest in-degree of player i that can arise in some equilibrium: for any equilibrium network g', $k_i \geq k'_i = k_i(g')$. Note that the in-degrees k_i and k_j, $i \neq j$, in principle, might be achieved in different equilibrium networks. To prove an existence of a maximal equilibrium, I need to show that the in-degrees $k_1, ..., k_n$ might be achieved in the same equilibrium, i.e., that there exists an equilibrium network g such that $k_i = k_i(g)$ for all i. In order to do this, I construct the equilibrium in the following way: for each $i \in N$ consider an equilibrium where k_i is achieved and let those (and only those) players who report to i truthfully in that equilibrium to report truthfully to i in the constructed equilibrium. Because the incentives to report truthfully depend only on the receiver's in-degree, the sender's and the receiver's preference biases, it is still incentive compatible for those players to report truthfully to i. Thus, this is, indeed, an equilibrium, and, by construction, it is maximal.

Conflicting preferences, Step 2: Maximality of a pairwise stable equilibrium. Consider some pairwise stable equilibrium and assume that it is not maximal. Then there exists player i whose in-degree in the equilibrium network is lower than his maximal in-degree. Fix some maximal equilibrium; then it must be the case that there is some agent j who reports truthfully to i in this maximal equilibrium, but not in the pairwise stable equilibrium. But then it is profitable for i and j to deviate and induce a truthful communication from j to i, which contradicts the pairwise stability. Hence, every pairwise stable equilibrium must be maximal.

Conflicting preferences, Step 3: Existence of a pairwise stable equilibrium. I illustrate this statement by constructing one of (possibly multiple) pairwise stable equilibria. For each $i \in N$ perform the following procedure: order other players $j \in N/\{i\}$ in the increasing absolute values of their preference divergence from i, $|b_j - b_i|$; let this order be $i_1, ..., i_{n-1}$. Consider the maximal in-degree of player i, k_i. If $k_i = 0$, then nobody can report truthfully to i in equilibrium. If $k_i > 0$, then let the closest k_i players report truthfully to i (clearly, it is incentive compatible, because closer biases relax the incentive condition of the truth-telling). Since k_i is the maximal possible in-degree of player i, players $i_{k_i+1}, ..., i_{n-1}$ cannot be credible in communicating to i; hence, set $g_{ji} = 0$ for these players.

Conflicting opinions, Step 1: Existence and uniqueness of a maximal equilibrium. The proof of existence is the same as in conflicting preferences case. I prove the uniqueness by contradiction. Assume that the set of maximal in-degrees is $k_1, ..., k_n$ and suppose that there are two different maximal equilibria with communication networks $g \neq g'$. Then there is a link ij in g that is not present in g'. Player i can truthfully communicate his signal to player j when j gets $k_j - 1$ truthful messages from other players in g. Hence, by the positive externality effect, i can be credible to j when j gets k_j truthful messages from other players in g'. Thus, there must exist an equilibrium where j's in-degree is $k_j + 1$, which contradicts the maximality of k_j. Hence, $g = g'$.

Conflicting opinions, Step 2: Pairwise stability of a maximal equilibrium. Consider some maximal equilibrium and assume that it is not pairwise stable. Then there exist players i and j such that $g_{ji} = 0$, but who can improve their communication pattern to truth-telling. By the positive externality effect, making communication through ji truthful, doesn't alter the credibility of communication through other links. Hence, there exists an equilibrium where player i receives more truthful signals, which contradicts the maximality condition. Thus, every maximal equilibrium must be pairwise stable.

Conflicting opinions, Step 3: Existence of a pairwise stable equilibrium. Steps 1 and 2 immediately imply the existence of a pairwise stable equilibrium. **QED.**

Chapter 3

Paternalism, libertarianism, and the nature of disagreement

Uliana Loginova and Petra Persson

Abstract

Regulation justified on the grounds that it prevents physical or moral self-harm is wide-spread, yet controversial. Some favor restriction of individual liberty; others that the government provide advice, but let each individual choose action. We model a benevolent authority's decision to constrain or inform a population of agents, each of whom must choose an action that may cause self-harm, when the authority has private, relevant information. We show that her decision to regulate an activity depends on whether she deems it a matter of preference or opinion. In the former case, she gives truthful advice and safeguards liberty; in the latter, she constrains liberty, believing that she acts in the population's interest. This contrasts with the commonly held view that politicians who wish to regulate simply place a lower value on individual liberty, and provides a precise prediction for what issues a benevolent authority regulates. We apply the model to regulatory issues that display a tension between individual liberty and coercion, e.g., safety mandates and assisted suicide prohibitions.

3.1 Introduction

> *Before the prayer warriors massed outside her window, before gavels pounded in six courts, before the Vatican issued a statement, before the president signed a midnight law and the Supreme Court turned its head, Terri Schiavo was just an ordinary girl...*

So begins the obituary of an ordinary woman with an extraordinary wish: to die.[1] When Schiavo's husband made his appeal to cease the treatment that kept her alive, a controversy broke out. The request divided the country, the world, even. Euthanasia – defined as "a deliberate intervention undertaken with the express intention of ending a life, to relieve intractable suffering" – is a criminal homicide in most jurisdictions, even if it is committed at the request of the patient ([Harris, 2001]). Yet, debates on the topic are ongoing in several countries, e.g. the U.S. and France: some deem it the right of an incurably ill to end her own suffering; others repudiate such requests, demanding that the government protect the requestor from herself.

When a restriction on individual freedom is justified solely on the grounds that it makes him better off, it represents an instance of *paternalism* ([Dworkin, 2010]). The word rings of benevolence – like a father (*lat.* Pater) disciplining his child out of love, the government constrains the populace in its best interest. But this raises a central question: Wouldn't a better informed government – a benevolent one, at least – just provide advice, and then let each individual decide for himself? This objection is neither novel nor ours; for centuries, *libertarians* have argued that individual liberty cannot legitimately be restricted to prevent self-harm ([Locke, 1689], [Mill, 1859]). This controversy is at the heart of contemporary debates that pit individual liberty against (supposed) safety: Should the government require drivers to wear seat-belts or motorcyclists to wear helmets? Should it forbid swimming at public beaches when lifeguards are not present, prevent women from taking heavy duty jobs, or require minors to have life-saving blood transfusions even when their religious beliefs forbid it? Today, lawmakers debate protecting minors against melanoma by banning the use of tanning beds (in Idaho), protecting the health of sex film workers by requiring the use of condoms (in California), and the Supreme Court is due to discuss the constitutionality of a mandate to purchase health insurance.[2]

[1] There was no living will. It was, however, affirmed as her wish in court after the testimony from eighteen witnesses on her end-of-life wishes ([Greer, 2000]). The obituary was written by [Benham, 2005].

[2] See [Lovett, 2012]; and [Yardley, 2012]. If health insurance would be mandated in order to alleviate externalities

We use economic theory to analyze when an authority restricts her subjects' autonomy, and when she instead simply gives advice. In particular, we start from an authority with no regard for her subjects, and ask how her behavior changes as her benevolence increases. In our framework, each individual must choose an action; for example, whether to wear a seat belt, or whether to buy health insurance. By assumption, this activity exerts no externality on others; this rules out non-paternalistic regulation, and makes the case for regulation as weak as possible.[3] The authority has private information about an exogenous state of the world that is relevant for the action choice; e.g., the risks associated with inaction. Even if the authority were to transfer this information to the individual, they would (still) disagree on the proper course of action, either because they have *different preferences* or because they have *different opinions* (priors) about the true state of the world. In the context of our example, given *the same* information about risks, (i) if the individual agrees with the authority on the risks, but he simply loves living on the edge, then they have different preferences; (ii) if they disagree on the interpretation of the information about risks, but are equally risk-loving, then they have different opinions. The authority has two tools at her disposal. She can issue a recommendation, and then give him the liberty to choose his course of action. Alternatively, she can coerce him, by mandating a certain action.

Our main insight is that the authority's choice between advising and coercing a subject crucially depends on (i) the nature of disagreement between them, and (ii) her regard for the subject. Under preference disagreement, a self-interested authority coerces the individual. An authority who is sufficiently altruistic, however, gives truthful advice, and then lets the individual decide; we say that she is *libertarian*. Under opinion disagreement, a self-interested authority communicates truthfully. An authority who is sufficiently altruistic, however, constrains the individual, against his will, believing that she acts in his best interest; we say that she is *paternalistic*. Thus, while restrictions on individual liberty can reflect both self-interest and benevolence, there is a clear dividing line between the two – the nature of disagreement.

from uninsured on public health, relatives, or insurance markets (e.g., through adverse selection), then this law would not be paternalistic, as it would restrict individual liberty in order to prevent harm to *others*. If, instead, it is justified by a worry that individuals who do not purchase health insurance fail to act in their own best interests, e.g., due to cognitive constraints ([Abaluck and Gruber, 2011], [Fang *et al.*, 2008], [Cutler and Zeckhauser, 2004]), then the law would be paternalistic.

[3] Regulation against behaviors that cause harm to third parties is non-paternalistic, and deemed legitimate also by libertarians ([Mill, 1859], [Locke, 1689]). [Hanson, 2003] considers non-paternalistic regulation; see the literature review for a discussion.

After an illustrative example in Section 3.2, where an authority is faced with a single individual, Section 3.3 develops our general framework, where she is faced with a continuum of individuals, forming a population. We start by considering an *advisor* who can issue recommendations, but not enact mandates. She privately obtains an imprecise piece of information s about an unknown state of the world $\theta \in \{0,1\}$, and sends a public message m, before each individual i chooses action $a_i \in \mathbb{R}$. Lying entails a cost $c \geq 0$; modulo this cost, talk is cheap (m is non-verifiable). The advisor's material payoff is maximized by each individual choosing $a_i = \theta$, and individual i's ideal action is $a_i = \theta + b_i$. The advisor's opinion (prior belief) about θ is $\pi_A = \Pr(\theta = 1)$; each individual's opinion is π_i. Under *preference disagreement*, the population's preferences are characterized by a distribution $f(b)$, but their opinions concur with that of the advisor, $\pi_i = \pi_A$ $\forall i$. Under *opinion disagreement*, the population's opinions are characterized by a distribution $g(\pi)$, but their preferences concur with those of the advisor, $b_i = 0$ $\forall i$.[4] For each type of disagreement, we analyze how the ability to sustain truthful communication depends on the advisor's *benevolence*, or altruism, modeled as the share of the individuals' material payoffs that she internalizes, $\varphi \geq 0$ ([Becker, 1974], [Camerer, 2003]).

Under preference disagreement, altruism improves communication. The stronger the advisor's altruism, the more she values that each individual gets to implement *his* preferred action. Hence, when φ increases, the action that the advisor wants each individual to choose approaches the individual's own preferred action. Higher altruism is thus akin to smaller preference disagreement ([Crawford and Sobel, 1982]); as disagreement lessens, truthful communication becomes attainable. By contrast, under conflicting priors, altruism can destroy communication. In this case, the advisor is convinced that her preferred action, given the signal s, maximizes both her own *and* each individual's expected welfare. Each individual, however, would interpret a truthfully revealed signal in light of his own prior, and choose a different action. Even though the advisor represents the median opinion in the population, she may believe that, on average, the individuals' are better off with the action choices they take when she lies. Lying then protects them from misinterpreting a truthful report. When φ increases, the advisor internalizes more of the disutility that she expects each individual to suffer from his (in her view) suboptimal choice of action. Paradoxically, a sufficiently

[4] On possible reasons for the existence and persistence of different priors see, e.g., [Aumann, 1976], [Tversky and Kahneman, 1974], [Acemoglu *et al.*, 2007], [Sethi and Yildiz, 2009]. Also see [Morris, 1995] for a discussion of the assumption of different priors in the economics literature.

altruistic advisor may therefore lie.[5]

We then consider an *authority* who, after observing the signal s, either can send a public message, or incur some cost $q \geq 0$ to mandate one action for all individuals. Under preference disagreement, enacting a mandate is unattractive to a benevolent authority, for two reasons: First, truthful communication can be sustained, so the authority has the ability to transfer all the relevant information to the individuals before they make their decisions. Second, if the authority lets each individual i choose his action, then i implements an action that is close to the action that the highly altruistic authority would want him to choose. Consequently, while a self-interested authority may enact a mandate, a benevolent authority instead communicates truthfully, and gives each individual the liberty to choose. By contrast, under opinion disagreement, mandating an action is attractive to the altruistic authority for two reasons: First, she may not be credible; if she would allow the individuals to choose their actions, they would thus base their choices on less information than she would do. Second, when the authority is sufficiently altruistic, she would enact a mandate even if truthful communication were possible, because she knows that the actions that the individuals would take differ from the action that she deems optimal for them. Consequently, while a self-interested authority may communicate truthfully, a benevolent authority instead mandates an action, believing that she acts in the population's interest.

While Section 3.4 considers an advisor who can send a single public message to the population, and an authority who (also) has the option to mandate a single action, Section 3.5 considers targeted advice or mandates. We start by asking whether an advisor is more credible when she can send a private signal to each individual than when she issues a single public message. We say that she is *more credible* when truthful communication can be sustained with a larger share of the population. Under preference disagreement, a self-interested advisor is more credible with private messages, but a sufficiently altruistic advisor is more credible with a public message. Intuitively, the advisor issues credible private advice to individuals with " sufficiently moderate" preferences, and as altruism strengthens, the set of such individuals increases. When altruism is strong enough to induce the advisor to communicate truthfully with the population's average-biased individual, she is also willing to send a truthful public signal. When altruism is too weak to sustain truthful

[5] This is particularly remarkable given that, before observing the signal s, the advisor's ex ante utility is higher in a truthful equilibrium. A strongly altruistic advisor is nevertheless non-credible, as she, ex interim, would prefer to lie – out of benevolence – given that the individual believes her report.

public communication, the advisor is more credible with private messages – as she still issues truthful private advice to some individuals. Otherwise, she is more credible with a public signal. By contrast, under opinion disagreement, a self-interested advisor is more credible with a public signal, whereas a sufficiently altruistic advisor is more credible with private messages. Intuitively, a weakly altruistic advisor may be truthful under public communication, even though she, in private, would issue false advice (to some, or all, individuals). For strong enough altruism, however, she may prefer to issue a false public message. Under both preference and opinion disagreement, an increase in altruism may thus yield a *credibility reversal*, where the relative credibility of public and private messages reverses. The direction of this reversal, however, depends on the nature of disagreement.

We then analyze how the authority's use of targeted mandates depends on her altruism. Under preference disagreement, any mandates are individual-specific, as the action that the authority wants a given individual to take partly reflects his individual preference. A more altruistic authority enacts fewer mandates, since the actions that she would mandate approaches those taken by the individuals in the absence of mandates. By contrast, under opinion disagreement, the authority would enact a single mandate for all individuals whose actions she restricts, as she deems the same action optimal for all individuals. This mandate is enacted for the individuals whose opinions are the most extreme; that is, for those who, if left to choose their own actions, would make the largest mistakes (in her view). For strong enough altruism, her mandate applies to all individuals whom she believes it would protect.

Equipped with these results, in Section 3.6 we return to the central issue that we set out to answer: In debates on regulation of activities that can cause self-harm, some (politicians) advocate a laissez-faire approach – to provide public information, but then let each individual make his own action choice – whereas others advocate constraining individual liberty, through mandates or prohibitions. What determines whether a politician advocates one or the other, and why do those who advocate restrictions in individual liberty deem such restrictions better than information provision? One commonly held view is that those who advocate restrictions simply place a lower value on individual liberty. This view, however, is hard to reconcile with the fact that some are in favor of regulation that others oppose, *and vice versa*.[6]

[6] Indeed, if, say, R places a higher value on individual liberty than does L, then all mandates supported by R

Our results offer an alternative, precise prediction: *A benevolent politician enacts regulation on issues that she believes is a matter of opinion, but leaves individuals to make their own choices on issues that she believes is a matter of preference.* To understand what issues a benevolent authority regulates, it thus suffices to ask what issues she deems a matter of of opinion. For example, a benevolent politician who believes that an individual's request for euthanasia reflects a true preference to die would resist a prohibition. A politician would instead advocate a prohibition if she believes that the requestor has an incorrect understanding of his own wish to die – e.g., if the politician deems suicide a sin that precludes the individual from afterlife benefits that he, if were aware of this, would not want to give up. Similarly, a politician who believes that failure to take up health insurance reflects preferences resists a mandate; a politician who instead fears that it reflects an underestimation of the true benefits of insurance favors a mandate.

Whenever the authority deems the individuals' actions to be driven by incorrect beliefs, restricting liberty is consistent with benevolence. In fact, simply transferring information is not enough – it is better, even necessary, to coerce. The distinction between preferences and opinions is thus imperative for whether intervention is socially beneficial or harmful. As an illustration of this insight, let us contrast the findings of several recent papers that departs from the empirical observation that take-up of health insurance cannot be explained by individual risk types alone. [Cohen and Einav, 2007] and [Einav *et al.*, 2010] attribute this unexplained variation in the demand for insurance to differences in individuals' *preferences for risk*. [Spinnewijn, 2012] suggests that another reason why risks do not fully explain the demand for insurance may be that individuals perceive their own risks incorrectly; in our terminology, that (incorrect) *opinions* drives the action choices. In the spirit of our general insights, [Spinnewijn, 2012] finds that when erroneous risk perceptions (opinions) play a larger role in insurance decisions relative to preferences for risk, the welfare gain of a universal mandate is higher.

3.1.1 Related Literature

This chapter is related to [Che and Kartik, 2009], who analyze a communication game with difference of opinion and contrast differences in opinion with differences in preferences. They show that difference of opinion between a principal and an agent can incentivize the agent to exert effort to

should be (a strict subset of the mandates) supported by L. We discuss this further in Section 3.6.

persuade the principal; however, differences in preferences cannot induce such a persuasion motive for effort provision. [Hirsch, 2011], [Van den Steen, 2006] and [Van den Steen, 2009] also analyze this mechanism. [Hirsch, 2011] illustrates how open disagreement in opinions between the principal and the agent creates a persuasion-based rationale for short-term deference: the principal might find it optimal to allow the agent to implement the agent's preferred policy. [Van den Steen, 2009] shows that a principal may incur a cost to alter the agent's beliefs in order to boost the agent's effort, in particular when this effort is critical for the project's outcome. Relatedly, [Van den Steen, 2006] shows that a principal may exploit the effect of differing opinions on the agent's effort by transferring decision rights to the agent. In our model, the uninformed party does not make an effort choice; hence, we do not rely on the mechanism that drives the results of these papers. Nevertheless, our result underscores one key message from [Che and Kartik, 2009]: the distinction between differing opinions and differing preferences may be crucial.[7]

The key mechanism in this chapter, instead, is altruism. This relates the paper to the emerging literature on communication and altruism. [Lee and Persson, 2011] analyze how friends transmit hard information to each other when sharing information dilutes its value. [Carlin *et al.*, 2010] show that an altruistic (unbiased) principal may share information with an uninformed set of agents so as to help improve their action choices, but that this may hamper the agents' individual incentives to acquire information. This, in turn, may lower aggregate welfare if one agent's privately acquired information can be accessed by others, either through social learning or on a market for information. We add to this literature by analyzing difference of opinion between the communicating parties. Moreover, we contrast communication with coercive measures to affect the individual's action choice. Intuitively, this corresponds to the distinction between *libertarian paternalism* ([Thaler and Sunstein, 2003], [Thaler and Sunstein, 2008], [Carlin *et al.*, 2010]) – whereby the government may recommend a default choice, but not constrain the individual's choice set – and (hard) *paternalism*.

Our analysis of targeted advice (but not mandates) relates to [Farrell and Gibbons, 1989] and [Goltsman and Pavlov, 2011] that compare the credibility of private and public messages of a non-altruistic advisor in settings with two individuals and preference disagreement. When we shut down altruism in our model, we replicate their result that the relative credibility of private versus

[7] Also related is [Loginova, 2012a], who shows that communication patterns in a network of individuals crucially depend on the nature of disagreement between the agents.

public messages depends on the preference distribution. We also obtain an analogous insight under opinion disagreement. In this sense, we extend their results to populations with more than two individuals, and to opinion disagreement. Further, while the relative credibility of different modes of communication varies in the absence of altruism, we show that introducing altruism eliminates or reduces this indeterminacy.

The most closely related paper in spirit is [Hanson, 2003], who models a regulator who is empowered to ban an activity or to warn the public about it. He shows that when a government is concerned about some market imperfection, cheap talk may not be credible, so the government may resort to a prohibition. Our mechanism is distinct in two ways. First, we study a setting without any externalities, to rule out any motives for regulation else than to prevent self-harm; in [Hanson, 2003], the government would never ban an activity in the absence of market imperfections. Second, and perhaps more fundamentally, in [Hanson, 2003], regulation is a solution to an information problem: if it were possible to issue a truthful recommendation, no regulation would be necessary. In our setting under differences of opinion, however, a benevolent government would regulate even if it were able to transfer its superior information to the individual. This is because it knows that the individual – who interprets the recommendation in light of his own distinct prior – will take an action that differs from the one that the government wishes him to take.

3.2 Example

Before introducing the general model, we present a simple setting with one individual that illustrates our key results and their underlying mechanisms. Appendix 3.11 presents a more general version of the model discussed here and proves all results. An individual (I, he) must choose an action $a \in \mathbb{R}$. His payoff from a depends on an unknown state of the world, $\theta \in \{0, 1\}$. Before he chooses, the individual's advisor (A, she) privately observes a signal s about the state, with precision $\Pr(s = \theta|\theta) \equiv \gamma \in (0.5, 1)$, and sends him a message $m \in \{0, 1\}$. If the advisor lies, $m \neq s$, she incurs a cost $c \geq 0$. The players' material (non-altruistic) payoffs are given by $u_A(a, \theta) = -(a - \theta)^2 - c\mathbf{I}_{\{m \neq s\}}$ and $u_I(a, \theta) = -(a - \theta - b)^2$, and their priors on the state of the world are given by $\Pr_i(\theta = 1) = \pi_i$, for $i \in \{I, A\}$.

In this standard model of communication, we allow the advisor to be altruistic: in addition to her own material payoff, she internalizes a share φ of the individual's payoff ([Becker, 1974]). Her utility is thus given by $U_A(a,\theta) = u_A(a,\theta) + \varphi u_I(a,\theta)$. An altruistic advisor not only cares about the action choice for her own sake, but also because the action affects the individual. We ask how communication is affected by the strength of the advisor's regard for the individual, φ, and how this depends on the nature of disagreement. We isolate two pure forms of disagreement. Under *preference disagreement*, the players' material payoffs from a differ ($b \neq 0$, w.l.o.g. $b > 0$), but they have a common prior, or opinion, on the state of the world ($\pi_A = \pi_I = 0.5$); under *opinion disagreement*, the players' material payoffs are identical ($b = 0$), but their opinions diverge ($\pi_I \neq \pi_A = 0.5$, w.l.o.g. $\pi_I > 0.5$). All of the above is common knowledge. A pure strategy of the advisor, $m(s)$, specifies, for each signal s, the message m that she sends. A pure strategy of the individual, $a_I(m)$, specifies, for each message m, the action that he takes. We solve the game for pure strategies Perfect Bayesian Equilibria.

Our first key result is that the impact of the advisor's altruism on communication depends crucially on the nature of disagreement. Under preference disagreement, truthful communication can arise if and only if altruism is *strong* enough. Under opinion disagreement, whenever altruism impacts communication, truthful communication can arise if and only if altruism is *weak* enough. The impact of the lying cost c, however, is independent of the nature of disagreement: raising c always improves the prospects to achieve truthful communication.

The logic driving this result is as follows. In any truthful equilibrium, the individual chooses $a_I(s) = p_I(s) + b$, where $p_I(s)$ is his posterior given the (truthfully reported) signal s. This action always exceeds the advisor's ideal action given the signal s, $a_A(s)$; under preference disagreement because $b > 0$, and under opinion disagreement because $\pi_I > \pi_A$. Consequently, the advisor always reports the signal $s = 0$ truthfully. When lying is costless ($c = 0$), she also reports the signal $s = 1$ truthfully if and only if (iff) her ideal action, $a_A(1)$, is closer to the action induced by a truthful message, $a_I(1)$, than to the action induced by a false message, $a_I(0)$.[8] A truthful equilibrium thus exists iff $a_I(1) - a_A(1) \leq a_A(1) - a_I(0)$, which can be written

$$2\left(a_I(1) - a_A(1)\right) - \tau \leq 0, \qquad (\mathrm{TT}_{c\,=\,0})$$

[8] This obtains because the advisor's loss function is monotonic in the distance between $a_A(1)$ and $a_I(1)$.

where $\tau \equiv a_I(1) - a_I(0)$ is a constant (given γ).

Under preference disagreement, the advisor's ideal action depends on the strength of altruism, $a_A(1) = a_I(1) - \frac{b}{1+\varphi}$. Intuitively, the stronger the advisor's regard for the individual, the more she values that he gets to implement *his* preferred action, $a_I(1)$. Thus, ($\text{TT}_{c\,=\,0}$) reduces to

$$2b - \tau(1+\varphi) \leq 0. \tag{$\text{TT}_{\text{pr},c\,=\,0}$}$$

Clearly, higher altruism is akin to a lower preference bias b ([Crawford and Sobel, 1982]); as φ increases, their disagreement lessens, and truthful communication becomes attainable. When lying is costly, $c > 0$, a truthful equilibrium exists iff

$$2b - \tau(1+\varphi) \leq \frac{c}{\tau}. \tag{$\text{TT}_{\text{pr},\ c\,>\,0}$}$$

A higher cost of lying and higher altruism thus both make truthful reporting more attractive.

Under opinion disagreement, even though the players' preferences are perfectly aligned, their preferred actions differ in any truthful equilibrium, as they interpret the signal s in light of their (different) priors. The advisor believes that $a_A(1)$ maximizes both her own *and* the individual's expected material payoff; consequently, $a_A(1)$ does not approach $a_I(1)$ as φ increases. Defining $K \equiv a_I(1) - a_A(1)$, we can thus write ($\text{TT}_{c\,=\,0}$) as

$$2K - \tau \leq 0. \tag{$\text{TT}_{\text{op},c\,=\,0}$}$$

When $c = 0$, the existence of a truth-telling equilibrium is independent of φ. The advisor reveals $s = 1$ when her ideal action $a_A(1)$ is closer to $a_I(1)$ than to $a_I(0)$. This occurs when opinion disagreement is minor; for example, when $\gamma = 0.6$, a truthful equilibrium exists for $\pi_I \leq 0.604$. When $c > 0$, a truthful equilibrium exists iff

$$(2K - \tau)(1+\varphi) \leq \frac{c}{\tau}, \tag{$\text{TT}_{\text{op},\ c\,>\,0}$}$$

The lying cost matters only if the advisor prefers to lie when $c = 0$, i.e., if $(2K - \tau) > 0$. Then, a higher lying cost makes truthful reporting more attractive, as under preference disagreement. Stronger altruism, however, makes truthful reporting *less* attractive. The logic behind this result is as follows. The lying cost induces the advisor to sometimes reveal $s = 1$ truthfully even when she believes that the action induced by a false message, $a_I(0)$, is better. This occurs if her benefit from lying – inducing $a_I(0)$ instead of $a_I(1)$ – is too small to outweigh the cost. Crucially, however, because the advisor believes that inducing the better action will benefit not only herself, *but also the individual*, her expected benefit from lying increases with φ. More precisely, when φ increases, the advisor internalizes more of the disutility that she expects the individual to suffer from his (in her view) suboptimal choice of action following a truthful report. Stronger altruism therefore makes lying more worthwhile. Importantly, whenever an increase in φ induces the advisor to switch from truth-telling to lying, she lies to protect *the individual* from the consequences of his misjudgment; the advisor's own material benefit from lying is too small to motivate the lie ($2K - \tau > 0$). In general, whenever the advisor believes that $a_I(0)$ dominates $a_I(1)$, a sufficiently altruistic advisor lies. In the context of our example, when $\pi_I = 0.85$, the advisor believes that $a_I(0)$ dominates $a_I(1)$, so she lies if $c = 0$. For $c = 0.06$, she reports truthfully so long $\varphi \leq 0.188$; when she cares more about the individual, she lies.

The second part of the chapter replaces the advisor with an authority (A, she). After observing the signal s, the authority can either behave like an advisor – send a message m to the individual, who then implements his preferred action, $a_I(m)$ – or incur a cost $q > c$ to coerce the individual to implement *her* desired action, $a_A(s)$.[9] Our second main result is that the impact of altruism on the authority depends crucially on the nature of disagreement. Under preference disagreement, a non-altruistic authority may prefer coercion; under sufficiently strong altruism, however, the authority strictly prefers truthful communication over all other equilibria. We say that the altruistic authority is *libertarian*, as she wants to inform the individual and then let him choose action. Under opinion disagreement, a non-altruistic authority may communicate truthfully; under sufficiently strong altruism, however, the authority always coerces. We say that the altruistic authority is *paternalistic*, as she constrains the individual's liberty in, as we shall see, his supposed self-interest.

The logic driving this result is as follows. After getting the signal s, the advisor prefers to send

[9] In most applications we discuss in Section 3.6, constraining the individual's liberty may be costlier than withholding information. To reflect this, we let $q > c$. Note that coercion differs from delegation; we discuss how these concepts are related in Appendix 3.11.

some message m, which induces $a_I(m)$, over imposing $a_A(s)$ iff

$$\mathbb{E}_A\left\{(u_A(a_A(s),\theta)+\varphi u_I(a_A(s),\theta))-(u_A(a_I(m),\theta)+\varphi u_I(a_I(m),\theta))\right\}<q-c\mathbf{I}_{\{m\neq s\}}. \tag{3.1}$$

A truthful equilibrium exists if it exists in the advisor game (above) and if truthful communication is preferred to coercion, i.e., if (3.1) is satisfied for $m=s$, for both signals.

Under preference disagreement, $a_A(s)$ approaches $a_I(s)$ as φ increases, so the benefit of coercion decreases; formally, when $m=s$, (3.1) reduces to

$$\varphi \geq \frac{b^2}{q}-1. \tag{TT$_{\text{pr, Authority}}$}$$

Combining this with ($\mathrm{TT}_{\text{pr, } c>0}$) yields that a truthful equilibrium exists when altruism is sufficiently strong. Intuitively, the authority transfers her information to the individual, who then makes an informed decision which is very close to what the authority would implement under coercion, but she need not incur the cost q. In the context of our example, where $\gamma=0.6$ and $c=0.06$, if we further let $q=0.10$ and characterize preference conflict by $b=1/3$, then (3.1) holds iff $\varphi\geq 0.11$ and ($\mathrm{TT}_{\text{pr, } c>0}$) holds iff $\varphi\geq 0.83$; hence, a truthful equilibrium exists (and, it can be shown, is preferred) iff $\varphi\geq 0.83$.

Under opinion disagreement, $a_A(s)$ does not approach $a_I(s)$ as φ increases. Instead, the authority's expected benefit from coercion – implementing $a_A(s)$ instead of $a_I(s)$ – increases with her regard for the individual. Formally, when $m=s$, (3.1) reduces to

$$\varphi \leq \frac{q}{(a_A(s)-a_I(s))^2}-1. \tag{TT$_{\text{op, Authority}}$}$$

Combining this with ($\mathrm{TT}_{\text{op, } c>0}$) yields that a truthful equilibrium exists when altruism is sufficiently weak. When altruism is strong enough, however, the authority always forces the individual to implement $a_A(s)$, as she believes that this protects the individual from his (in her view) erroneous action choice, and thus ultimately benefits him. As coercion after both signals is the only equilibrium that implements the authority's preferred action after both signals, this is the unique equilibrium, for sufficiently strong altruism. In the context of our example, for $\pi_I=0.85$, coercion after both signals is the unique equilibrium outcome for $\varphi\geq 0.6$.

In Appendix 3.11, we formally show that all of the results discussed above (also) obtain in a richer setting and with general (mixed and pure) strategies; further, we show that the main insights arise in the presence of both preference and opinion disagreement. In the remainder of the chapter, we study a more general model where the advisor or authority is faced with a population – a continuum of individuals with heterogenous preferences or beliefs. We show that the above insights remain applicable, and we develop additional results. All proofs are in Appendix 3.9. In Appendix 3.10, we also demonstrate that the main results, driven by intuitively analogous mechanisms, continue to apply a setting that is closely related in spirit, but where all results arise in a dominance solvable setting.

3.3 Model and Preliminaries

There is a continuum of individuals of unit mass, indexed by the unit interval $[0,1]$. Each individual (i, he) must take an action $a_i \in \mathbb{R}$ that renders him payoff $U_i(a_i, \theta) = -(a_i - \theta - b_i)^2$, where b_i is his preference bias, and $\theta \in \{0,1\}$ is an unknown state of the world. The preference biases are described by a general (continuous or discrete) distribution with density $f(b)$, so that $\bar{b} = \int_{-\infty}^{+\infty} bf(b)db$ and $\overline{b^2} = \int_{-\infty}^{+\infty} b^2 f(b)db$ are finite. Individual i's prior belief about the state of the world is given by $\Pr_i(\theta = 1) = \pi_i \in [0,1]$; the beliefs are characterized by a general distribution with density $g(\pi)$.[10]

3.3.1 The Altruistic Advisor

The advisor (A, she), who holds a prior belief $\Pr_A(\theta = 1) = \pi_A \in (0,1)$, privately observes one signal s about the state, with precision $\Pr(s = \theta|\theta) \equiv \gamma \in (0.5, 1)$, and sends a public message $m \in \{0,1\}$. Sending a false message entails the cost $c \geq 0$. After observing m, each individual chooses his action a_i.

Preference disagreement. The preference distribution $f(b)$ is not entirely concentrated at 0, i.e., $\int_{b\neq 0} f(b)db > 0$; the opinion distribution $g(\pi)$ is degenerate and satisfies $\pi_i = \pi_A$ for all $i \in [0,1]$. The advisor's material payoff is given by $u_A(a, \theta) = -\int_{-\infty}^{+\infty}(a_i - \theta)^2 f(b_i)db_i - c\cdot\mathbf{I}(m \neq s)$, where a denotes the set $\{a_i\}_{i\in[0,1]}$. Her material benefit is thus maximized when each individual's

[10] If F and G are discrete, the integrals below should be substituted by summations.

action a_i matches the state of the world, θ. We allow the advisor to be altruistic, i.e., to internalize a share φ of the individuals' payoffs ([Becker, 1974]).[11] Her utility is thus given by

$$U_A(a,\theta) = u_A(a,\theta) - \varphi \int_{-\infty}^{+\infty} (a_i - \theta - b_i)^2 f(b_i) db_i. \tag{3.2}$$

Opinion disagreement. The opinion distribution is not entirely concentrated at 0.5, that is, $\int_{\pi \neq 0.5} g(\pi) d\pi > 0$; the preference distribution $f(b)$ is degenerate and satisfies $b_i = 0$ for all $i \in [0,1]$. The advisor's prior belief is equal to $\pi_A = 0.5$, which is the median of the distribution $g(\pi)$, i.e., $\int_0^{0.5} g(\pi) d\pi \geq 0.5$ and $\int_{0.5}^1 g(\pi) d\pi \geq 0.5$. This assumption implies that the advisor is representative of the median opinion, which is motivated by our interpretation of the advisor as a government. As we shall see in the analysis, this assumption makes lying and coercion more unattractive than if the advisor can hold an extreme, unrepresentative opinion; thus, it stakes the game "against" intervention.[12] The advisor's material payoff is given by $u_A(a,\theta) = -c \cdot \mathbf{I}(m \neq s)$.[13] Her material payoff is thus independent of the individuals' action choices. This captures the fact that she does not care about the individuals' action choices *per se*. The utility of an altruistic advisor, who internalizes a share φ of the individuals' payoffs, is given by

$$U_A(a,\theta) = u_A - \varphi \int_0^1 (a_i - \theta)^2 g(\pi_i) d\pi_i.$$

The preference and opinion distributions $f(b)$ and $g(\pi)$, the authority's prior π_A, the signal precision γ, the lying cost c, and the strength of altruism φ are common knowledge.

Strategies and equilibrium A pure strategy of the advisor specifies, for each signal s, the message $m(s)$ that she sends, $m : \{0,1\} \to \{0,1\}$. The individuals' posterior beliefs conditional on message m are described by $\Pr_i(\theta = 1|m) = p_i(m)$, where superscript i signifies that individual i forms his beliefs using his prior π_i. A pure strategy of individual i specifies, for each message m,

[11] In Section 3.7, we discuss the alternative formulation where the advisor places a weight $(1-\varphi)$ on herself and φ on the individuals; we show that our main insights carry through to this setting as well.

[12] Note that, in the setting with preference disagreement, we do not assume that the advisor is representative of the median preference. As our results in the case of preference disagreement hold for any distribution $f(b)$, they (trivially) hold for the particular distributions that have a median equal to zero.

[13] Main results still hold if the material payoff of the advisor is defined similarly to the case of preference disagreement, i.e., $u_A(a,\theta) = -\int_0^1 (a_i - \theta)^2 g(\pi_i) d\pi_i - c \cdot \mathbf{I}(m \neq s)$.

the action $a_i(m)$ that he takes, $a_i : \{0, 1\} \to \mathbb{R}$. We solve for Perfect Bayesian Equilibria (PBE). Under opinion disagreement, each individual evaluates his expected utility, $\mathbb{E}[U_i(a_i, \theta)]$, using his own prior, π_i, and the advisor evaluates her expected utility, $\mathbb{E}[U_A(a, \theta)]$, using her prior, π_A. Importantly, the advisor thus uses her own prior when forming her expectation of the individuals' payoffs from a. This captures that the advisor may deem another action optimal for an individual than he does himself. If the advisor instead would evaluate an individual's expected payoff using *his* prior, she would derive utility from him choosing an action that he believes is optimal, although the advisor is convinced that he makes a mistake, and that he later will come to regret this choice. This distinction is essential, and it speaks to the notion of altruism that we apply: We say that a more altruistic advisor cares more about *her own valuation of* the individuals' payoffs. We discuss this further in Section 3.7.

3.3.2 The Altruistic Authority

After observing the signal, s, the authority (A, she) chooses between sending a public message m, after which the individuals choose actions, and engaging in *coercion*, whereby the authority mandates an action for all individuals, $a_A \in \mathbb{R}$. To capture that coercion may be costly, we let the advisor's cost of coercion be given by q, where $q \geq c$.[14] For simplicity, we assume $c = 0$.

Under preference disagreement, the authority's material (non-altruistic) payoff is given by $u_A(a, \theta) = -\int_{-\infty}^{+\infty}(a_i - \theta)^2 f(b_i) db_i - q \cdot \mathbf{I}(\text{coercion})$; under opinion disagreement, by $u_A = -q \cdot \mathbf{I}(\text{coercion})$. The individuals' material payoff functions, under preference and opinion disagreement, respectively, remain as specified in the game with an altruistic advisor.

Strategies and equilibrium. A pure strategy of the authority specifies, for each signal s, whether she chooses to coerce or not coerce, $C(s) \in \{\text{Coerce, Not coerce}\}$, what action $a_A(s) \in \mathbb{R}$ she mandates under coercion, and what message $m(s) \in \{0, 1\}$ she sends if she does not coerce. The individuals' beliefs and strategies remain as specified above. We solve for PBE.

[14] Depending on the application, this cost may reflect the instrumental cost of active intervention, or the authority's intrinsic aversion against removing the individual's liberty to choose. In most applications we discuss in Section 3.6, constraining the individual's liberty may be costlier than withholding information. To reflect this, we let $q > c$.

3.4 Libertarianism and Paternalism

3.4.1 Altruism and truthful advice

We search for fully revealing equilibria (FRE), where she transmits each signal truthfully to the population.

Proposition 8. *The impact of stronger altruism on the advisor's incentives to report truthfully depend on the nature of disagreement:*

- *Under preference disagreement, for any bias distribution $f(b)$ and lying cost c, there exists a threshold $\overline{\varphi}(c, \bar{b})$ s.t. a FRE exists if and only if $\varphi \geq \overline{\varphi}(c, \bar{b})$.*
- *Under opinion disagreement, there exists a non-degenerate set of opinion distributions $\mathcal{G}$ such that the advisor prefers to misreport at least one signal when $c = 0$. For any $g(\pi) \in \mathcal{G}$ and $c > 0$, there exists a threshold level of altruism $\overline{\varphi}(c, g)$ such that a FRE exists if and only if $\varphi \leq \overline{\varphi}(c, g)$. For any $g(\pi) \notin \mathcal{G}$ a FRE exists for any $c \geq 0$ and $\varphi \geq 0$.*

Regardless of the nature of disagreement, a higher cost of lying weakens the advisor's incentives to report truthfully: $\overline{\varphi}(c, \bar{b})$ (weakly) decreases in c and $\overline{\varphi}(c, g)$ (weakly) increases in c.

Under preference disagreement, truthful communication can thus arise iff altruism is *strong* enough; under opinion disagreement, whenever altruism impacts communication, truthful communication can arise iff altruism is *weak* enough. We discuss each setting in turn.

Preference Disagreement. From the setting with a single individual in Section 3.2 we know that the advisor's ideal action for individual i, given the signal s, is given by $a_{i,A}(s) = p(s) + \frac{\varphi}{1+\varphi} b_i = a_i(s) - \frac{1}{(1+\varphi)} b_i$. As her altruism strengthens, her disagreement lessens with each individual in the population. With quadratic utility functions, the precise strength of altruism necessary for a truthful equilibrium to exist when the advisor is faced with a population is equal to the strength of altruism necessary to sustain a FRE when the advisor is faced with a single individual with bias $\bar{b}$, the population's average preference bias. The proof of Proposition 8 establishes this formally; to see the logic of this result, w.l.o.g. suppose that $\bar{b} > 0$. Then, the advisor reveals the signal $s = 0$ to individual $\bar{b}$, and reveals the signal $s = 1$ to him iff the incentive compatibility condition, ($\text{TT}_{\text{pr},\, c\,>\,0}$), is satisfied for $b = \bar{b}$. When she is indifferent between revealing and misreporting the signal $s = 1$ to individual $\bar{b}$, her gain from misreporting it to all types $b_i > \bar{b}$ exactly offsets her loss from misreporting it to all types $b_i < \bar{b}$; hence, all that matters for her decision to reveal the signal truthfully is $\bar{b}$.[15] Our result thus follows immediately from the discussion in Section 3.2: For strong enough altruism, truthful reporting of the signal $s = 1$ is incentive compatible, and a FRE exists.

Opinion Disagreement. As the advisor does not care about the action choices for her own sake, she always reports truthfully when $\varphi = 0$; thus we henceforth consider $\varphi > 0$. First consider a simple opinion distribution with two (equally prevalent types of) individuals, t_1 and t_2, with priors $\pi_1 \leq \pi_A = 0.5 \leq \pi_2$, $\pi_1 < \pi_2$. After observing the signal s, the advisor would like all individuals to take the action $p_A(s)$. Given the individuals' strategies and beliefs, sending some message m induces actions $p_1(m)$ and $p_2(m)$. Suppose that the individuals' believe that the advisor reports truthfully. Then, if $s = 1$, sending a truthful message induces the actions $p_1(1) < p_A(1) < p_2(1)$. Clearly, the advisor always prefers to report the high signal, $s = 1$, truthfully to t_1, but she may prefer to lie to t_2, whom she deems too optimistic. The advisor can, however, send only one public message. When lying is costless, she simply trades off the disutility that a lie causes t_1 and the benefit that (she believes that) it brings t_2. Figure 3.1 illustrates that this induces her to misreport $s = 1$ in two regions of the (π_1, π_2) space, denoted by $L1_{\text{left}}$ and $L1_{\text{right}}$, when the signal precision is $\gamma = 0.6$. An analogous argument yields that she misreports $s = 0$ in the regions $L0_{\text{up}}$ and $L0_{\text{low}}$.[16]

Consider the region $L1_{\text{left}}$. When $\pi_1 = 0$, the action choice of t_1 is unaffected by the advisor's message. This brings us back to the setting with a single individual, where the advisor reveals $s = 1$ when her ideal action $a_A(1)$ is closer to $a_2(1)$ than to $a_2(0)$, i.e., when ($\text{TT}_{c\,=\,0}$) is satisfied. This occurs when opinion disagreement is minor; in particular, as noted in our example in Section 3.2, when $\gamma = 0.6$, a truthful equilibrium exists for $\pi_2 \leq 0.604$.

Now consider some $\pi_2 \in (0.604, 1]$. The advisor's benefit from lying derives from the fact that (she believes that) the lie improves the action choice of t_2. When $\pi_1 > 0$, the advisor also suffers a loss from lying, which derives from the fact that lying worsens the action choice of t_1. The proof of

[15] In a setting with two individuals and general loss functions, [Goltsman and Pavlov, 2011] establish that truthful public communication can be sustained iff it can be sustained with an individual with average bias. This suggests that our result can be generalized to other loss functions.

[16] When the priors π_1 and π_2 are equidistant from the advisor's prior, $0.5 - \pi_1 = \pi_2 - 0.5$, the loss from misreporting the signal $s = 1$ always exceeds the benefit. In general, for any distribution $g(\pi)$ that is symmetric around 0.5, a FRE exists for any $\varphi \geq 0$ and $c \geq 0$.

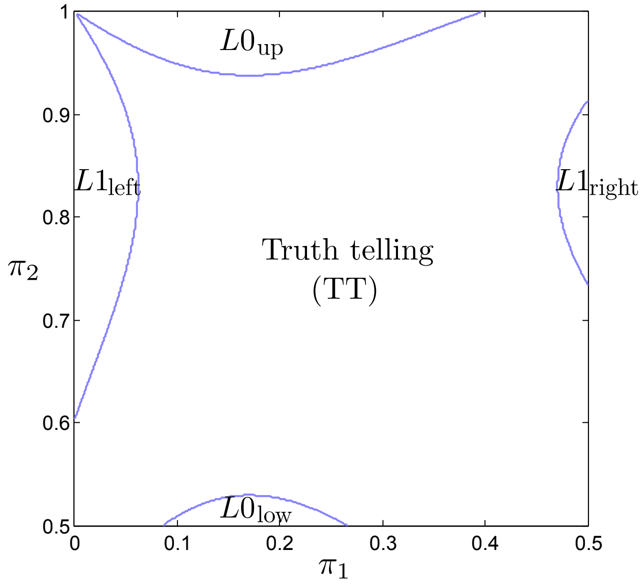


Figure 3.1: Incentives to misreport the signal, $\gamma = 0.6$.

Proposition 8 establishes that, as π_1 increases from zero to 0.5, the advisor's loss from lying to t_1 first increases, and then decreases. The loss is outweighed by the (fixed) benefit from lying when π_1 is close to zero (region $L1_{\mathrm{left}}$) and, potentially, when π_1 is close to 0.5 (region $L1_{\mathrm{right}}$).

The non-monotonicity of the loss from lying to t_1 is due to two opposing effects. As π_1 increases, t_1 gets more responsive to the public message; that is, the distortion induced by the lie, $a_1(1)-a_1(0)$, increases. This *distortion effect* raises the advisor's loss from lying. As π_1 approaches 0.5, however, the disagreement between t_1 and the advisor also lessens; formally, $a_1(s)$ approaches $a_A(s)$ for both signals $s \in \{0,1\}$. As the advisor's loss function is flatter close to her own ideal action, her loss from any given distortion in action choice thus decreases with π_1. This *disagreement effect*, which reduces her loss from lying, outweighs the distortion effect for π_1 close to 0.5.

Finally, the region $L1_{\mathrm{left}}$ is non-empty for all $\gamma \in (0.5, 1]$; intuitively, there always exists some region of extreme π_2 for which t_2 would benefit from a lie, and when $\pi_1 = 0$, lying entails no loss. The region $L1_{\mathrm{right}}$, however, disappears as γ increases. Intuitively, when $\pi_1 = 0.5$, lying harms t_1; when the signal is precise enough, this loss outweighs the benefit from lying to t_2.

Now consider a general opinion distribution $g(\pi)$ with median $1/2 = \pi_A$. Clearly, when $c = 0$ the advisor misreports the signal $s = 1$ iff enough mass of the distribution is contained in $L1 \in$

$\{L1_{\text{left}}, L1_{\text{right}}\}$, and the signal $s = 0$ iff enough mass is contained in $L0 \in \{L0_{\text{up}}, L0_{\text{down}}\}$.[17] As the regions $L1_{\text{left}}$ and $L0_{\text{up}}$ are non-empty for all γ, the set of such distributions $\mathcal{G}$ is uncountable.

For distributions $g \in \mathcal{G}$, a FRE fails to exists when $c = 0$, but may exist when $c > 0$. For any given cost of lying, however, increasing the strength of altruism eventually destroys truthful communication, *even though the advisor represents the median opinion in the population.* Intuitively, while the advisor knows that withholding information harms some individuals – whose informed choices would dominate their misinformed ones – she is also convinced that it benefits others, who would misinterpret a truthful report. For all distributions $g \in \mathcal{G}$ she believes that, on average, the individuals' are better off with their misinformed action choices. This conviction makes her more inclined to lie, the stronger is her altruism: in essence, the more she cares about the individuals, the more willing she is to bear the cost of lying, as lying protects the individuals from their own informed (erroneous) choices. As the population anticipates that the altruistic advisor lies, no informative communication can take place. Paradoxically, the population may thus prefer a disinterested advisor, whom can be trusted to tell the truth.

3.4.2 Altruism and Coercion

We now consider an authority who, after observing the signal s, chooses between sending a public signal or mandating a single action for the population.

Proposition 9. *The impact of stronger altruism on the authority's incentives to report truthfully or coerce depend on the nature of disagreement:*

- *Under preference disagreement, for any preference distribution $f(b)$ with $\bar{b}^2 \neq \overline{b^2}$ and any $q < \bar{q}$, there exist finite thresholds $\varphi_{CC}(q, \bar{b}, \overline{b^2}) < \varphi_C(q, \bar{b}, \overline{b^2}) < \varphi_{TT}(q, \bar{b}, \overline{b^2})$. For $\varphi < \varphi_{CC}(q, \bar{b}, \overline{b^2})$ there exists a unique equilibrium, in which the authority coerces with probability one after each signal s. For $\varphi < \varphi_C(q, \bar{b}, \overline{b^2})$ every equilibrium involves the authority coercing with strictly positive probability. A FRE exists if and only if $\varphi \geq \varphi_{TT}(q, \bar{b}, \overline{b^2})$. In pure strategies, the FRE is strictly preferred, by both players, to any other equilibrium.*

- *Under opinion disagreement, for any opinion distribution $g(\pi)$ and $q \geq 0$, there exist thresholds $\varphi_C(q, g) \leq \varphi_{CC}(q, g)$. For $\varphi > \varphi_{CC}(q, g)$, there exists a unique equilibrium, in which the*

[17] This is discussed in more detail in the proof of Proposition 8.

> *authority coerces with probability one after each signal s. For $\varphi > \varphi_C(q, g)$, every equilibrium involves the authority coercing with strictly positive probability. For any $g \notin \mathcal{G}$, there exists a threshold $\varphi_{TT}(q, g)$, s.t. a FRE exists if and only if $\varphi \leq \varphi_{TT}(q, g)$.*

Regardless of the nature of disagreement, a higher cost of coercion weakens the authority's incentives to coerce: $\varphi_{CC}(q, \bar{b}, \overline{b^2})$, $\varphi_C(q, \bar{b}, \overline{b^2})$, and $\varphi_{TT}(q, \bar{b}, \overline{b^2})$ are decreasing in q; $\varphi_{TT}(q, g)$, $\varphi_C(q, g)$, and $\varphi_{CC}(q, g)$ are increasing in q.

Under preference disagreement, for sufficiently strong altruism, truthful communication is possible; moreover, in pure strategies, it is strictly preferred by the authority.[18] Under opinion disagreement, a non-altruistic authority communicates truthfully; under sufficiently strong altruism, however, she always coerces. We discuss each setting in turn.

Preference Disagreement. The more altruistic the authority, the more she values if each individual gets to implement the action that *he* prefers; that is, $a_{i,A}(s)$ approaches $a_i(s)$ as φ increases. This makes the benefit of coercion – the ability to mandate another action than the individuals' own choices, $a_i(s)$ – decreasing in the level of altruism, φ. If the authority mandates an action, she chooses $a_A(s) = p(s) + \frac{\varphi}{1+\varphi}\bar{b}$, as an authority faced with a single individual of type $\bar{b}$. She would, however, prefer to impose different actions on different individuals, $a_{i,A}(s) = p(s) + \frac{\varphi}{1+\varphi}b_i$. When she enacts a uniform mandate, she cannot take into account the nuances in the population's preferences even though she would want to. This "indirect cost" of coercion does not arise in the single individual setting. The strength of altruism necessary for truthful reporting to dominate coercion is therefore weakly smaller in the population setting than in the setting with an individual of type $\bar{b}$.

Combining the fact that truth-telling dominates coercion for strong enough altruism with Proposition 8 yields that a FRE exists iff the authority's altruism is sufficiently strong. In the FRE, the individuals choose actions that she eventually *prefers* to the action that she would mandate; further, she need not incur the cost q. A sufficiently benevolent authority thus prefers to behave in a *libertarian* fashion – to transfer the information at her disposal, thereby giving each individual the means to make as informed a choice as possible, and then give them the liberty to choose the

[18] When $\bar{b}^2 = \overline{b^2}$, the preference distribution is degenerate; this case is thus equivalent to the single-individual setting that we analyze in Appendix 3.11 (and in the example in Section 3.2).

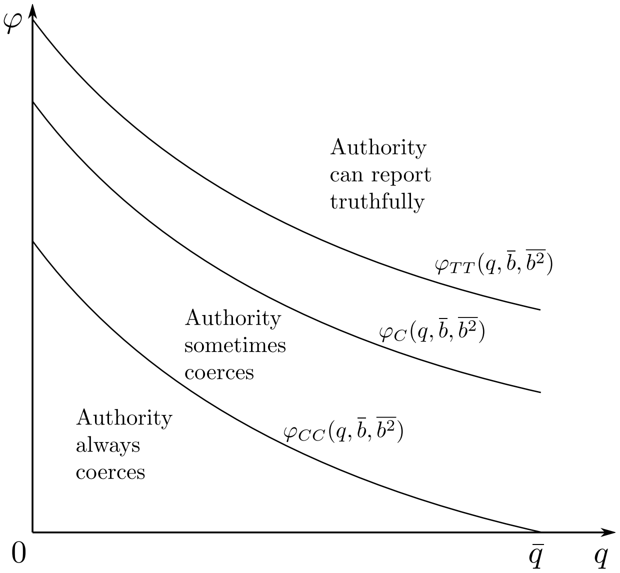

φ
Authority can report truthfully
$\varphi_{TT}(q,\bar{b},\overline{b^2})$
Authority sometimes coerces
$\varphi_C(q,\bar{b},\overline{b^2})$
Authority always coerces
$\varphi_{CC}(q,\bar{b},\overline{b^2})$
0
$\bar{q}$
q

Figure 3.2: Equilibrium outcomes under preference disagreement $f(b)$, where $\bar{b}^2 \neq \overline{b^2}$.

actions that they want.

Coercion becomes more viable, however, the weaker is altruism. More precisely, for $\varphi < \varphi_C(q,\bar{b},\overline{b^2})$, the authority coerces with strictly positive probability for at least one signal s, and for $\varphi < \varphi_{CC}(q,\bar{b},\overline{b^2}) < \varphi_C(q,\bar{b},\overline{b^2})$ coercion after both signals is the unique equilibrium outcome. Figure 3.2 gives an approximate representation of these sets of (q,φ), and illustrates that coercion is less viable, the higher is q.[19]

Opinion Disagreement. As the authority does not care about the action choices for her own sake, communication strictly dominates coercion when $\varphi = 0$, for any $q > 0$. When $\varphi > 0$, however, the advisor cares about (the actions of) the population. She is convinced that, given s, the action $a_A(s)$ is ideal for all individuals. Thus, she believes that mandating it (strictly) benefits all individuals who, in the absence of a mandate, would take an action $a_i \neq a_A(s)$. As the benefit accrues to the population, her valuation of this benefit increases with her altruism, φ, which makes coercion increasingly viable. Indeed, for $\varphi > \varphi_C(q, f(\pi))$, every equilibrium involves coercion with positive probability after at least one signal s. When altruism strengthens further,

[19] Note, that if $\varphi_C(q,\bar{b},\overline{b^2}) \leq \varphi < \varphi_{TT}(q,\bar{b},\overline{b^2})$, then the authority cannot behave as a truth-telling advisor. Instead, there are (possibly, mixed strategy) equilibria, in which the authority communicates some information and/or coerces.

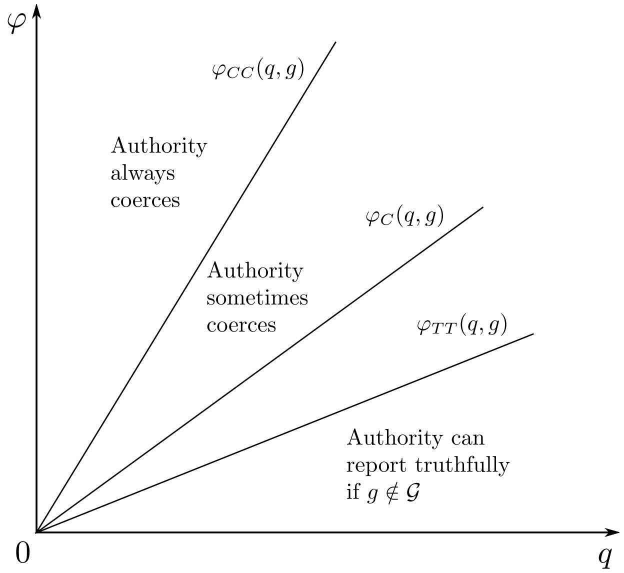


Figure 3.3: Equilibrium outcomes under opinion disagreement $g(\pi)$.

for $\varphi > \varphi_{CC}(q, f(\pi))$, coercion after both signals is the unique equilibrium; this is true even if a FRE is sustainable in the advisor game. Intuitively, it is the only equilibrium that implements $a_A(s)$ after both signals, and for sufficiently strong altruism, she is willing to bear the cost q to mandate these actions, thereby protecting the individuals who otherwise would make (in her view) erroneous action choices. We say that the altruistic authority is *paternalistic*, as she constrains the individuals' liberty *out of affection*; i.e., in their supposed best interest.

For weak enough altruism – and in its absence – however, the authority cares too little to intervene by enacting a costly mandate. Instead, she communicates, and thus a FRE exists for opinion distributions $g(\pi) \notin \mathcal{G}$. Figure 3.3 gives an approximate representation of these sets of (q, φ), and illustrates that coercion is less viable, the higher is q.[20]

3.5 Targeted Advise and Targeted Mandates

Above we considered an advisor who can send a single public message to the population, and an authority who (also) has the option to mandate a single action. We now allow the advisor to engage

[20] Note, that for $\varphi_{TT}(q,g) < \varphi \leq \varphi_C(q,g)$ and any distribution $g(\pi)$, the authority cannot behave as a truth-telling advisor. Instead, there are (possibly, mixed strategy) equilibria, in which the authority communicates some information and/or coerces.

in targeted communication, whereby she sends different messages to different individuals, and the authority to also enact targeted mandates.

3.5.1 Altruism and targeted vs. public advice

Under targeted communication, the advisor privately observes the signal s about the state, and sends one message $m_i \in \{0,1\}$ to each individual i. Then, each individual chooses his action a_i. To capture that lying may be costly, we define the measure of individuals to whom the advisor sends a false message, η, and let the advisor's cost of lying be given by ηc, where $c \geq 0$. With the exception of the redefinition of the cost of lying, all utility functions remain as specified above. A pure strategy of the advisor now specifies, for each signal s, the message $m_i(s)$ that she sends to each individual i, $m_i : \{0,1\} \to \{0,1\}$; the individuals' posterior beliefs and pure strategies are redefined accordingly. We solve for pure strategies PBE.

We say that the advisor is *more credible* when truthful communication can be sustained with a larger share of the population. For either type of disagreement, in pure strategies, more credible communication is preferred by the individuals and the advisor. Whether the advisor is more credible in this setting with private messages than when she sends a single public message, depends on the strength of the advisor's altruism and the nature of disagreement.

Proposition 10. *The impact of stronger altruism on the relative credibility of public and private messages depend on the nature of disagreement:*

- *Under preference disagreement, for any preference distribution $f(b)$ and cost of lying c, the advisor is more credible with private messages when $\varphi \leq \overline{\varphi}(c, \bar{b})$, but more credible with a public message otherwise.*

- *Under opinion disagreement, for any opinion distribution $g(\pi) \in \mathcal{G}$ and cost of lying c, the advisor is more credible with a public message when $\varphi \leq \overline{\varphi}(c, g)$, but more credible with private messages otherwise. For any opinion distribution $g(\pi) \notin \mathcal{G}$, public communication is (more) credible for all φ.*

Preference Disagreement. The setting with a single individual yields that the advisor's message to individual i is credible iff ($\text{TT}_{\text{pr},\, c > 0}$) is satisfied (for $b > 0$). For any given level of altruism,

φ, the advisor's incentive compatibility conditions for truthful reporting of $s \in \{0,1\}$ define the set of individuals to whom the advisor can send a credible message, $b_i \in (\underline{b}(\varphi), \bar{b}(\varphi))$. Under targeted communication, she can thus be credible to the subset of the population with moderate preferences, but is non-credible to individuals with extreme biases. As altruism strengthens, the advisor is willing to communicate truthfully to individuals with more extreme preferences ($\underline{b}(\varphi)$ decreases in φ and $\bar{b}(\varphi)$ increases in φ); that is, she gains credibility.

When the advisor sends a public message, Proposition 8 yields that she can be credible whenever she would send a truthful message to an individual with bias $\bar{b}$, i.e., when $\varphi \geq \overline{\varphi}(c, \bar{b})$. Consider $\overline{\varphi}(c, \bar{b}) > 0$.[21] For $\varphi < \overline{\varphi}(c, \bar{b})$, the advisor is more credible under private messages, as she communicates truthfully with some individuals; namely, those with $b_i \in (\underline{b}(\varphi), \bar{b}(\varphi))$. As φ increases in this region, the set of individuals to whom the advisor can communicate truthfully under private messages increases – as well as the credibility advantage of private over public messages. At $\varphi = \overline{\varphi}(c, \bar{b})$, however, the advisor becomes credible under public messages; hence, she can communicate truthfully to *all* individuals when communication is public. Under targeted communication, she remains non-credible to individuals with more extreme biases than $\bar{b}$. An increase in altruism that makes the advisor willing to send a truthful message in public thus strengthens her credibility among individuals with whom preference divergence is too large for her to be credible in private.

These results relate to those of [Farrell and Gibbons, 1989], and [Goltsman and Pavlov, 2011], who compare the credibility of private and public messages of a non-altruistic advisor in settings with two individuals. When we shut down altruism in our model ($\varphi = 0$), we replicate their result that the relative credibility of private versus public messages depends on the preference distribution, $f(b)$.[22] In this sense, we extend their results to populations with more than two individuals. More substantively, while the relative credibility of different modes of communication varies with $f(b)$ when $\varphi = 0$, we show that introducing altruism eliminates this indeterminacy: for all $f(b)$, when the advisor is sufficiently altruistic, public communication is more credible.[23] Lastly, we note

[21] If instead $\overline{\varphi}(c, \bar{b}) = 0$, public communication is (more) credible than targeted communication for all φ.

[22] More precisely, all cases that are discussed in these papers – (i) credible communication with both public and private messages, (ii) no credible communication (in either case), (iii) subversion of (credibility in) private communication under public communication, and (iv) (one-sided or mutual) discipline under public communication – can arise, depending on the distribution $f(b)$.

[23] In the language of [Farrell and Gibbons, 1989] and [Goltsman and Pavlov, 2011], when $\varphi < \overline{\varphi}(c, \bar{b})$, the advisor's private communication with individuals with $b_i \in (\underline{b}(\varphi), \bar{b}(\varphi))$ is *subverted* under public communication; when $\varphi \geq$

that considering combined communication – where the advisor can send both public and private messages – does not improve advisor's credibility (as in [Farrell and Gibbons, 1989]). That is, for any pure strategies PBE of the combined communication setting, there exists a PBE with either public or targeted communication that allows the advisor to be (weakly) more credible.

Opinion Disagreement. First, consider the simple opinion distribution with two (equally prevalent types of) individuals, t_1 and t_2, with priors $\pi_1 < \pi_A = 1/2 < \pi_2$, and let lying be costless, $c = 0$. When the advisor can target her messages, a message that (she believes) benefits one type does not exert any negative externality on the other type. She thus simply treats each type as she would if faced with a single individual with prior π_i, for $i \in \{0, 1\}$: she misreports the signal $s = 1$ to t_2 if he is too optimistic, and misreports the signal $s = 0$ to t_1 if he is too pessimistic (provided that they believe the messages). From our example in Section 3.2, we recall that, when the signal precision is $\gamma = 0.6$, the advisor misreports the signal $s = 1$ to t_2 iff $\pi_2 > 0.604$; similarly, she misreports $s = 0$ to t_1 iff $\pi_1 < 0.396$. This gives rise to four regions in the (π_1, π_2) space, illustrated in Figure 3.4, where targeted advice is credible to both types ($T_{\pi_1}T_{\pi_2}$), only to t_2 ($L0_{\pi_1}T_{\pi_2}$), only to t_1 ($T_{\pi_1}L1_{\pi_2}$), and to neither type ($L0_{\pi_1}L1_{\pi_2}$).[24]

Consider the point $A \in L0_{\pi_1}L1_{\pi_2}/\{L1_{\text{left}}, L0_{\text{up}}\}$. While the advisor cannot send credible targeted messages to any type, she can issue credible public advice. Similarly, the advisor's credibility is higher with public advice in $L0_{\pi_1}T_{\pi_2}/L0_{\text{low}}$ and $T_{\pi_1}L1_{\pi_2}/L1_{\text{right}}$, as she can provide credible private advice to only half of the population. The converse can also arise, however: Consider $B \in L1_{\text{right}}$. Here, the advisor cannot provide credible public advice, but she can provide credible private advice to half of the population; thus, she is more credible with private messages.

When lying is costly, $c > 0$, altruism may influence the advisor's behavior, and thereby the relative credibility of public and private messages. Again consider $B \in L1_{\text{right}}$.[25] Regardless of the strength of altruism, the advisor provides credible private advice to t_1, as she believes that the action she induces t_1 to take when sending a truthful message is better than the action she induces when misreporting. With a single public signal, Proposition 8 yields that she is credible (to the entire

$\overline{\varphi}(c, \bar{b})$, for types $b_i \notin (\underline{b}(\varphi), \bar{b}(\varphi))$ the advisor is *disciplined* by the presence of others.

[24] Here $L1_{\pi_i}$ and $L0_{\pi_i}$ denote the incentives to misreport the signals 1 and 0 to individual i.

[25] Figure 3.4 plots all regions for $c = 0$. When $c > 0$, the size of the regions where the advisor misreports some signal(s) shrink(s); however, the regions' relative positions largely remain as in Figure 3.4.

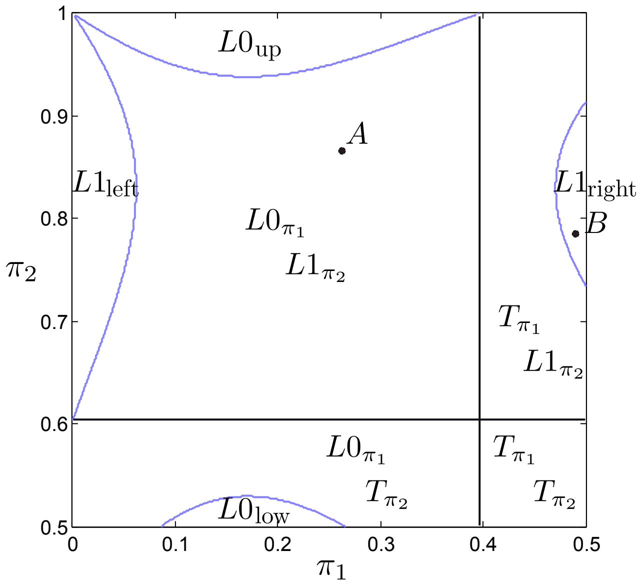


Figure 3.4: Incentives to misreport the signal under opinion disagreement, $\gamma = 0.6$.

population) iff altruism is sufficiently weak. A weakly altruistic advisor is thus more credible when issuing public advice, whereas a strongly altruistic advisor is more credible with private messages. An analogous argument applies to the region $L0_{\mathrm{low}}$. In a similar vein, in the regions $L1_{\mathrm{left}}$ and $L0_{\mathrm{up}}$, the advisor is more credible under public communication when altruism is weak, and equally (non-)credible under both types of communication when altruism is strong. When we consider general opinion distributions, similar (weak or strong) credibility reversals arise for all distributions such that altruism affects public communication, $g(\pi) \in \mathcal{G}$. For opinion distributions such that altruism does not impact the credibility of public messages, $g(\pi) \notin \mathcal{G}$, public communication is credible regardless of φ.

When we shut down altruism in our model ($\varphi = 0$), we obtain, in our setting with opinion conflict, an insight akin to that obtained by [Farrell and Gibbons, 1989] and [Goltsman and Pavlov, 2011] in settings with preference disagreement: the relative credibility of private versus public signals depends on the *opinion* distribution, $g(\pi)$.[26] Furthermore, we show that in the presence of altruism, for any opinion distribution $g(\pi) \in \mathcal{G}$, a sufficiently altruistic advisor is more credible when

[26] In this sense, we extend their results to the case of opinion conflict (and allow for populations with more than two individuals).

communication is private.[27] Finally, as in case of preference disagreement, combined communication does not improve the advisor's credibility in pure strategies PBE.

To conclude, under both preference and opinion disagreement, an increase in φ may thus yield a *credibility reversal*, where the relative credibility of public and private messages reverses. The direction of this reversal, however, depends on the nature of disagreement. Under preference disagreement, the relative credibility of a public message always increases with altruism; under opinion disagreement, the reverse is true for all opinion distributions $g(\pi) \in \mathcal{G}$.[28]

3.5.2 Altruism and Targeted Mandates

Setting. After observing the signal, s, the authority (A, she) chooses between either sending a message $m_i(s)$ to individual i, after which the individual chooses action himself, or engaging in *coercion*, whereby the authority makes the action choice on behalf of individual i, $a_{i,A}(s) \in \mathbb{R}$. To capture that coercion may be costly, we define the measure of individuals whom the advisor coerces, ϵ, and let the advisor's cost of coercion be given by ϵq, where $q \geq c = 0$; else, all utility functions remain as specified above. A pure strategy of the advisor now specifies, for each signal s and for each individual i, whether she chooses to coerce or not coerce i, $C_i(s) \in \{\text{Coerce}, \text{ Not coerce}\}$, what action $a_{i,A}(s) \in \mathbb{R}$ she mandates under coercion, and what message $m_i(s) \in \{0, 1\}$ she sends if she does not coerce. The individuals' posterior beliefs and pure strategies are redefined accordingly. Again, we solve for PBE.

Proposition 11. *The authority's use of targeted mandates depends on the nature of disagreement:*

- *Under preference disagreement, any mandates are individual-specific. The share of the population subjected to (tailored) mandates is decreasing with the advisor's altruism.*

- *Under opinion disagreement, a single mandate is applied to all individuals whose action*

[27] In the language of [Farrell and Gibbons, 1989] and [Goltsman and Pavlov, 2011], when $\varphi \leq \overline{\varphi}(c, g)$, the advisor is *disciplined* by the presence of others; when $\varphi > \overline{\varphi}(c, g)$, either credible reporting to one individual is subverted, or truthful communication arises with neither of them.

[28] More precisely, all cases that are discussed in these papers can arise, depending on the distribution $g(\pi)$. In the two-individual case depicted in Figure 3.4, (i) credible communication with both public and private signals occur in $T_{\pi_1}T_{\pi_2}$; (ii) no credible communication (in either case) in $L1_{\text{left}}$ and $L0_{\text{up}}$; (iii) subversion of private communication in $L1_{\text{right}}$ and $L0_{\text{low}}$; (iv) one-sided discipline in $L0_{\pi_1}T_{\pi_2}/L0_{\text{low}}$ and $T_{\pi_1}L1_{\pi_2}/L1_{\text{right}}$; and (v) mutual discipline in $LL/\{L1_{\text{left}}, L0_{\text{up}}\}$.

choices are restricted. The share of the population subjected to the (single) mandate is increasing with the advisor's altruism.

When the authority can target her advice and mandates, she treats each individual i as she would if she were faced with a single individual with preference b_i or prior π_i. Our analysis in Section 3.2 (and Appendix 3.11) of the single individual setting thus applies.

Preference Disagreement. The authority's ideal action for individual i, given the signal s, is given by $a_{i,A}(s) = a_i(s) - \frac{b_i}{1+\varphi}$. If she mandates an action for individual i, she chooses $a_{i,A}(s)$; that is, mandates are individual-specific. If coercion is costless, $q = 0$, she enacts mandates for all individuals whose preferences differ from her own, $b_i \neq 0$. When coercion is costly, $q > 0$, the authority weighs the cost of each mandate against her expected benefit. For a given strength of altruism, her expected benefit from imposing the action $a_{i,A}(s)$ on an individual is higher, the larger is their preference divergence $|b_i|$. For a given cost of coercion $q > 0$, the share of the population that is subjected to mandates decreases with the advisor's altruism. Intuitively, as φ increases, $a_{i,A}(s)$ approaches $a_i(s)$, which reduces her benefit from mandating $a_{i,A}(s)$, for each i; eventually, she behaves in a libertarian fashion towards all individuals.

Opinion Disagreement. After observing the signal s, the advisor deems the action $a_A(s) = p_A(s)$ optimal for all individuals. If coercion is costless, $q = 0$, she mandates this action for all individuals whose opinions differ from her own, $\pi_i \neq \pi_A$. When coercion is costly, $q > 0$, the authority weighs the cost of each mandate against her expected benefit. For a given strength of altruism, her expected benefit from imposing the action $a_A(s)$ on an individual is higher, the larger is their opinion divergence $|\pi_A - \pi_i|$. Whenever she mandates $a_A(s)$ for only a subset of the population, her mandate therefore applies to the individuals whose opinions are the most extreme. Intuitively, she prioritizes to constrain the individuals who, if left to choose their own actions, would make the largest mistakes (in her view). For a given cost of coercion $q > 0$, the share of the population that is subjected to the mandate increases with the advisor's altruism; when she cares more about any given individual, her willingness to intervene – and help improve his action choice – increases. From strong enough altruism, she mandates the action $a_A(s)$ for all individuals with $\pi_i \neq \pi_A$.

3.6 Applications

On any given issue, some (politicians) advocate a laissez-faire approach – to provide public information, but then let each individual make his own action choice – whereas others advocate constraining individual liberty, through mandates or prohibitions. Why do those who advocate restrictions in individual liberty deem such restrictions better than information provision? What determines whether a politician advocates one or the other, and what distinguishes arguments in favor of regulation from those in favor of information provision?

One commonly held view is that politicians' differ in their valuations of personal liberty. Those with higher valuations of liberty, then, are less prone to intervene in constituents' lives. This view has an immediate implication: if R places a higher value on individual liberty than does L, then all mandates supported by R should be (a strict subset of the mandates) supported by L. In reality, however, politicians' desired interventions into individual liberty do not always satisfy this property. The following stylized example illustrates a violation of the "cost-of-liberty" hypothesis: L argues in favor of a citizens right to enter marriage regardless of sexual orientation, whereas R wants to reserve this right for heterosexual couples; and, at the same time, L favors of a mandate to purchase health insurance, whereas R argues that a citizen has the right to choose.

Our results offer an alternative, precise lens through which we can understand politicians' views on the government's right to regulate activities that can cause self-harm: they are intimately linked to whether the conflicted issue is, consciously or subconsciously, framed in terms of preferences or opinions. We obtain the following, empirically testable predictions:

Regulation vs. Liberty. *Any given benevolent politician wants to enact regulation on issues that she believes is a matter of opinion, but leave individuals to make their own choices on issues that she believes is a matter of preferences.* To understand what issues a benevolent authority regulates, it thus suffices to ask what issues she views in terms of differing opinions. This prediction contrasts with the commonly held view that politicians' differing views on regulation emanates from differences in their valuations of liberty. Further, this prediction is consistent with the fact that some politicians are in favor of regulation that others oppose, and vice versa.

Justifications for Enacted Regulation. *Advocates of regulation view differences in observed individual actions as a result of differing opinions, and thus use arguments based on beliefs; opponents of regulation view differences in observed actions as a result of differing preferences, and thus use arguments based on liberty.* These debates are, literally, clashes of ideas of what drives individual choices. When regulation does emerge, it is justified on the basis of differing opinions.

3.6.1 Benevolent paternalism

We discuss these predictions in light of concrete, real-world examples of regulation.

Protection from physical self-harm: Euthanasia and assisted suicide. Our framework suggests that a benevolent authority denies an individual's request for euthanasia if she believes that he holds an incorrect belief about his own wish to die. Such differences in opinion can arise, e.g., if the authority believes (i) that the individual experiences pain that he, if denied euthanasia, will learn to endure over time, and that (ii) he, then, will be happy that his request was denied. In contrast, the authority allows euthanasia if she deems the individual to be fully conscious of the consequences of his decision, and to simply prefer to die.

In the few places where assisted suicide is legal – Belgium, Luxembourg, the Netherlands, Switzerland and the American states of Oregon, Washington, and Montana – the legal provisions are precise, and they shed light on precisely the distinction between preferences and opinions. That is, they indicate a desire to disallow requests for euthanasia that are made by patients who hold incorrect beliefs, but to accommodate requests that reflect (true) preferences. In Switzerland, for example, a person who assists a suicide can avoid conviction by proving that the deceased knew what he or she was doing, was capable of making the decision, and had requested death several times [Whiting, 2002].[29]

Valid Consent and Restrictions on Minors. At the heart of the distinction between preferences and opinions is the question of what constitutes a valid consent: paternalism arises when the benevolent authority disqualifies an individual's consent to an action or activity, believing that

[29] Similarly, in Oregon, the patient must have made one written and two oral requests in order for the assisting physician to escape criminal liability; moreover, the physician must make a written confirmation that the act is voluntary and informed (Oregon Death with Dignity Act).

he would change his mind if he were of a sound opinion.[30] This highlights the tenuous nature of paternalism: How can a government know better than a single individual what he or she actually wants? When can consent be disqualified? Our analysis shows that, when mandates can be targeted, a benevolent authority enacts restrictions that apply to those whose opinions are the farthest from π_A; that is, the individuals who are the least qualified to make decisions in their own interests and thus, in the absence of regulation, would make the largest mistakes. A prominent example of disqualification of consent only for some individuals – that is, targeted mandates – is restrictions on minors. Many governments require minors to have blood transfusions even when their religious beliefs forbid it; no such restriction is imposed on adults. Similarly, there is a legal drinking age, driving age, voting age, an age of criminal responsibility, etc. Our framework suggests that these distinctions are justified by a view that an adult knows what he or she is doing; the government may, however, protect a minor from his own misjudgment. This is illustrated by the below statement, made by Dr. Steven Mings, a past president of the Idaho Dermatology Society, when discussing a bill to restrict tanning salon use among minors in order to protect them from skin cancer ([Yardley, 2012]).

> *"If you want to be 19 and not wear a motorcycle helmet, if you want to be 19 and smoke a cigarette, if you want to be 19 and go get a sunburn, that's one issue," he said. "But I think it's hard to argue against protecting someone under 18 from making bad decisions."*

This also relates to how opinions are formed. Over time, individuals may update their beliefs about, e.g., the dangers of riding a motorcycle without a helmet. If learning brings the individual's opinion closer to that of the authority – which may occur when there exists an "objective truth," e.g., the actual risk of death from riding a motorcycle without a helmet, which corresponds to π_A – then individual's own decisions improve over time. Consequently, she would want to target only those who, at a given point in time, hold (the most) erroneous beliefs; this is accomplished through minority regulation, as the law no longer applies when the individual becomes an adult.

[30] Indeed, dueling was outlawed because lawmakers believed that even those who consented to a duel were giving invalid consents procured through extreme pressure. Similarly, it is contemporaneously debated whether prostitutes – even if they earn a decent living and are protected against disease – are giving valid consents ([Suber, 1999]). We note that such disqualifications are inherently inconsistent with the revealed preference axiom.

Protection from moral self-harm. If opinions can evolve over time, restrictions on the liberty of adults may arise primarily on matters where the authority deems it unlikely that individuals' beliefs approach her own (in her view, correct) belief over time. This may occur when learning is unlikely to bring individuals' beliefs (much) closer to some "objective truth," perhaps because no such truth exists. For example, an authority with a strong religious faith may deem it unlikely that individuals who are atheists will update their (in her view erroneous) beliefs over time.

Indeed, many restrictions on the liberty of adults are motivated on moral grounds. If a benevolent authority is convinced that some behavior is sinful or morally corrupting – so that the individual would be better off resisting, and otherwise would come to regret his behavior in the future – criminalization can be a benevolent act to improve the citizens' well-being (in this life, or in the afterlife). In the absence of such (religious) beliefs on the part of the authority, she would allow the behavior, thereby giving those who take pleasure from it the liberty to choose. Behaviors that are or have been banned on the grounds of protecting individuals against morally corrupting behavior include consumption of (adult) pornography (Time 1969), and some sexual acts between consenting adults, e.g., of the same gender, or relatives. Externalities arising from these activities are arguably small; but more to the point, regulation often explicitly rely on moral justifications. The fact that laws against homosexual acts were justified on moral grounds, for example, was made explicit in the U.S. Supreme Court ruling *Lawrence v. Texas*, in 2003, which deemed laws that criminalize homosexual acts *on the grounds of morality* unconstitutional, and therefore repealed them.

3.6.2 Coercion and Benevolence

Our results by no means imply that coercion always is an indication of benevolence. On the contrary, our framework predicts that coercion arises in two cases: (i) under conflicting opinions when altruism is sufficiently strong; and (ii) under conflicting preferences when altruism is sufficiently weak.

To illustrate this fact, consider the issue of whether gay and lesbian couples should have the legal right to marry. Support for legislation that prohibits homosexual marriage may be consistent with benevolence if the authority is convinced that the spouses would come to regret their marriage later. For example, if she believes that homosexual marriage is an offense against God that would preclude the spouses from afterlife benefits that they, if they were aware of this, would not want to give up, then the benevolent authority would behave in a paternalistic fashion and remove this possibility.

In sharp contrast, among authorities who view an individual's choice of spouse as a matter of preferences, only authorities with little regard for the individual would engage in coercion. Coercion by a tyrannical authority does not represent an instance of paternalism, because she engages in coercion to promote her own self-interest (rather than the individual's self-interest). This would arise, for example, if an authority, for her own sake, dislikes the idea that the institution marriage is made available to homosexual couples; or if she has a preference for discrimination against homosexuals.

This highlights that whether a prohibition is enacted by a benevolent or tyrannical authority cannot be inferred by the prohibition itself; it is necessary to probe the authority's motivation or observe the strength of her altruism. Moreover, the same prohibition may be viewed as discrimination by some – who deem the issue as one of differing preferences – and as protection by others – who deem the issue as one of differing opinions. Those who view a woman's occupational choice as a matter of preferences, for example, would view legislation that precludes women from certain jobs as discrimination; on the contrary, those who fear that a woman who takes such a job is misinformed about its negative health consequences view the legislation as protection. In the 1970s, a law precluding women from heavy duty jobs was repealed in the U.S. on the grounds that it was discriminating. This highlights that the line between safety and discrimination is not only vague, but it may change over time, as predominant beliefs in society change.

3.7 Discussion of assumptions

Altruism. We use the simplest formulation for our purpose. A few choices are noteworthy. First, A evaluates the individuals' expected utilities using *her own prior*, π_A. This is essential. If A instead uses the individuals' priors, the results under opinion disagreement would be similar to those under preference disagreement. This is not a qualification of our results; on the contrary, it underscores our message. If A evaluates an individual's expected payoff using his prior, then A would want the individual to take the action that he deems best for himself, even though A is convinced that the individual later will come to regret this choice. Naturally, this reduces (opinion) disagreement to one of preferences. This speaks to the notion of altruism that we apply: We say that a more altruistic A cares more about *her own valuation of* the individuals' payoffs.

Second, A places the same weight on each individual. This is merely a simplification; suitable versions of our results arise so long as disagreement is non-negligible between A and individuals' she cares (more) about.

Third, an alternative formulation would be to let A place the weight $(1 - \varphi)$ on her own non-altruistic payoff (instead of the weight 1). Our main insights would continue to apply. Indeed, under opinion disagreement, the key feature driving our results – that an increase in φ raises A's valuation of the individuals' payoffs relative to her own non-altruistic payoff, and in particular relative to the cost of lying (or coercion) – remains in both formulations. Then, when A believes that lying benefits the population, stronger altruism makes her more willing to bear a given cost of lying. Under preference disagreement, it can be easily verified that greater altruism still makes truth-telling more attractive to lying (or coercion). Our results thus remain as long as, when φ increases, A's valuation of the cost of lying (or coercion) remains constant (as in our main formulation), decreases (as in the alternative formulation), or, more generally, does not rise too fast.

The costs of lying and coercion. We consider all possible costs of lying, $c \geq 0$. This encompasses the canonical models of cheap talk ($c = 0$) and verifiable information ($c = \infty$). In the authority game, we consider the case of $q > c$. Relaxing this assumption would not alter the insight that a sufficiently altruistic authority would prefer to be libertarian under preference disagreement, and would be paternalistic under opinion disagreement.

The representativeness of the advisor (authority). In the setting with preference disagreement, we consider general distributions, hence, all our results hold for cases when the advisor (authority) is representative of the population's median preference, i.e., when $f(b)$ has a median equal to zero. In the setting with opinion disagreement, we assume that the advisor (authority) is representative of the median opinion. This assumption makes lying and coercion more unattractive than if the advisor can hold an extreme, unrepresentative opinion. Indeed, fix some opinion distribution $g(\pi)$ and consider an advisor who gets $s = 1$. The advisor would always prefer to report $s = 1$ truthfully to more pessimistic individuals, with priors $\pi_i \leq \pi_A$, and to sufficiently close individuals, with priors $\pi_i > \pi_A$. Thus, an advisor with a median opinion π_A faces a greater loss from lying when sending a public message than does an advisor with an extreme pessimistic opinion. Similarly, a pessimistic authority may derive a greater benefit from mandating her preferred action, because she deems the individuals' beliefs skewed in the optimistic direction, and that intervention thus yields a considerable improvement in their expected welfare. As a result, for a given distribution $g(\pi)$, the advisor with a more extreme opinion would be more tempted to lie and coerce. Trivially, all our main results remain with the lower thresholds for lying and coercion.

3.8 Conclusion

We study a model where a population must rely on an altruistic advisor for information before making a decision. We show that the impact of altruism on communication fundamentally depends on the nature of disagreement. Altruism improves communication when the parties have different underlying preferences. In contrast, altruism destroys communication when the parties have different opinions: the advisor believes that the population (on average) will misinterpret a truthful report, so an altruistic advisor is inclined to lie in order to protect the population. If the advisor can force the individuals' actions, the altruistic advisor is *libertarian* under preference disagreement: she communicates truthfully and gives the individuals the liberty to choose. In contrast, under differences of opinion, the altruistic advisor is *paternalistic*: believing that she acts in the individuals' best interest, the advisor forces another action than the individuals would choose for themselves. Thus, whether coercion is perceived as justifiable, and deemed socially desirable, is intimately linked to whether a conflict is, consciously or subconsciously, framed in terms of preferences or opinions.

Our model offers a lens through which we can understand differing views on a government's right to regulate actions that can cause self-harm, and sheds light on why those who advocate restrictions in individual liberty deem such restrictions strictly better than information provision. The framework yields a precise prediction about what issues a regulator chooses to regulate, and on what issues she instead adopts a laissez-faire approach.

3.9 Appendix: Proofs

Proof of Proposition 8. We consider each type of disagreement in turn.

Preference disagreement. Assume that the population believes the message announced by the advisor. If the advisor received the signal $s = 1$ and reports it truthfully to the population, then agent i optimally picks the action $a_i(1) = p(1) + b_i$, where $p(1)$ is the posterior belief that $\theta = 1$, i.e., $p(1) = \Pr(\theta = 1|s = 1)$. Hence, the advisor's expected utility is

$$\begin{aligned}
\mathbb{E}\left[U_A(a(1),\theta)|s=1\right] &= -p(1)\left[\int_{-\infty}^{+\infty}[(p(1)+b_i-1)^2+\varphi(p(1)-1)^2]f(b_i)db_i\right] \\
&\quad -(1-p(1))\left[\int_{-\infty}^{+\infty}[(p(1)+b_i)^2+\varphi p(1)^2]f(b_i)db_i\right] \\
&= -p(1)\left[(p(1)-1)^2+2\bar{b}(p(1)-1)+\overline{b^2}+\varphi(p(1)-1)^2\right] \\
&\quad -(1-p(1))\left[p(1)^2+2\bar{b}p(1)+\overline{b^2}+\varphi p(1)^2\right] \\
&= -p(1)\left[(p(1)+\bar{b}-1)^2+\varphi(p(1)-1)^2\right] \\
&\quad -(1-p(1))\left[(p(1)+\bar{b})^2+\varphi p(1)^2\right]+\bar{b}^2-\overline{b^2}.
\end{aligned}$$

If the advisor decides to misreport signal $s = 1$, it induces individual i to choose $a_i(0) = p(0) + b_i$ and results in the following advisor's expected utility:

$$\begin{aligned}
\mathbb{E}\left[U_A(a(0),\theta)|s=1\right] &= -p(1)\left[(p(0)+\bar{b}-1)^2+\varphi(p(0)-1)^2\right] \\
&\quad -(1-p(1))\left[(p(0)+\bar{b})^2+\varphi p(0)^2\right]+\bar{b}^2-\overline{b^2}-c.
\end{aligned}$$

Truth-telling is preferred when

$$\begin{aligned}
& -p(1)\left[(p(1)+\bar{b}-1)^2+\varphi(p(1)-1)^2\right]-(1-p(1))\left[(p(1)+\bar{b})^2+\varphi p(1)^2\right] \\
\geq\ & -p(1)\left[(p(0)+\bar{b}-1)^2+\varphi(p(0)-1)^2\right]-(1-p(1))\left[(p(0)+\bar{b})^2+\varphi p(0)^2\right]-c.
\end{aligned}$$

This condition is equivalent to the one, where the advisor communicates to just one individual with the preference difference of $\bar{b}$, and can be rewritten as

$$\varphi \geq -1+\frac{2\bar{b}}{p(1)-p(0)}-\frac{c}{(p(1)-p(0))^2}. \tag{TT1$_{\text{pr}}$}$$

Similarly, the advisor will report $s = 0$ truthfully whenever

$$\varphi \geq -1 - \frac{2\bar{b}}{p(1) - p(0)} - \frac{c}{(p(1) - p(0))^2}. \tag{TT0$_{\mathrm{pr}}$}$$

Thus, truth-telling is an equilibrium outcome if and only if the altruism level is substantial, $\varphi \geq \overline{\varphi}(c, \bar{b}) = \max\{-1 + \frac{2|\bar{b}|}{p(1)-p(0)} - \frac{c}{(p(1)-p(0))^2},\}$. Clearly, $\overline{\varphi}(c, \bar{b})$ (weakly) decreases in c.

Opinion disagreement. Start from a simple opinion distribution with two types of individuals, t_1 and t_2, with priors $\pi_1 \leq \pi_A = 0.5 \leq \pi_2$. The types are equally prevalent, i.e., $g(\pi_1) = g(\pi_2) = 0.5$. Assume that lying is costless, $c = 0$, and the population believes the reported signal. Consider the advisor who gets $s = 1$, which results in her posterior belief equal to γ. If the advisor could differentiate messages between the people, then she would always prefer to report $s = 1$ truthfully to t_1 with the prior $\pi_1 \leq 1/2$; and might want to misreport to t_2 when π_2 is sufficiently close to 1. Overall, the advisor would prefer to lie if the benefit from lying to t_2 outweighs the loss from lying to t_1. To argue that the lying region $L1$ has the form as shown in Figure 3.1, below we consider the properties of the loss and the benefit functions (the analysis of $s = 0$ case and a respective region $L0$ can be performed in a similar way).

First, study the loss from lying. Denote the action choices of t_1 after messages 1 and 0 by $a_1 = p_{\pi_1}(1) = \frac{\pi_1\gamma}{\pi_1\gamma+(1-\pi_1)(1-\gamma)}$ and $a_0 = p_{\pi_1}(0) = \frac{\pi_1(1-\gamma)}{\pi_1(1-\gamma)+(1-\pi_1)\gamma}$.[31] The advisor's expected loss from sending the message $m = 0$ is

$$\begin{aligned} l(\pi_1) &= -\gamma(a_1 - 1)^2 - (1-\gamma)a_1^2 + \gamma(a_0 - 1)^2 + (1-\gamma)a_0^2 \\ &= -(a_1 - \gamma)^2 + (a_0 - \gamma)^2. \end{aligned}$$

First, we show that the loss function is strictly concave. Consider the second derivative of $l(\pi_1)$ (in the following derivations the prime and the double prime symbols denote the corresponding

[31] Similarly to the previously introduced notation, $p_\pi(s)$ stands for the posterior belief of an individual with prior π about θ: $\Pr_\pi(\theta = 1|s)$.

derivatives with respect to π_1):

$$\begin{aligned} l''(\pi_1) &= 2\left[-a_1'(a_1-\gamma)+a_0'(a_0-\gamma)\right]' \\ &= 2\left[-(a_1')^2-a_1''(a_1-\gamma)+(a_0')^2+a_0''(a_0-\gamma)\right]. \end{aligned}$$

In this expression,

$$\begin{aligned} (a_1')^2+a_1''(a_1-\gamma) &= \frac{\gamma^2(1-\gamma)^2}{[\pi_1\gamma+(1-\pi_1)(1-\gamma)]^4}-\frac{2\gamma(1-\gamma)(2\gamma-1)}{[\pi_1\gamma+(1-\pi_1)(1-\gamma)]^3}(a_1-\gamma) \\ &= \frac{\gamma^2(1-\gamma)}{[\pi_1\gamma+(1-\pi_1)(1-\gamma)]^4}\left[1-\gamma-2(2\gamma-1)(1-\gamma)(2\pi_1-1)\right] \\ &= \frac{\gamma^2(1-\gamma)^2}{[\pi_1\gamma+(1-\pi_1)(1-\gamma)]^4}\left[1+2(2\gamma-1)(1-2\pi_1)\right] \geq 0. \end{aligned}$$

Similarly,

$$\begin{aligned} (a_0')^2+a_0''(a_0-\gamma) &= \frac{\gamma^2(1-\gamma)^2}{[\pi_1(1-\gamma)+(1-\pi_1)\gamma]^4}-\frac{2\gamma(1-\gamma)(1-2\gamma)}{[\pi_1(1-\gamma)+(1-\pi_1)\gamma]^3}(a_0-\gamma) \\ &= \frac{\gamma(1-\gamma)}{[\pi_1(1-\gamma)+(1-\pi_1)\gamma]^4}\left[\gamma(1-\gamma)+2(2\gamma-1)(\pi_1(1-\gamma)^2-(1-\pi_1)\gamma^2)\right]. \end{aligned}$$

Condition $\pi_1 \leq 0.5$ ensures that

$$\pi_1\gamma+(1-\pi_1)(1-\gamma) \leq \pi_1(1-\gamma)+(1-\pi_1)\gamma$$

and

$$\gamma(1-\gamma)\left[1+2(2\gamma-1)(1-2\pi_1)\right] > \gamma(1-\gamma)+2(2\gamma-1)(\pi_1(1-\gamma)^2-(1-\pi_1)\gamma^2).$$

Hence, $(a_1')^2+a_1''(a_1-\gamma) > (a_0')^2+a_0''(a_0-\gamma)$, meaning that $l''(\pi_1)<0$.

Second, the loss function achieves its minimum of 0 at $\pi_1=0$. This, in particular, means that $l(\pi_1)$ increases at $\pi_1=0$. Indeed, t_1 with the prior $\pi_1=0$ does not respond to the revealed signal and always chooses action 0. For any other $0<\pi_1\leq 0.5$ different messages induce different actions $a_0 \neq a_1$, leading to a strictly positive loss from misreporting.

Finally, $l(\pi_1)$ decreases at $\pi_1=0.5$. To see this, note that at $\pi_1=0.5$ the chosen actions are

$a_1 = \gamma$ and $a_0 = 1-\gamma$. Hence, $l'(\pi_1) = -a_1'(a_1-\gamma) + a_0'(a_0-\gamma)$ becomes $a_0'(1-2\gamma) < 0$.

Now consider the potential benefit of misreporting to t_2 with the prior of π_2. As before, let $\tilde{a}_1 = p_{\pi_2}(1)$ and $\tilde{a}_0 = p_{\pi_2}(0)$ denote the actions choices of t_2 after messages 1 and 0, respectively. The potential benefit from misreporting is

$$b(\pi_2) = (\tilde{a}_1-\gamma)^2 - (\tilde{a}_0-\gamma)^2.$$

The potential benefit realizes with positive values for sufficiently extreme priors π_2 and translates into a loss when the prior π_2 is close to the advisor's priors of 0.5. Regarding the properties of the potential benefit function, we, first, show that it is strictly increasing until $\pi_2 = \frac{\gamma^2}{\gamma^2+(1-\gamma)^2}$, s.t. $\tilde{a}_0 = \gamma$ under this prior. Indeed, consider the first derivative:

$$b'(\pi_2) = 2\left[\tilde{a}_1'(\tilde{a}_1-\gamma) - \tilde{a}_0'(\tilde{a}_0-\gamma)\right].$$

It is strictly greater than 0 when $\pi_2 \leq \frac{\gamma^2}{\gamma^2+(1-\gamma)^2}$, because $\tilde{a}_1'$, $\tilde{a}_0' > 0$, $\tilde{a}_1 \geq \gamma$ and $\tilde{a}_0 \leq \gamma$.

Next, we show that $b(\pi_2)$ is strictly concave for priors π_2 s.t. $\tilde{a}_0 \geq \gamma$, i.e., $\pi_2 \geq \frac{\gamma^2}{\gamma^2+(1-\gamma)^2}$. The second derivative of $b(\pi_2)$ is

$$b''(\pi_2) = 2\left[(\tilde{a}_1')^2 + \tilde{a}_1''(\tilde{a}_1-\gamma) - (\tilde{a}_0')^2 - \tilde{a}_0''(\tilde{a}_0-\gamma)\right].$$

Similarly to what was derived before for t_1,

$$(\tilde{a}_1')^2 + \tilde{a}_1''(\tilde{a}_1-\gamma) = \frac{\gamma^2(1-\gamma)^2}{[\pi_2\gamma+(1-\pi_2)(1-\gamma)]^4}\left[1+2(2\gamma-1)(1-2\pi_2)\right]$$

and

$$(\tilde{a}_0')^2 + \tilde{a}_0''(\tilde{a}_0-\gamma) = \frac{\gamma(1-\gamma)}{[\pi_2(1-\gamma)+(1-\pi_2)\gamma]^4}\left[\gamma(1-\gamma)+2(2\gamma-1)(\pi_2(1-\gamma)^2-(1-\pi_2)\gamma^2)\right].$$

The fact that $\pi_2 \geq 0.5$ ensures

$$\pi_2\gamma + (1-\pi_2)(1-\gamma) \geq \pi_2(1-\gamma) + (1-\pi_2)\gamma,$$

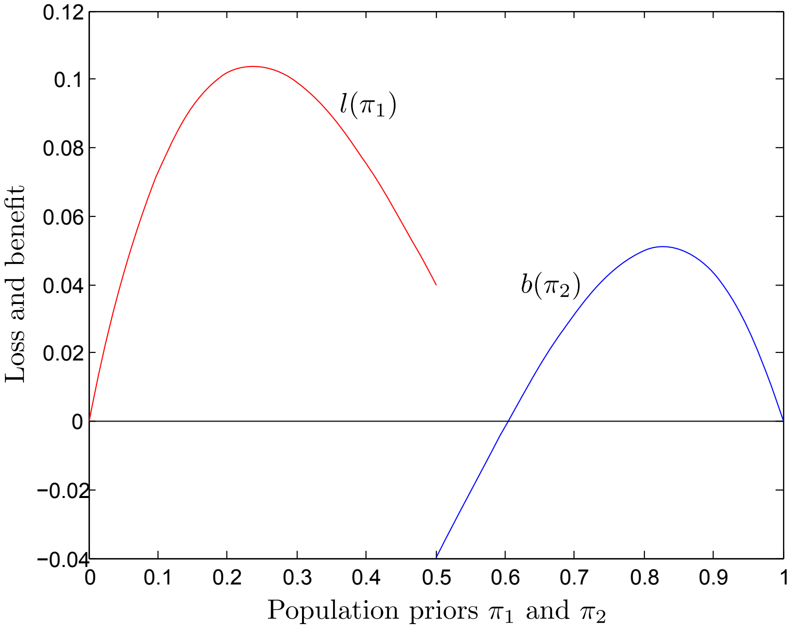


Figure 3.5: Loss and benefit functions from misreporting $s = 1$, $\gamma = 0.6$.

while condition $\pi_2 \geq \frac{\gamma^2}{\gamma^2+(1-\gamma)^2} > \gamma$ guarantees that

$$\max\{0, \gamma(1-\gamma)[1+2(2\gamma-1)(1-2\pi_2)]\} < \gamma(1-\gamma)+2(2\gamma-1)(\pi_2(1-\gamma)^2-(1-\pi_2)\gamma^2).$$

Hence, $(\tilde{a}_1')^2 + \tilde{a}_1''(\tilde{a}_1 - \gamma) < (\tilde{a}_0')^2 + \tilde{a}_0''(\tilde{a}_0 - \gamma)$, meaning that $b''(\pi_2) < 0$.

Finally, note that the benefit from lying decreases for π_2 sufficiently close to 1 and is 0 when $\pi_2 = 1$: $b'(1) < 0$ and $b(1) = 0$.

The described properties of loss and benefit functions, in particular, imply that $l(\pi_1)$ and $b(\pi_2)$ achieve their unique points of maximum in interiors $(0, 0.5)$ and $(0.5, 1)$, respectively. Moreover, the maximum of $l(\pi_1)$ exceeds the maximum of $b(\pi_2)$, because parabola $-(\pi-\gamma)^2$ has its peak at $\gamma > 0.5$. Typical loss and benefit functions are illustrated in Figure 3.5.

The difference in the advisor's expected payoffs from lying and truthful reporting is $\frac{1}{2}b(\pi_2) - \frac{1}{2}l(\pi_1)$. The summarized properties of $l(\pi_1)$ and $b(\pi_2)$ allow to see why the region $L1$ where $b(\pi_2) > l(\pi_1)$ has a typical form as shown in Figure 3.1.

Start from $L1_{\text{left}}$. If $\pi_1 = 0$, then $l(0) = 0$ but $b(\pi_2) \geq 0$ for sufficiently large π_2. Now raise π_1 by a little bit. Then $l(\pi_1)$ becomes strictly positive, while the properties of $b(\pi_2)$ ensure that the interval of π_2 for which $b(\pi_2) > l(\pi_1)$ shrinks. Because $l(\pi_1)$ increases until it reaches its maximum, raising π_1 further leads to greater shrinking of the interval $\{\pi_2 : \ b(\pi_2) > l(\pi_1)\}$, until it disappears completely (provided that the maximum of $l(\pi_1)$ exceeds the maximum of $b(\pi_2)$).

Now consider $L1_{\text{right}}$. This region is non-empty if and only if $l(0.5) < \max_{\pi_2} b(\pi_2)$. Assume that this condition is satisfied. To understand the shape of $L1_{\text{right}}$, start by considering $\pi_1 = 0.5$. The properties of $b(\pi_2)$ ensure that $\{\pi_2 : \ b(\pi_2) > l(\pi_1)\}$ is a non-empty interval inside $(0.5, 1)$. Now start decreasing π_1. Then the interval $\{\pi_2 : \ b(\pi_2) > l(\pi_1)\}$ shrinks until it disappears completely.

Now we show that the condition $l(0.5) < \max_{\pi_2} b(\pi_2)$ holds if γ sufficiently close to 0.5. The loss at $\pi_1 = 0.5$ is

$$l(0.5) = (a_0 - \gamma)^2 = (2\gamma - 1)^2.$$

Now consider the benefit when:

$$\begin{aligned} b(\pi_2) &= (\tilde{a}_1 - \gamma)^2 - (\tilde{a}_0 - \gamma)^2 \\ &= (\tilde{a}_1 - \tilde{a}_0)(\tilde{a}_1 + \tilde{a}_0 - 2\gamma) \\ &= \frac{(2\gamma - 1)\pi_2(1 - \pi_2)}{(\pi_2\gamma + (1 - \pi_2)(1 - \gamma))(\pi_2(1 - \gamma) + (1 - \pi_2)\gamma)}(\tilde{a}_1 + \tilde{a}_0 - 2\gamma). \end{aligned}$$

Because $\frac{l(0.5)}{2\gamma - 1} = 2\gamma - 1 \to 0$ as $\gamma \to 0.5$, while $\frac{b(\pi_2)}{2\gamma - 1} \to \pi_2(1 - \pi_2)(2\pi_2 - 1) > 0$ as $\gamma \to 0.5$, then $l(0.5) < \max_{\pi_2} b(\pi_2)$ if γ is small enough.

On the other hand, $l(0.5) > \max_{\pi_2} b(\pi_2)$ for any sufficiently large γ. Indeed, $l(0.5) = (2\gamma - 1)^2$ increases as γ increases, while $\max_{\pi_2} b(\pi_2)$ is strictly less than $(1 - \gamma)^2$. As a result, when $\gamma \geq \frac{2}{3}$, the loss $l(0.5)$ strictly exceeds the benefit for $b(\pi_2)$ for any π_2.[32]

When lying is costly, $c > 0$, the advisor will misreport the signal $s = 1$ whenever $\varphi(\frac{1}{2}b(\pi_2) - \frac{1}{2}l(\pi_1)) > c$. Clearly, the advisor remains credible for all $(\pi_1, \pi_2) \in TT$ (see Figure 3.1) independently of the levels of $\varphi > 0$ and $c > 0$. Further, for any $c > 0$ and $(\pi_1, \pi_2) \in L1 \cup L0$, there exists a threshold level of altruism $\overline{\varphi}(c, \pi_1, \pi_2)$ such that a FRE exists if and only if $\varphi \leq \overline{\varphi}(c, \pi_1, \pi_2)$.

This result that a FRE exists if and only if φ is below some threshold can be easily generalized to some other distributions of priors $g(\pi)$ as well. To see this, take some distribution $g(\pi)$ and define the two distribution functions g_1 and g_2 so that $g_1(\pi_1) = 2g(\pi_1)$ if $\pi_1 < 0.5$ and $g_1(\pi_1) = 0$

[32] While we do not show analytically that $l(0.5) < \max_{\pi_2} b(\pi_2)$ *if only if* γ is below a specific threshold, computational exercises suggest that this is the case.

if $\pi_1 > 0.5$; similarly, $g_2(\pi_2) = 2g(\pi_2)$ if $\pi_2 > 0.5$ and $g_2(\pi_2) = 0$ if $\pi_2 < 0.5$.[33] Clearly, functions g_1 and g_2 are uniquely defined for each distribution g. When lying is costless, $c = 0$, the advisor will lie after the signal $s = 1$ as long as enough mass of joint distribution $g_1(\pi_1)g_2(\pi_2)$ is inside $L1$. Similarly, the advisor will misreport $s = 0$ provided that sufficient mass of the joint distribution is concentrated inside region $L0$. This implies that the set $\mathcal{G}$ in Proposition 8 is non-empty and consists of uncountably many members. For distributions $g \in \mathcal{G}$, a FRE may exist when $c > 0$. For any given cost of lying, however, increasing the level of altruism eventually destroys truthful communication, because greater φ makes the net benefit from lying larger relative to the cost c. Clearly, $\overline{\varphi}(c, g)$ (weakly) increases in c.

Consider now some opinion distribution $g(\pi) \notin \mathcal{G}$. Truth-telling is incentive compatible for some $\varphi > 0$ when $c = 0$, hence, it is incentive compatible for any $\varphi \geq 0$ (when $c = 0$). Because, for any given φ, an increase in c makes truth-telling even more attractive, a FRE exists for all $c \geq 0$ and $\varphi \geq 0$. **QED.**

Proof of Proposition 9. We consider each type of disagreement in turn.

Preference disagreement. The proof consists of two steps, each described by a corresponding lemma.

Lemma 10. *For any cost of coercion $q \geq 0$ and preference distribution $f(b)$ there exists a threshold $\varphi_{TT}(q, \bar{b}, \overline{b^2})$, s.t. a FRE exists if and only if $\varphi \geq \varphi_{TT}(q, \bar{b}, \overline{b^2})$.*

Proof. Assume that the authority gets the signal s. If the authority coerces, she chooses the action a that maximizes her expected utility given by:

$$\begin{aligned}\mathbb{E}(U_A(a,\theta)|s) \quad &= \quad -p(s)\left[\int_{-\infty}^{+\infty}[(a-1)^2 + \varphi(a-1-b_i)^2]f(b_i)db_i\right] \\ &\quad -(1-p(s))\left[\int_{-\infty}^{+\infty}[a^2 + \varphi(a-b_i)^2]f(b_i)db_i\right] - q.\end{aligned}$$

[33] The values $g_1(0.5)$ and $g_2(0.5)$ are defined so that the functions g_1 and g_2 integrate (in discrete case, sum up) to 1 over the interval [0.1].

Thus, the authority optimally imposes the action:

$$a_A(s) = p(s) + \frac{\varphi}{1+\varphi}\bar{b}.$$

Her expected utility from coercion is

$$\begin{aligned}
\mathbb{E}(U_A(a,\theta)|s) &= -p(s)\left[(1+\varphi)(a_A(s)-1)^2 - 2\varphi\bar{b}(a_A(s)-1) + \varphi\overline{b^2}\right] \\
&\quad -(1-p(s))\left[(1+\varphi)(a_A(s))^2 - 2\varphi\bar{b}a_A(s) + \varphi\overline{b^2}\right] - q \\
&= -(1+\varphi)\mathbb{E}\left((a_A(s)-\theta)^2|s\right) + 2\varphi\bar{b}\left[a_A(s) - p(s)\right] - \varphi\overline{b^2} - q \\
&= -(1+\varphi)\left[\left(\frac{\varphi}{1+\varphi}\bar{b}\right)^2 + p(s)(1-p(s))\right] + 2\frac{\varphi^2}{1+\varphi}\bar{b}^2 - \varphi\overline{b^2} - q \\
&= \frac{\varphi^2}{1+\varphi}\bar{b}^2 - (1+\varphi)p(s)(1-p(s)) - \varphi\overline{b^2} - q.
\end{aligned}$$

If instead the authority communicates some message m to the population, then the population's posterior becomes $\hat{p} = p(m)$ (by symmetry, in any equilibrium all individuals hold the same posteriors after observing the authority's message). Hence, each individual i picks his preferred action $a_i(m) = \hat{p} + b_i$, which generates the following expected payoff to the authority:

$$\begin{aligned}
\mathbb{E}\left(U_A(a(m),\theta)|s\right) &= -p(s)\left[\int_{-\infty}^{+\infty}(\hat{p}+b_i-1)^2 f(b_i)db_i + \varphi(\hat{p}-1)^2\right] \\
&\quad -(1-p(s))\left[\int_{-\infty}^{+\infty}(\hat{p}+b_i)^2 f(b_i)db_i + \varphi\hat{p}^2\right] \\
&= -p(s)\left[(1+\varphi)(\hat{p}-1)^2 + 2\bar{b}(\hat{p}-1) + \overline{b^2}\right] \\
&\quad -(1-p(s))\left[(1+\varphi)\hat{p}^2 + 2\bar{b}\hat{p} + \overline{b^2}\right] \\
&= -(1+\varphi)(\hat{p}^2 - 2\hat{p}p(s) + p(s)) - 2\bar{b}(\hat{p}-p(s)) - \overline{b^2} \\
&= -(1+\varphi)(\hat{p}-p(s))^2 - (1+\varphi)p(s)(1-p(s)) - 2\bar{b}(\hat{p}-p(s)) - \overline{b^2}.
\end{aligned}$$

Communication generates greater expected payoff to the authority than coercion if

$$\frac{\varphi^2}{1+\varphi}\bar{b}^2 - \varphi\overline{b^2} - q \leq -(1+\varphi)(\hat{p}-p(s))^2 - 2\bar{b}(\hat{p}-p(s)) - \overline{b^2}. \tag{3.3}$$

In particular case of truthful advising this boils down to

$$\begin{aligned}\frac{\varphi^2}{1+\varphi}\bar{b}^2 - \varphi\overline{b^2} + \overline{b^2} - q &\leq 0, \\ -\varphi^2(\overline{b^2} - \bar{b}^2) - \varphi q + \overline{b^2} - q &\leq 0. \end{aligned} \tag{3.4}$$

The parabola in the LHS is concave and achieves its maximum at $\varphi \leq 0$. This implies that the LHS decreases in φ for $\varphi \geq 0$. In particular, the authority prefers to act as a truth-telling advisor if and only if the level of altruism is sufficiently large: $\varphi \geq \varphi_{TT}(q, \bar{b}, \overline{b^2})$. The threshold $\varphi_{TT}(q, \bar{b}, \overline{b^2}) \leq 1$ if the average bias is sufficiently small:$\frac{\bar{b}^2}{2} - q \leq 0$.

Note, that if coercion is costless, $q = 0$, then the truthful advising is preferred for sufficiently large φ, for any non-degenerate distribution ($\bar{b}^2 \neq \overline{b^2}$), i.e., $\varphi_{TT}(0, \bar{b}, \overline{b^2}) < +\infty$; while coercion is always preferred for any degenerate distribution ($\bar{b}^2 = \overline{b^2}$), i.e., $\varphi_{TT}(0, b, b^2) = +\infty$.

Clearly, $\varphi_{TT}(q, \bar{b}, \overline{b^2})$ strictly exceeds 0 and decreases in q for $q < \overline{b^2}$, and $\varphi_{TT}(q, \bar{b}, \overline{b^2}) = 0$ when $q \geq \overline{b^2}$ □

Lemma 11. *For every $f(b)$ and $q < q_C(\bar{b}, \overline{b^2})$ there exists a threshold $\varphi_C(q, \bar{b}, \overline{b^2}) > 0$, s.t. for any $\varphi < \varphi_C(q, \bar{b}, \overline{b^2})$ every equilibrium outcome involves the authority imposing her preferred action with non-zero probability. Moreover, for every b and $q < q_{CC}(\bar{b}, \overline{b^2})$ there exists a threshold $\varphi_{CC}(q, \bar{b}, \overline{b^2}) > 0$, s.t. for any $\varphi < \varphi_{CC}(q, \bar{b}, \overline{b^2})$ there is a unique equilibrium, in which the authority coerces the individual for each signal $s \in \{0, 1\}$.*

Proof. As regards the first part of the Lemma, we show that sufficiently low cost q and altruism φ preclude the existence of communication equilibria. We study two different types of communication equilibria: (i) only one population's posterior belief $\hat{p}$ is induced, and (ii) two different posterior beliefs $\hat{p}_0 < \hat{p}_1$ are induced.

In case (i), the authority's communication leads to the same population's posterior belief independently of the underlying signal, i.e., no informative communication takes place and the posterior belief necessarily coincides with the prior, $\hat{p} = \pi$. This means that each individual i picks an action $a_i = \pi + b_i$. Such a communication equilibrium fails to exist when, after some signal s, the authority would rather exercise her power to choose her preferred action, i.e., when inequality (3.3) hold with the opposite sign:

$$\frac{\varphi^2}{1+\varphi}\bar{b}^2 - \varphi\overline{b^2} - q > -(1+\varphi)(\pi - p(s))^2 - 2\bar{b}(\pi - p(s)) - \overline{b^2}.$$

Assume that the authority does not incorporate the individual's payoff into her utility: $\varphi = 0$. Then, for sufficiently low cost $q < \max_s[(\pi - p(s))^2 + 2\bar{b}(\pi - p(s)) + \overline{b^2}]$, the authority coerces the individual for at least one signal s.

Now consider case (ii), in which the induced posterior beliefs of the population are different, $\hat{p}_0 < \hat{p}_1$. The authority cannot be indifferent between these beliefs after both signal realizations. Thus, for these actions to be induced in equilibrium, it must be the case that the authority mixes between messages for no more than one signal. It implies that $\hat{p}_s = p(s)$ for at least one signal s. Hence, a non-altruistic authority who obtained s will choose to impose her preferred action whenever inequality (3.4) breaks down, i.e., $q < \overline{b^2}$. Clearly, she will still do so for sufficiently low levels of altruism.

Together, cases (i) and (ii) imply that for a sufficiently low coercion cost, $q < q_C(\bar{b}, \overline{b^2})$, and altruism level, $\varphi < \varphi_C(q, \bar{b}, \overline{b^2})$, each equilibrium involves coercion with a non-zero probability. Clearly, $\varphi_C(q, \bar{b}, \overline{b^2})$ decreases in q for $q < q_C(\bar{b}, \overline{b^2})$.

As regards the second part of the Proposition, we argue that for sufficiently low q and φ no communication is sustainable in equilibrium. Below we consider two possibilities of how communication can arise.

First, assume that the authority communicates with positive probability for both signal realizations $s = 0, 1$. If only one population's posterior $\hat{p} \in [p(0), p(1)]$ is induced in equilibrium, then a non-altruistic authority would strictly prefer to coerce for at least one signal when

$$q < q(\hat{p}, \bar{b}, \overline{b^2}) = \max_s[(\hat{p} - p(s))^2 + 2\bar{b}(\hat{p} - p(s)) + \overline{b^2}]. \tag{3.5}$$

Clearly, $q(\hat{p}, \bar{b}, \overline{b^2}) > 0$ for every $\hat{p}$. Because $q(\hat{p}, \bar{b}, \overline{b^2})$ is a continuous function of $\hat{p}$, there exists $0 < \bar{q}(\bar{b}, \overline{b^2}) = \min_{\hat{p} \in [p(0),p(1)]} q(\hat{b}, \bar{b}, \overline{b^2})$. Inequality (3.5) is strict for any $q < \bar{q}(\bar{b}, \overline{b^2})$ and $\hat{p} \in [p(0), p(1)]$; hence, the authority still prefers to coerce whenever the level of altruism is below the threshold

$\varphi(q,\hat{p},\bar{b},\overline{b^2}) > 0$.[34] Because $\varphi(q,\hat{p},\bar{b},\overline{b^2})$ is continuous in $\hat{p} \in [p(0),p(1)]$, there exists

$$0 < \overline{\varphi}(q,\bar{b},\overline{b^2}) = \min_{\hat{p}\in[p(0),p(1)]} \varphi(q,\hat{p},\bar{b},\overline{b^2}).$$

As a result, for $q < \bar{q}(\bar{b},\overline{b^2})$ and $\varphi < \overline{\varphi}(q,\bar{b},\overline{b^2})$, there exists no equilibrium with non-zero communication after both signals and only one induced posterior belief. Next, if two different posterior beliefs $\hat{p}_0 < \hat{p}_1$ are induced in equilibrium, then the authority strictly prefers inducing one action over the other for at least one signal, meaning that $\hat{p}_s = p(s)$ for some $s \in \{0,1\}$. Clearly, such an equilibrium cannot exist for q lower than $\bar{b}^2$ and sufficiently low φ.

Second, suppose that the authority communicates with non-zero probability for one signal and coerces with certainty for the other signal. For example, assume that the authority sometimes communicates after $s = 1$ and always coerces after $s = 0$. In this case, the mere fact that the authority didn't impose an action indicates to the population that the signal is $s = 1$. As a result, the induced population's posterior $\hat{p}$ necessarily equals $p(1)$. Such an equilibrium fails to exist for q lower than $\overline{b^2}$ and sufficiently low φ. The case when the authority sometimes communicates the low signal $s = 0$ and always coerces when $s = 1$ is analogous. It follows that for every $q < \overline{b^2}$, communicating after only one signal cannot be an equilibrium if φ is sufficiently low.

Combining the results of the two cases, when the coercion cost q and altruism level φ are below the respective bounds $q_{CC}(\bar{b},\overline{b^2})$ and $\varphi_{CC}(q,\bar{b},\overline{b^2})$, the unique equilibrium is the one where the authority always coerces the individual. □

Denoting $\bar{q} = q_{CC}(\bar{b},\overline{b^2})$, Lemma 10 and Lemma 11 imply the preference disagreement result of Proposition 9. Evidently, $\varphi_{CC}(q,\bar{b},\overline{b^2})$, $\varphi_{C}(q,\bar{b},\overline{b^2})$, and $\varphi_{TT}(q,\bar{b},\overline{b^2})$ are decreasing in q.

Clearly, a FRE is preferred by the individuals to any other (pure or mixed strategies) PBE. Now we show that a FRE is preferred by the authority to any other pure strategies PBE, when φ exceeds $\varphi_{TT}(q,\bar{b},\overline{b^2})$. First, the FRE is always preferred to a purely uninformative advising, where the same message is sent after both signals. Indeed, the FRE generates a greater expected utility

[34] Assuming that the maximum in (3.5) is achieved at $\hat{s}$, the threshold $\varphi(q,\hat{p},\bar{b},\overline{b^2})$ is determined as a closest to 0 positive root of $\frac{\varphi^2}{1+\varphi}\bar{b}^2 - \varphi\overline{b^2} - q = -(1+\varphi)(\hat{p}-p(s))^2 - 2\bar{b}(\hat{p}-p(s)) - \overline{b^2}$.

to the advisor whenever

$$\begin{aligned} & (\pi\gamma+(1-\pi)(1-\gamma))\left[-(1+\varphi)(\pi-p(1))^2-2\bar{b}(\pi-p(1))\right] \\ + \quad & (\pi(1-\gamma)+(1-\pi)\gamma)\left[-(1+\varphi)(\pi-p(0))^2-2\bar{b}(\pi-p(0))\right] \leq 0. \end{aligned}$$

This inequality can be rewritten as

$$-(1+\varphi)\left[(\pi\gamma+(1-\pi)(1-\gamma))(\pi-p(1))^2+(\pi(1-\gamma)+(1-\pi)\gamma)(\pi-p(0))^2\right] \leq 0,$$

which satisfied for all φ. Second, the FRE is preferred to equilibria where the authority coerces after some or after both signals. Indeed, as the previous analysis illustrates, after any signal s the authority prefers truth-telling to coercion for $\varphi \geq \varphi_{TT}(q, \bar{b}, \overline{b^2})$.

Opinion disagreement. The proof is split in two steps; each step is formulated as a corresponding lemma.

Lemma 12. *For any opinion distribution $g \notin \mathcal{G}$, there exists a threshold $\varphi_{TT}(q, g)$, s.t. a FRE exists if and only if $\varphi \leq \varphi_{TT}(q, g)$.*

Proof. Consider some opinion distribution $g(\pi) \notin \mathcal{G}$. That is, truthful reporting to population described by $g(\pi)$ is incentive compatible when $c = 0$, given that the individuals believe the message reported. Truth-telling is preferred to coercing after signal s if

$$-\varphi \int_0^1 \left[p_A(p_i-1)^2+(1-p_A)(p_i)^2\right] \geq -q-\varphi \int_0^1 \left[p_A(p_A-1)^2+(1-p_A)(p_A)^2\right] g(\pi_i) d\pi_i,$$

where $p_A = p_A(s)$ and $p_i = p_i(s) = a_i(s)$ are the authority's and individual i's posterior beliefs that $\theta = 1$ after the signal s. This condition can be rewritten as

$$\varphi \int_{\pi_i \neq 0.5} \left[p_A(p_i-1)^2+(1-p_A)(p_i)^2-p_A(p_A-1)^2-(1-p_A)(p_A)^2\right] g(\pi_i) d\pi_i \leq q.$$

The integral in the LHS is strictly positive, because $p_A(p_i-1)^2+(1-p_A)(p_i)^2 > p_A(p_A-1)^2+(1-p_A)(p_A)^2$ for any $p_i \neq p_A$ (i.e., $\pi_i \neq 0.5$) and $\int_{\pi_i \neq 0.5} g(\pi_i) d\pi_i > 0$. Thus, a FRE exists if and only if φ is below a threshold $\varphi_{TT}(q, g)$. Note that $\varphi_{TT}(q, g) > 0$ when $q > 0$ and $\varphi_{TT}(0, g) = 0$.

□

Lemma 13. *For any opinion distribution $g(\pi)$ and $q \geq 0$, there exist thresholds $\varphi_C(q,g) \leq \varphi_{CC}(q,g)$. For $\varphi > \varphi_{CC}(q,g)$, there exists a unique equilibrium, in which the authority coerces with probability one after each signal s. For $\varphi > \varphi_C(q,g)$, every equilibrium involves the authority coercing with strictly positive probability.*

Proof. First, we prove that all equilibria with communication after both signal realizations break down for a sufficiently high level of altruism $\varphi > \varphi_C(q,g)$. To show this, we distinguish between (i) communication equilibria where each individual i takes only one action $\hat{a}_i$, and (ii) communication equilibria where each individual i takes two different actions $\hat{a}_{i,0} < \hat{a}_{i,1}$ with positive probabilities.

In case (i), the communication is necessarily uninformative, and hence, each individual i chooses an action equal to his prior belief, $\hat{a}_i = \pi_i$. Coercion after some signal s is preferred to such communication whenever

$$\varphi \int_{\pi_i \neq p_A} \left[p_A(\pi_i - 1)^2 + (1-p_A)(\pi_i)^2 - p_A(p_A-1)^2 - (1-p_A)(p_A)^2\right] g(\pi_i) d\pi_i \geq q,$$

where $p_A = p_A(s)$. The integral in the LHS is strictly positive, because $p_A(\pi_i - 1)^2 + (1-p_A)(\pi_i)^2 > p_A(p_A-1)^2 + (1-p_A)(p_A)^2$ for any $\pi_i \neq p_A$ and $\int_{\pi_i \neq p_A} g(\pi_i) d\pi_i > 0$. Thus, coercion is strictly preferred for sufficiently large φ, so such a communication equilibrium breaks down.

In case (ii), the communicating authority induces the individuals to take different actions $\hat{a}_{i,0} < \hat{a}_{i,1}$. Clearly, rational behavior on the part of each individual i ensures that $\hat{a}_{i,0}$, $\hat{a}_{i,1} \in [p_i(0), p_i(1)]$. Because $\int_{\pi \neq 0.5} g(\pi) d\pi > 0$ and $0.5 = \pi_A$ is a median of $g(\pi)$, w.l.o.g. we assume that $\int_{\pi > 0.5+\varepsilon} g(\pi) d\pi > 0$ for some $\varepsilon > 0$. Coercion after signal $s = 0$ is preferred to such communication whenever

$$\max_{m \in \{0,1\}} \varphi \int_0^1 \left[p_A(\hat{a}_{i,m} - 1)^2 + (1-p_A)(\hat{a}_{i,m})^2 - p_A(p_A-1)^2 - (1-p_A)(p_A)^2\right] g(\pi_i) d\pi_i \geq q,$$

where $p_A = p_A(0)$. Now we show that the integral in the LHS has a lower bound $\delta > 0$ that is independent of particular actions $\hat{a}_{i,0}$ and $\hat{a}_{i,1}$. Indeed,

$$\begin{aligned}
& \max_{m\in\{0,1\}} \int_0^1 \left[p_A(\hat{a}_{i,m}-1)^2+(1-p_A)(\hat{a}_{i,m})^2-p_A(p_A-1)^2-(1-p_A)(p_A)^2\right] g(\pi_i)d\pi_i \\
\geq \; & \int_{\pi_i>0.5+\varepsilon} \left[p_A(\hat{a}_{i,0}-1)^2+(1-p_A)(\hat{a}_{i,0})^2-p_A(p_A-1)^2-(1-p_A)(p_A)^2\right] g(\pi_i)d\pi_i \\
\geq \; & \int_{\pi_i>0.5+\varepsilon} \left[p_A(p_{0.5+\varepsilon}(0)-1)^2+(1-p_A)(p_{0.5+\varepsilon}(0))^2-p_A(p_A-1)^2-(1-p_A)(p_A)^2\right] g(\pi_i)d\pi_i \\
= \; & \delta > 0,
\end{aligned}$$

where $p_{0.5+\varepsilon}(0)$ is the posterior of the individual with the prior $0.5+\varepsilon$ after getting the signal $s=0$.[35] Here the first inequality holds, because the expression inside the integral is always positive and for every i the action $a_{i,0}$ is closer to the advisor's preferred action p_A than $a_{i,1}$ is. The second inequality holds, because $p_A(0) < p_{0.5+\varepsilon}(0) < a_{i,0}$ for any communication equilibrium of this type. As a result, the authority prefers coercion over communication when her altruism level exceeds some threshold (which is independent of particular actions $\hat{a}_{i,0}$ and $\hat{a}_{i,1}$).

Combining the results of the two cases, we obtain that for a sufficiently high level of altruism, $\varphi > \varphi_C(q,g)$ (where the subscript C denotes coercion), the authority cannot communicate with probability one in equilibrium.

Second, we argue that for a sufficiently large φ no communication is sustainable in equilibrium. In order to do this, we consider two manners in which information transmission can arise: (i) the authority communicates with positive probability after both signal realizations, and (ii) the authority communicates with non-zero probability after one signal realization and coerces with certainty after the other.

In case (i), we consider two possibilities: either each individual i takes only one action $\hat{a}_i$, or each individual i takes two different actions $\hat{a}_{i,0} < \hat{a}_{i,1}$ with positive probabilities. In the first case the authority prefers to coerce after the signal $s=0$ if

$$\varphi \int_0^1 \left[p_A(\hat{a}_i-1)^2+(1-p_A)(\hat{a}_i)^2-p_A(p_A-1)^2-(1-p_A)(p_A)^2\right] g(\pi_i)d\pi_i \geq q,$$

where $p_A = p_A(0)$, and we continue to assume that $\int_{\pi>0.5+\varepsilon} g(\pi)d\pi > 0$ for some $\varepsilon > 0$. As it was shown, the integral in the LHS exceeds some $\delta > 0$ that is independent of particular actions

[35] Note that the individual with the prior $0.5+\varepsilon$ is hypothetical, i.e., he might not be a part of the population.

$\hat{a}_i$. Hence, for φ exceeding some threshold (which is independent of particular actions $\hat{a}_i$), there exists no equilibrium with non-zero communication for both signals where each individual i takes only one action. Second, every individual i takes different actions $\hat{a}_{i,0} < \hat{a}_{i,1}$, then, by the similar analysis, the authority will prefer to coerce for sure when φ is greater than some threshold (which is independent of particular actions $\hat{a}_{i,0}$ and $\hat{a}_{i,1}$).

In case (ii), suppose, for example, that the authority coerces for sure after getting the signal $s = 0$. In this equilibrium, the authority's decision to communicate perfectly reveals that she observed the signal $s = 1$, so the individuals implement $a_i(1) = p_i(1)$. Clearly, for a high enough level of altruism

$$\varphi \int_{\pi_i \neq 0.5} \left[p_A(p_i - 1)^2 + (1 - p_A)(p_i)^2 - p_A(p_A - 1)^2 - (1 - p_A)(p_A)^2 \right] g(\pi_i) d\pi_i > q.$$

where $p_A = p_A(1)$ and $p_i = p_i(1)$, and the authority would choose to improve upon this action; consequently, such an equilibrium breaks down.

Combining the results of (i) and (ii), for sufficiently high $\varphi > \varphi_{CC}(q, g)$ the unique equilibrium is the one where the authority always coerces the individual. Note that $\varphi_{CC}(q, g) \geq \varphi_C(q, g) \geq \varphi_{TT}(q, g) > 0$ when $q > 0$ and $\varphi_{CC}(0, g) = \varphi_C(q, g) = 0$. □

Lemma 12 and Lemma 13 imply the opinion disagreement result of Proposition 9. Evidently, $\varphi_{CC}(q, g)$, $\varphi_C(q, g)$, and $\varphi_{TT}(q, g)$ are increasing in q. **QED.**

Proof of Proposition 10. The case of preference disagreement is considered in the main text; here we present a proof for opinion disagreement case. Consider some distribution $g(\pi) \in \mathcal{G}$ and the cost of lying $c > 0$. By Proposition 8, the advisor can be credible (to entire population) under public communication iff $\varphi \leq \overline{\varphi}(c, g)$. At the same time, under targeted communication, the advisor can report truthfully to a part of population with sufficiently close opinions. Thus, the advisor is (weakly) more credible with public messages when $\varphi \leq \overline{\varphi}(c, g)$, and is (weakly) more credible with private messages when $\varphi > \overline{\varphi}(c, g)$.

Assume that $\overline{\varphi}(c, g) > 0$ and that every individual believes the message he gets (either public or private). If $\varphi = \overline{\varphi}(c, g)$, then the advisor can be credible (to entire population) in public communication, but can not be credible to all individuals with private messages. Indeed, under public communication, the advisor is indifferent between reporting some signal $s \in \{0,1\}$ truthfully or lying (the advisor weakly prefers to report the other signal truthfully). That is, the benefit from lying about s to extreme individuals of measure η equals the loss from lying to other $1-\eta$ less extreme individuals plus the cost c. This implies that there is a set of individuals of measure η', $0 < \eta' \leq \eta$ such that the individual benefit of lying to these individuals exceeds c. Hence, under targeted communication, the advisor will strictly prefer to misreport the signal s to η' extreme individuals.

Now consider some distribution $g(\pi) \notin \mathcal{G}$. Then for any the cost of lying $c \geq 0$ and altruism level $\varphi \geq 0$ public communication is credible. At the same time, the advisor can be non-credible to some individuals with private messages. Thus, for all $\varphi \geq 0$ the advisor is (weakly) more credible under public communication.

Clearly, all individuals prefer a FRE to any other (pure or mixed strategies) PBE. Now we show that in private communication with individual i, a FRE is also preferred by the advisor to any other pure strategies PBE for any $\pi_i \in (0,1)$. Hence, a FRE is preferred under public communication as well; and more credible communication is always beneficial to the advisor with the prior $\pi_A = 0.5$ compared to less credible communication.

Assume that $\pi_i > 0.5$ (case of $\pi_i < 0.5$ is similar and case of $\pi_i = 0.5$ is straightforward). In a FRE, individual i takes actions $p_0 = p_i(0)$ and $p_1 = p_i(1)$ after messages 0 and 1, respectively. In the uninformative equilibrium the advisor sends the same message independently on the signal, and i takes action $\pi = \pi_i$. The FRE generates a greater expected payoff utility to the advisor whenever

$$\begin{aligned} & \frac{1}{2}\left[-(1-\gamma)(p_0-1)^2 - \gamma p_0^2 + (1-\gamma)(\pi-1)^2 + \gamma\pi^2\right] \\ > \; & \frac{1}{2}\left[\gamma(p_1-1)^2 + (1-\gamma)p_1^2 - \gamma(\pi-1)^2 - (1-\gamma)\pi^2\right]. \end{aligned}$$

This inequality can be rewritten as

$$(\pi - p_0)(\pi + p_0 - 2(1-\gamma)) > (p_1 - \pi)(\pi + p_1 - 2\gamma).$$

First, note that $\pi - p_0 > p_1 - \pi > 0$, i.e., the individual with the prior $\pi > 0.5$ is more responsive to a low signal $s = 0$ than to $s = 1$. Second, it is easily verified that $p_1 - p_0 < 2\gamma - 1$ (the individual with the prior $\pi > 0.5$ is on average less responsive than the individual with the prior 0.5). Hence,

$\pi + p_0 - 2(1-\gamma) > \pi + p_1 - 2\gamma$, which by $\pi + p_0 - 2(1-\gamma) > 0$, implies that the FRE in private communication with i is preferred to the uninformative equilibrium. **QED.**

3.10 Appendix: Dominance solvable setting

3.10.1 The altruistic advisor

Here we consider the following modification of the baseline model of Section 3.3. In the first stage of the game, a signal about the state, s, with precision $\Pr(s = \theta|\theta) \equiv \gamma \in (0.5, 1)$ is realized. This signal is privately observed by the advisor with probability $\alpha \in (0, 1)$; the advisor remains uninformed with a complementary probability $(1 - \alpha)$. Whether the advisor gets to know the signal does not depend on a particular signal realization, and is not observable by the population. Second, if the advisor did not obtain the signal, no information transmission happens. If the advisor is informed, she can either pass the signal s on to all individuals, or hide it at a cost $c \geq 0$, in which case the individuals learn nothing about the state – denoted as $\emptyset$. That is, the advisor can incur a costly effort to hide information from the population. Importantly, when the individual observes $\emptyset$, he can not tell apart whether the advisor did not receive a signal or just hid it. Third, upon learning the sinal s or observing no information $\emptyset$, the individuals choose an action $a \in \mathbb{R}$. The assumptions about individuals' preferences, opinions and payoffs are the same as in the baseline model, and the advisor's utility is given by

$$U_A(a, \theta) = -\int_0^1 (a_i - \theta)^2 di - c\mathbf{I}_{A \text{ hides info}} - \varphi \int_0^1 (a_i - \theta - b_i)^2 di.$$

Consider the action that individual i optimally chooses given the available information. In case individual i gets to know the signal, his action is $a_i(1) = p_i(1) + b_i$ if $s = 1$ and $a_i(0) = p_i(0) + b_i$ if $s = 0$. When no information is revealed to i, he rationally chooses $a_i(\emptyset) = p_i(\emptyset) + b_i$, where $p_i(\emptyset) \in [\overline{p}_{i,0}, \overline{p}_{i,1}] \subset (p_i(0), p_i(1))$, where the upper bound $\overline{p}_{i,1}$ corresponds to i's posterior given that the advisor passes on the signal $s = 0$ and incurs effort to hide $s = 1$. Similarly, the lower bound $\overline{p}_{i,0}$ is i's posterior when that the advisor reveals $s = 1$ and incurs effort to hide $s = 0$. The expressions for $\overline{p}_{i,1}$ and $\overline{p}_{i,0}$ are given by:

$$\begin{aligned}\overline{p}_{i,1} &= \frac{\pi_i(1-\alpha+\alpha\gamma)}{\pi_i(1-\alpha+\alpha\gamma)+(1-\pi_i)(1-\alpha+\alpha(1-\gamma))},\\ \overline{p}_{i,0} &= \frac{\pi_i(1-\alpha+\alpha(1-\gamma))}{\pi_i(1-\alpha+\alpha(1-\gamma))+(1-\pi_i)(1-\alpha+\alpha\gamma)}.\end{aligned}$$

Preference disagreement. Under preference disagreement, if the level of altruism is sufficiently large, the advisor necessarily passes on the signal to the population whenever she has one.

Lemma 14. *For any cost $c \geq 0$ preference distribution $f(b)$ with the mean $\bar{b}$ there exists $\overline{\varphi}(c,\bar{b})$, s.t. for every $\varphi > \overline{\varphi}(c,\bar{b})$ there is a unique rationalizable outcome, in which the advisor reveals the signal if she has one.*

Proof. Assume that the population's belief after observing no signal $\emptyset$ is given by some $p(\emptyset) \in [\overline{p}_0, \overline{p}_1]$. If the informed advisor does not hide the signal $s = 1$, she gets an expected payoff of:[36]

$$\begin{aligned}\mathbb{E}\left[U_A(a(1),\theta)|s=1\right] &= -p(1)\left[(p(1)+\bar{b}-1)^2+\varphi(p(1)-1)^2\right]\\ &\quad -(1-p(1))\left[(p(1)+\bar{b})^2+\varphi p(1)^2\right]+\bar{b}^2-\overline{b^2}.\end{aligned}$$

If the advisor exerts an effort to hide $s = 1$, her expected payoff becomes

$$\begin{aligned}\mathbb{E}\left[U_A(a(\emptyset),\theta)|s=1\right] &= -p(1)\left[(p(\emptyset)+\bar{b}-1)^2+\varphi(p(\emptyset)-1)^2\right]\\ &\quad -(1-p(1))\left[(p(\emptyset)+\bar{b})^2+\varphi p(\emptyset)^2\right]+\bar{b}^2-\overline{b^2}-c.\end{aligned}$$

The advisor prefers to pass the signal $s = 1$ on to the population whenever

$$\begin{aligned}& -p(1)\left[(p(1)+\bar{b}-1)^2+\varphi(p(1)-1)^2\right]-(1-p(1))\left[(p(1)+\bar{b})^2+\varphi p(1)^2\right]\\ > \;& -p(1)\left[(p(\emptyset)+\bar{b}-1)^2+\varphi(p(\emptyset)-1)^2\right]-(1-p(1))\left[(p(\emptyset)+\bar{b})^2+\varphi p(\emptyset)^2\right]-c.\end{aligned}$$

This condition is equivalent to the one, where the advisor communicates to just one individual with

[36] See the proof of Proposition 8 in Appendix 3.9 for a more detailed derivation of this and the following expressions.

the preference difference of $\bar{b}$, and because $p(\emptyset) < p(1)$, it can be rewritten as

$$\varphi > -1 + \frac{2\bar{b}}{p(1) - p(\emptyset)} - \frac{c}{(p(1) - p(\emptyset))^2}.$$

Similarly, using $p(\emptyset) > p(0)$, the advisor's incentive to pass on the signal $s = 0$ to the population can be expressed as

$$\varphi > -1 - \frac{2\bar{b}}{p(\emptyset) - p(0)} - \frac{c}{(p(0) - p(\emptyset))^2}.$$

Because $p(\emptyset) \in [\overline{p}_0, \overline{p}_1] \subset (p(0), p(1))$, for sufficiently large $\varphi > \overline{\varphi}(c, \bar{b})$ it is a dominant strategy for the informed advisor to reveal the signal independently of $p(\emptyset)$. □

Opinion disagreement. Under opinion disagreement, there exists a non-degenerate set of distributions, for which the the advisor necessarily hides some signal when the level of altruism is sufficiently large.

Lemma 15. *There exists a non-degenerate set of opinion distributions $\widehat{\mathcal{G}}$ such that hiding some signal is a strictly dominant strategy for the advisor when $c = 0$. For any $g(\pi) \in \widehat{\mathcal{G}}$ and $c > 0$, there exists a threshold level of altruism $\overline{\varphi}(c, g)$ such that the unique rationalizable outcome involves the advisor hiding some signal for sure when $\varphi \leq \overline{\varphi}(c, g)$.*

Proof. First, we show that $\widehat{\mathcal{G}}$ is non-empty and is, in fact, uncountable. Consider a simple opinion distribution with two types of individuals, t_1 and t_2, with priors $\pi_1 < \pi_A = 0.5 < \pi_2$. The types are equally prevalent, i.e., $g(\pi_1) = g(\pi_2) = 0.5$. Assume that type t_1 is extreme and is not responsive to the signal s, $\pi_1 = 0$. Assume that π_2 is sufficiently high to ensure that any $p(\emptyset) \in [\overline{p}_{2,0}, \overline{p}_{2,1}]$ is closer to $p_A(1) = \gamma$ than $p_2(1)$. This implies that in the case of $c = 0$ a strictly dominant strategy for the advisor is to always hide the signal $s = 1$. Clearly, $\widehat{\mathcal{G}}$ contains all such (and, by continuity, many more other) distributions, hence it is uncountable.

Second, we note that for any $g(\pi) \in \widehat{\mathcal{G}}$ and $c > 0$, the advisor's strictly dominant strategy is to hide some signal when φ is sufficiently large. Indeed, for any given cost of lying, increasing the level of altruism makes the net benefit from hiding the signal larger relative to the cost c.

□

3.10.2 The altruistic authority

After observing the signal, s, or observing no signal, $\emptyset$, the authority (A, she) chooses between sending a public message m, after which the individuals choose actions, and mandating an action for all individuals, $a_A \in \mathbb{R}$. As before, the cost of coercion is q, where $q \geq c = 0$.

Preference disagreement. Under preference disagreement, for sufficiently high levels of altruism, the authority never coerces or hides information.

Lemma 16. *For any for any preference distribution $f(b)$ with $\bar{b}^2 \neq \overline{b^2}$ and any $q \geq 0$, there exists a threshold $\overline{\varphi}(q, \bar{b}^2, \overline{b^2})$, such that for any $\varphi > \overline{\varphi}(q, \bar{b}^2, \overline{b^2})$ there is a unique rationalizable outcome where the authority never coerces the individuals and reveals the signal if she has one.*

Proof. Assume that the authority obtained the signal. By Lemma 14, for sufficiently large $\varphi > \overline{\varphi}(c = 0, \bar{b})$ the authority always prefers revealing the signal to hiding it independently of $p(\emptyset) \in [\overline{p}_0, \overline{p}_1]$. Next, comparing signal revelation with coercing and imposing $a_A(s) = a_A(s) = p(s) + \frac{\varphi}{1+\varphi}\bar{b}$, signal revelation is preferred when:[37]

$$-\varphi^2(\overline{b^2} - \bar{b}^2) - \varphi q + \overline{b^2} - q \leq 0. \tag{3.6}$$

Thus, for sufficiently large $\varphi > \overline{\varphi}(q, \bar{b}^2, \overline{b^2}) \geq \varphi > \overline{\varphi}(c = 0, \bar{b})$ it is a dominant strategy of the informed authority to pass on the signal and never coerce the individual.

Now consider an uninformed authority and assume $\varphi > \overline{\varphi}(q, \bar{b}^2, \overline{b^2})$. If the authority does nothing, each individual i rationally chooses action $\pi + b_i$ (the individuals know that the authority never hides information, because $\varphi > \overline{\varphi}_1(q, \bar{b}^2, \overline{b^2})$). If the authority coerces, she chooses $a_A(\emptyset) = \pi + \frac{\varphi}{1+\varphi}\bar{b}$. Thus, the authority and the individuals hold the same beliefs about the state, in which case the authority prefers not to coerce whenever (3.6) is satisfied. As a result, provided the rational behavior of the population, whenever $\varphi > \overline{\varphi}(q, \bar{b}^2, \overline{b^2})$ it is a strictly dominant strategy for the authority not to coerce the individuals and pass the signal on to them when she is informed. □

Opinion disagreement. Under opinion disagreement, for sufficiently high levels of altruism, the authority necessarily coerces the population.

[37] See the proof of Proposition 9 in Appendix 3.9 for a more detailed derivation of this and the following expressions.

Lemma 17. *For any for any opinion distribution $g(\pi)$ and any $q \geq 0$, there exists a threshold $\overline{\varphi}(q,g)$, such that for any $\varphi > \overline{\varphi}(q,g)$ there is a unique rationalizable outcome where the authority coerces the individuals with probability one.*

Proof. W.l.o.g. we assume that $\int_{\pi>0.5+\varepsilon} g(\pi)d\pi > 0$. Consider the authority who obtained the signal $s = 0$. If the authority reveals it to the population, then every individual i optimally chooses an action $p_i(0) \neq p_A(0)$ if $\pi_i \neq 0.5$. Thus, for sufficiently large $\varphi > \overline{\varphi}_0(q,g)$ the benefit from coercing gets larger than the cost and the authority finds it strictly optimal to impose her preferred action $p_A(s)$ on the population. If the authority hides the signal $s = 0$ and does not coerce, then each individual i picks an action $p_i(\emptyset) \in [p_i(0), p_i(1)]$. The benefit from coercing after signal $s = 0$ is

$$\begin{aligned}
&\varphi \int_0^1 \left[p_A(p_i(\emptyset)-1)^2 + (1-p_A)(p_i(\emptyset))^2 - p_A(p_A-1)^2 - (1-p_A)(p_A)^2\right] g(\pi_i)d\pi_i \\
\geq\; &\varphi \int_{\pi_i>0.5+\varepsilon} \left[p_A(p_i(\emptyset)-1)^2 + (1-p_A)(p_i(\emptyset))^2 - p_A(p_A-1)^2 - (1-p_A)(p_A)^2\right] g(\pi_i)d\pi_i \\
\geq\; &\delta_0 > 0,
\end{aligned}$$

where $p_A = p_A(0)$. Here the first inequality holds because the expression inside the integral is always positive, and the second inequality holds because for every i the action $p_i(\emptyset) \geq p_{0.5+\varepsilon}(\emptyset) > p_A(0)$ when $\pi_i > 0.5 + \varepsilon$. Thus, for $\varphi > \overline{\varphi}_1(q,g) \geq \overline{\varphi}_0(q,g)$ the authority strictly prefers to coerce the population after the signal $s = 0$ independently of $p_i(\emptyset)$. This, in turn, implies that $p_i(\emptyset) \in [\pi_i, p_i(1)]$.

Assume that $\varphi > \overline{\varphi}_1(q,g)$. If an uninformed authority coerces the population, she derives an expected benefit of

$$\begin{aligned}
&\varphi \int_0^1 \left[\pi_A(p_i(\emptyset)-1)^2 + (1-\pi_A)(p_i(\emptyset))^2 - \pi_A(\pi_A-1)^2 - (1-\pi_A)(\pi_A)^2\right] g(\pi_i)d\pi_i \\
\geq\; &\varphi \int_{\pi_i>0.5} \left[\pi_A(p_i(\emptyset)-1)^2 + (1-\pi_A)(p_i(\emptyset))^2 - \pi_A(\pi_A-1)^2 - (1-\pi_A)(\pi_A)^2\right] g(\pi_i)d\pi_i \\
\geq\; &\delta_1 > 0.
\end{aligned}$$

Here the first inequality holds because the expression inside the integral is always positive, and the second inequality holds because for every i the action $p_i(\emptyset) \geq \pi_i > \pi_A$ when $\pi_i > 0.5$. Thus, the authority strictly prefers to coerce after observing $\emptyset$ if $\varphi > \overline{\varphi}_2(q,g) \geq \overline{\varphi}_1(q,g)$.

Assume that $\varphi > \overline{\varphi}_2(q,g)$ and that the authority observed the signal $s=1$. By the same logic as before, she would strictly prefer coercing to revealing the signal when $\varphi > \max\{\overline{\varphi}_2(q,g), \overline{\varphi}_3(q,g)\}$. If she hides the signal and does not coerce, then each individual i optimally chooses $p_i(\emptyset) = p_i(1)$ (because if the individuals get to choose the action, it necessarily means that $s=1$, provided that $\varphi > \overline{\varphi}_2(q,g)$). This is strictly dominated by coercion for $\varphi > \max\{\overline{\varphi}_2(q,g), \overline{\varphi}_3(q,g)\}$.

As a result, when $\varphi > \overline{\varphi}(q,g)$, the unique rationalizable outcome is the one in which the authority always coerces the individuals.

□

3.11 Appendix: Setting with one individual

3.11.1 Model

We first develop a framework where an *individual* obtains information from an altruistic *advisor* before making a decision. We then consider an alternative framework where the advisor in fact is an *authority*, who can communicate with the individual before he makes his decision, but who also can choose to simply coerce the individual to take a certain action. In both of these frameworks, we analyze the impact of altruism under two different assumptions on the nature of disagreement between the two parties: conflicting preferences and conflicting opinions.

3.11.1.1 The altruistic advisor

There are two players, an individual (I, he) and an advisor (A, she). The individual must take an action, $a \in \mathbb{R}$. His payoff from the action depends on an unknown state of the world, $\theta \in \{0,1\}$. While he is unable to obtain information about the state by himself, he can rely on the privately informed advisor for such information. The players' prior beliefs about the state of the world are characterized by $\Pr_i(\theta = 1) = \pi_i \in (0,1)$, for $i \in \{I, A\}$.[38] Following [Che and Kartik, 2009], we say that the players have different opinions when $\pi_I \neq \pi_A$.

In the first stage of the game, the advisor privately observes a signal about the state, s, with precision $\Pr(s = \theta|\theta) \equiv \gamma \in (0.5, 1)$. Second, she sends a message $m \in \{0,1\}$ to the individual. If

[38] Given that the state is binary, all beliefs can be characterized by $\Pr(\theta = 1)$.

the advisor lies, i.e., sends message $m \neq s$, she incurs a cost $c \geq 0$.[39] Third, upon receiving the advisor's signal, the individual chooses an action $a \in \mathbb{R}$.

The players' material payoffs from this action are given by $u_I(a, \theta) = -(a - \theta - b(\theta))^2$ and $u_A(a, \theta) = -(a - \theta)^2 - c\mathbf{I}_{\{m \neq s\}}$, where the bias $b(\theta) \geq 0$ captures their preference disagreement, and $\mathbf{I}_{\{m \neq s\}}$ is an indicator variable taking the value one if and only if the advisor lies, and zero otherwise. In this standard model of communication, we allow the advisor to be altruistic, i.e., the players' utilities are given by

$$\begin{aligned} U_I(a, \theta) = & \quad u_I(a, \theta) & = -(a - \theta - b(\theta))^2, \\ U_A(a, \theta) = & \quad u_A(a, \theta) + \varphi u_I(a, \theta) & = -(a - \theta)^2 - c\mathbf{I}_{\{m \neq s\}} - \varphi(a - \theta - b(\theta))^2, \end{aligned}$$

where $\varphi \geq 0$ captures the degree to which the advisor cares about the (material) well-being of the individual.[40]

The prior beliefs π_I and π_A, the signal precision γ, the lying cost c, the preference alignment $b(\theta)$, and the degree of altruism φ are common knowledge.

Strategies and equilibrium. A pure strategy of the advisor specifies, for each signal s, the message $m(s)$ that she sends, $m : \{0, 1\} \to \{0, 1\}$. The individual's posterior beliefs conditional on message m are described by $\Pr_I(\theta = 1|m)$, where superscript I signifies that the individual forms his beliefs using his prior π_I. A pure strategy of the individual specifies, for each message m, action $a_I(m)$ that he takes, $a_I : \{0, 1\} \to \mathbb{R}$. We use a solution concept of Perfect Bayesian Equilibria (PBE), where the advisor maximizes her expected utility for each signal s given the individual's strategy $a_I(m)$; the individual maximizes his expected utility given his beliefs $\Pr_I(\theta = 1|m)$ after each message m; and the beliefs $\Pr_I(\theta = 1|m)$ satisfy Bayes' rule whenever possible.

[39] This formulation encompasses cheap-talk ($c = 0$) and hard information ($c = \infty$). We maintain that $c \in (0, \infty)$ unless we explicitly consider one of these two cases.

[40] We show in Section 3.11.5 that all the main results remain valid if friendship is mutual. We choose unidirectional friendship as our baseline formulation not only for simplicity, but also because it better reflects a social planner who cares about a citizen.

3.11.1.2 The altruistic authority

We now replace the advisor with an *authority*. The authority can choose to behave like an advisor—that is, to provide the individual with information before he chooses $a \in \mathbb{R}$ himself—but the authority also has the option to instead impose an action on the individual. Formally, after observing the signal, s, the authority chooses between sending a message $m(s)$ to the individual (in which case the game proceeds as in the altruistic advisor framework), and engaging in coercion, whereby the authority simply picks *her* desired action $a \in \mathbb{R}$ for the individual.[41]

We assume that the authority must incur a cost $q > 0$ if she coerces the individual. Depending on the application, this cost may reflect the instrumental cost of active intervention, or the authority's intrinsic aversion against removing the individual's liberty to choose.

In this altruistic authority framework, we let $c = 0$. This reflects the fact that in many of the applications we consider, it is reasonable to assume that the lying cost is negligible relative to the cost of coercion.[42] This is merely a simplification; all insights remain valid for $c > 0$. The parties utilities are thus given by

$$\begin{aligned} U_I(a,\theta) = &\quad u_I &&= -(a-\theta-b(\theta))^2, \\ U_A(a,\theta) = &\quad u_A + \varphi u_I &&= -(a-\theta)^2 - q\mathbf{I}_{\{coercion\}} - \varphi(a-\theta-b(\theta))^2, \end{aligned}$$

where $\mathbf{I}_{\{coercion\}}$ takes value one if and only if coercion occurs.

Strategies and equilibrium. A pure strategy of the authority specifies, for each signal s, whether she chooses to coerce or not coerce the individual, $C(s) \in \{\text{Coerce, Not coerce}\}$, what action $a_A(s) \in \mathbb{R}$ she takes if she chooses to coerce, and what message $m(s) \in \{0,1\}$ she sends if she chooses not to coerce. The individual's posterior beliefs conditional on receiving message m are given by $\Pr_I(\theta = 1|m)$ and a pure strategy of the individual specifies an action $a_I(m) \in \mathbb{R}$ for each message m. As before, we use the solution concept of PBE, in which the authority maximizes

[41] Note that such coercion differs from delegation, whereby the individual decides that the authority should exercise the action choice. In contrast, under coercion, the authority herself decides when she should exercise the action choice on behalf of the individual. Moreover, because the authority observes the signal s before deciding whether to coerce, whereas the individual does not observe the signal s, the coercion choice is taken ex interim, whereas a delegation choice must be made ex ante. We discuss how these concepts are related in Section 3.11.4.

[42] This is discussed in depth in Section 3.11.3.

her expected utility for each signal given the individual's strategy; the individual maximizes his expected utility given his beliefs after each message; and the beliefs satisfy Bayes' rule whenever possible.

3.11.1.3 Nature of disagreement

We analyze this model under two different assumptions on the nature of disagreement between the individual and the advisor:

Conflicting preferences. Under conflicting preferences, the individual's preferences are biased relative to those of the advisor, i.e., $b(\theta) \neq \mathbf{0}$ for at least one value of θ; however, the parties have the same opinions, i.e., $\pi_I = \pi_A \equiv \pi$. In the main body of the study we assume that preference biases are independent on the state: $b(0) = b(1) = b$.[43]

Conflicting opinions. Under conflicting opinions, the parties have fully aligned preferences, i.e., $b(\theta) = \mathbf{0}$; however, they have different opinions, i.e., $\pi_I \neq \pi_A$.

3.11.2 Analysis: the altruistic advisor

In this section, we study communication between the individual and his altruistic advisor. The stages of this game are summarized in Figure 3.6 below. In particular, we analyze how the possibility to sustain truthful communication is influenced by the advisor's regard for the individual. Because this crucially depends on the nature of disagreement between the individual and his advisor, we consider each case in turn in order to isolate the mechanisms.

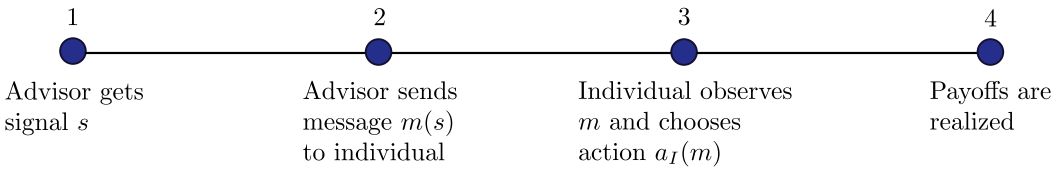

1
2
3
4
Advisor gets signal s
Advisor sends message $m(s)$ to individual
Individual observes m and chooses action $a_I(m)$
Payoffs are realized

Figure 3.6: Stages of the setting with altruistic advisor.

[43] This merely simplifies the exposition and makes the underlying intuition more transparent; as we show in the Appendix, all our results hold in the more general case of conflicting preferences.

3.11.2.1 Conflicting preferences

Because the parties share the same prior beliefs, $\pi_I = \pi_A \equiv \pi$, they hold the same posterior beliefs $p(s) = \Pr(\theta = 1|s)$ about the state of the world, conditional on the signal s,

$$\begin{aligned} p(1) &= \frac{\pi\gamma}{\pi\gamma + (1-\pi)(1-\gamma)}, \\ p(0) &= \frac{\pi(1-\gamma)}{\pi(1-\gamma) + (1-\pi)\gamma}. \end{aligned}$$

Under perfect information, when the state of the world θ is known, the individual and the advisor prefer different actions because $b \neq 0$. This disagreement carries on to the case of imperfect information when the individual and the advisor have the same beliefs about the state of the world.

We analyze the game backwards. In stage 3, the individual observes the message m, forms his posterior belief $\Pr(\theta = 1|m)$, and chooses the action that maximizes his expected payoff:

$$a_I(m) = \arg\max_{a\in\mathbb{R}} \mathbb{E}(U_I(a,\theta)|m).$$

Because the loss function is quadratic, the individual's maximization problem has a straightforward solution: $a_I(m) = \Pr(\theta = 1|m) + b$; that is, the individual optimally chooses the action that matches his expected state of the world plus the bias b. If the individual believes that the advisor reports truthfully at stage 2, i.e., that $m(s) = s$, then the individual optimally chooses $a_I(m) = p(m) + b$. In this case, a FRE exists whenever the advisor's incentive compatibility conditions for truthful reporting are satisfied.

Now consider stage 2. For the advisor to report a signal s truthfully, inducing the action $a_I(m = s)$ must generate a greater expected payoff than inducing the action $a_I(m \neq s)$. First, suppose that the advisor obtained the signal $s = 1$ in stage 1. In this case, her posterior belief that $\theta = 1$ is $p(1)$, and her posterior belief that $\theta = 0$ is $(1 - p(1))$. The advisor's expected utility from sending $m = 1$ and inducing action $a_I(1) = p(1) + b$ is given by

$$\begin{aligned} \mathbb{E}\left[U_A(a_I(1),\theta)|s=1\right] &= -p(1)\left[(a_I(1)-1)^2 + \varphi(a_I(1)-1-b)^2\right] \\ &\quad -(1-p(1))\left[a_I(1)^2 + \varphi(a_I(1)-b)^2\right]. \end{aligned}$$

If the advisor instead lies, i.e., sends the message $m = 0$, she incurs the cost c and induces the action $a_I(0) = p(0) + b < a_I(1)$. In this case, her expected utility is given by

$$\begin{aligned}\mathbb{E}\left(U_A(a_I(0), \theta)|s = 1\right) \quad = \quad & -p(1)\left[(a_I(0)-1)^2 + \varphi(a_I(0)-1-b)^2\right] \\ & -(1-p(1))\left[a_I(0)^2 + \varphi(a_I(0)-b)^2\right] - c.\end{aligned}$$

Hence, after getting the signal $s = 1$, the advisor prefers to send a truthful message over lying if and only if the following inequality (denoted (TT1$_{\text{pr}}$), where the subscript "pr" indicates the conflicting preferences setting) is satisfied:

$$\varphi \geq -1 + \frac{2b}{p(1)-p(0)} - \frac{c}{(p(1)-p(0))^2}. \tag{TT1$_{\text{pr}}$}$$

Second, suppose that the advisor gets the signal $s = 0$. An analogous analysis leads to the following incentive compatibility condition for truth-telling

$$\varphi \geq -1 - \frac{2b}{p(1)-p(0)} - \frac{c}{(p(1)-p(0))^2}. \tag{TT0$_{\text{pr}}$}$$

Combining (TT1$_{\text{pr}}$) and (TT0$_{\text{pr}}$) yields that altruism improves communication when the parties have different underlying preferences. In particular, truthful communication can be sustained when the level of altruism φ is sufficiently high. This result is formally stated in the following proposition:

Proposition 12. *For any bias b and lying cost c there exists a threshold $\overline{\varphi}(c, b)$ s.t. a fully revealing equilibrium exists if and only if $\varphi \geq \overline{\varphi}(c, b)$.*[44]

Clearly, if $b = 0$, the individual and the advisor prefer the same action, and truthful communication can be sustained. When the parties have different preferences, however, communication may break down. If so, strengthening the level of altruism eventually restores truthful communication, because altruism mitigates the preference conflict. To gain intuition for the result, consider the action preferred by the advisor, conditional on getting the signal s, $a_A(s)$:

$$a_A(s) = \frac{1}{1+\varphi}p(s) + \frac{\varphi}{1+\varphi}(p(s)+b).$$

[44] This result holds in the general case of state dependent preferences, i.e., when $b(0) \neq b(1)$.

This is a weighted sum of the preferred action of a non-altruistic advisor, $p(s)$, and the preferred action of the individual, $(p(s)+b)$. For her own sake, the advisor would like the action $p(s)$ to be implemented. Nevertheless, when she cares about the individual, her optimal action also reflects that he is better off with the action $(p(s)+b)$. The stronger the advisor's care for the individual, the more she internalizes the individual's preferences, and the closer is $a_A(s)$ to $a_I(s)$. Clearly, for a high enough level of altruism, truthful communication is attainable.

To better understand the exact form of the truth-telling conditions, consider the cheap-talk case with a zero cost of lying, $c = 0$. The symmetric quadratic loss function yields that the advisor wants to induce an action as close to her own preferred action, $a_A(s)$, as possible. If the individual's bias is positive, $b > 0$, then the advisor always reports the signal $s = 0$ truthfully, because $a_A(0) = a_I(0) + \frac{\varphi}{1+\varphi}b$ is closer to $a_I(0)$ than to $a_I(1) > a_I(0)$. If she obtains the signal $s = 1$, the advisor reports it truthfully if and only if action $a_A(1)$ is closer to $a_I(1)$ than to $a_I(0)$; that is, if $a_I(1) - a_A(1) \leq a_A(1) - a_I(0)$. This condition can be rewritten $\frac{2b}{1+\varphi} \leq a_I(1) - a_I(0)$, which is equivalent to (TT1_{pr}) for $c = 0$. Analogously, if the bias is negative, $b < 0$, the advisor always reports $s = 1$ truthfully, and reveals $s = 0$ if and only if the incentive constraint (TT0_{pr}) is satisfied.

Corollary 1. *The threshold $\overline{\varphi}(c, b)$ decreases with c and increases with $|b|$.*

The truth-telling threshold has intuitive properties. First, a higher cost of lying c makes truthful reporting more attractive for the advisor. When a FRE can be sustained for a larger range of altruism levels, the level of altruism necessary to induce truth-telling, $\overline{\varphi}(c, b)$, decreases. Second, a more severe bias $|b|$ intensifies the preference conflict, which increases the advisor's relative attractiveness of misreporting her signal. Consequently, the level of altruism required to mitigate the preference divergence, $\overline{\varphi}(c, b)$, increases.

3.11.2.2 Conflicting opinions

When the individual and the advisor have different opinions, $\pi_I \neq \pi_A$, they have different posterior beliefs $p_i(s) = \Pr_i(\theta = 1|s)$ given the same signal s:

$$p_i(1) = \frac{\pi_i\gamma}{\pi_i\gamma + (1-\pi_i)(1-\gamma)}, \; p_i(0) = \frac{\pi_i(1-\gamma)}{\pi_i(1-\gamma) + (1-\pi_i)\gamma}, \quad i \in \{I, A\}.$$

In terms of preferences over outcomes, the individual is fully aligned with the advisor, i.e., $b = 0$, which implies that they have the same material payoffs. Their resulting utilities are given by:

$$\begin{aligned} U_I(a,\theta) &= -(a-\theta)^2, \\ U_A(a,\theta) &= -(a-\theta)^2 - c\mathbf{I}_{\{m\neq s\}} - \varphi(a-\theta)^2. \end{aligned}$$

Under perfect information about the state of the world, there is no disagreement between the parties; they prefer the same action, $a = \theta$. In contrast, under imperfect information, the individual and the advisor prefer different actions even if they have the same signal s about the state of the world, because they interpret this signal in light of their respective (different) priors. In our framework, information is always imperfect, since the signal s is not fully informative ($\gamma < 1$).

We analyze the game backwards. In Stage 3, given the individual's belief about the state of the world, $\Pr_I(\theta = 1|m)$, he optimally chooses the action that matches this expected state of the world, $a_I(m) = \Pr_I(\theta = 1|m)$. If the individual believes that the advisor reports truthfully at stage 2, i.e., that $m(s) = s$, then the individual chooses action $a_I(m) = p_I(m)$. The necessary and sufficient conditions for a truth-telling equilibrium to exist thus coincide with the advisor's incentive compatibility conditions for truthful reporting.

Now consider stage 2. First, suppose that the advisor obtained the signal $s = 1$ in stage 1. In this case, her posterior belief that $\theta = 1$ is $p_A(1)$, and her posterior belief that $\theta = 0$ is $(1 - p_A(1))$. Given the individual's strategy, sending $m = 1$ induces action $a_I(1) = p_I(1)$, while sending $m = 0$ induces action $a_I(0) = p_I(0) < a_I(1)$. The advisor's expected payoff from sending $m = 1$ is given by

$$\begin{aligned} \mathbb{E}_A\left(U_A(a_I(1),\theta)|s=1\right) &= -p_A(1)\left((a_I(1)-1)^2 + \varphi(a_I(1)-1)^2\right) \\ &\quad -(1-p_A(1))\left(a_I(1)^2 + \varphi a_I(1)^2\right), \end{aligned}$$

where the subscript A on her expectation reflects that the advisor evaluates the expected utility using *her* posterior about the state of the world, whereas the individual chooses the action that is optimal given *his* posterior. The advisor's expected payoff from sending $m = 0$ is given by

$$\begin{aligned}\mathbb{E}_A\left(U_A(a_I(0),\theta)|s=1\right) &= -p_A(1)\left((a_I(0)-1)^2+\varphi(a_I(0)-1)^2\right)\\ &\quad -(1-p_A(1))\left((a_I(0))^2+\varphi a_I(0)^2\right)-c.\end{aligned}$$

The advisor's incentive compatibility condition for truthful reporting when $s = 1$ (denoted (TT1$_{\text{op}}$), where the subscript "op" indicates the conflicting opinions setting) is given by the inequality $\mathbb{E}_A\left(U_A(a_I(1),\theta)|s=1\right) \geq \mathbb{E}_A\left(U_A(a_I(0),\theta)|s=0\right)$, which can be rearranged as

$$\frac{a_I(1)+a_I(0)}{2} \leq p_A(1) + \frac{c}{\tau\left(1+\varphi\right)}, \tag{TT1$_{\text{op}}$}$$

where $\tau \equiv 2\left(a_I(1)-a_I(0)\right)$. An analogous calculation yields the truth-telling condition when the advisor gets the signal $s = 0$, and combining these two conditions yields that a fully revealing equilibrium can be sustained if and only if

$$p_A(0) - \frac{c}{\tau\left(1+\varphi\right)} \leq \frac{a_I(1)+a_I(0)}{2} \leq p_A(1) + \frac{c}{\tau\left(1+\varphi\right)}. \tag{TT$_{\text{op}}$}$$

As we shall see, this condition yields that greater level of altruism worsens the prospects to achieve truthful communication. In particular, if truthful communication can be sustained, raising the level of altruism may eventually destroy it; and if truthful communication cannot be sustained, raising the level of altruism cannot help. Before stating this result formally, we develop its intuition, starting from the special case when $c = 0$. Then, (TT$_{\text{op}}$) reduces to

$$p_A(0) \leq \frac{a_I(1)+a_I(0)}{2} \leq p_A(1). \tag{TT$'_{\text{op}}$}$$

Since $a_I(1) = p_I(1)$ and $a_I(0) = p_I(0)$, this condition identifies the set of (π_I, π_A) for which a FRE can be sustained, given γ. Figure 3.7 below displays this region when $\gamma = 0.7$.

The top solid line represents the advisor's truth-telling condition when getting the signal $s = 1$: she communicates $s = 1$ truthfully for all (π_A, π_I) below this line. Similarly, she communicates $s = 0$ truthfully for all (π_A, π_I) above the solid lower line. Hence, a FRE can be sustained in their intersection, which we denote by TT'.

The location of TT' illustrates that truthful communication requires the priors of the individual and his advisor to be sufficiently similar. Along the 45°-line, their priors are identical; hence, the action that the individual chooses if she gets signal s, $a_I(s) = p_I(s)$, coincides with the action that the advisor desires him to take, $a_A(s) = p_A(s)$. Everywhere else, their priors differ, so $a_I(s) \neq a_A(s)$. Nevertheless, the advisor prefers to tell the truth so long as $a_A(s)$ is closer to the action that a truthful report induces than it is to the action induced by a false report. To see this formally, we can re-write the right inequality in ($\mathrm{TT}'_{\mathrm{op}}$) as $a_I(1) - a_A(1) \leq a_A(1) - a_I(0)$, using the fact that $p_A(1) \equiv a_A(1)$. The advisor reports the signal $s = 1$ truthfully so long as the action that she wants the individual to take, $a_A(1)$, is closer to the action that the individual chooses under truthful reporting, $a_I(1)$, than to the action that he chooses if the advisor lies, $a_I(0)$. This obtains because the advisor's expected loss function is monotonic in the distance between the action that she prefers, $a_A(1)$, and the action that the individual takes, $a_I(m)$.

When the priors are so different that the advisor prefers the action that he induces by lying to the action that he induces by telling the truth in some state of the world, the FRE breaks down. For all (π_A, π_I) below the lower solid line, the advisor is considerably more convinced than the individual that the state of the world is $\theta = 1$ (from ex-ante perspective). In this case, the advisor would communicate truthfully when getting the signal $s = 1$; however, when getting the signal $s = 0$, she would prefer to lie and send the message $m = 1$. Intuitively, the advisor prefers lying over truth-telling—even though there is no preference conflict—because she believes that the individual is so misinformed about the true prior that the individual will take a worse action if the advisor sends a true message than if she lies. A similar logic applies to the region above the top solid line, where the FRE breaks down because the advisor would report the signal $s = 0$ truthfully, but lie when $s = 1$.

The shape of TT' does not depend on the advisor's care for the individual, φ. Intuitively, this is because altruism does not influence whether the advisor prefers the action that is induced by telling the truth, $a_I(m = s)$, or the action that is induced by lying, $a_I(m \neq s)$. Rather, the level of altruism influences *how much* the advisor suffers from the implementation of an action that deviates from her own preferred action, $a_A(s)$. Indeed, when the advisor cares about the individual, on top of *her own* material expected loss the advisor also internalizes a share of the disutility that she expects the individual to suffer from his erroneous choice of action.

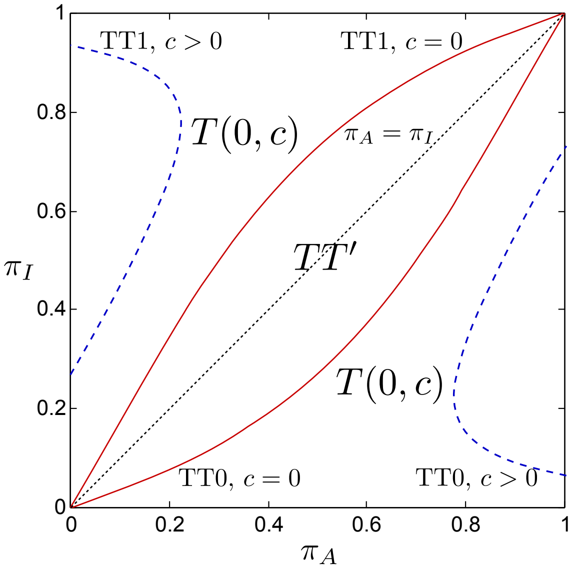


Figure 3.7: Truth-telling incentives for different priors (π_A, π_I).

We now consider the case when lying entails a cost, $c > 0$. In this case, truth-telling can be sustained for a larger set of (π_A, π_I). Formally, this is immediate from the truth-telling condition, (TT_{op}): for all (π_A, π_I) such that the advisor is indifferent between lying and telling the truth when $c = 0$—i.e., along the boundaries of TT'—she strictly prefers to tell the truth when $c > 0$. Intuitively, when lying entails no cost, the advisor would like to lie whenever she believes that lying induces the individual to take a better action than does telling the truth. However, when the advisor is averse to lying, she weighs the cost of incurring a lie against the cost of inducing the individual to take (what she believes to be) a suboptimal action. Clearly, in the presence of a lying cost, she will be more prone to reveal the true signal.

The above discussion yields that there exists a nonempty set of (π_A, π_I) such that a FRE can be sustained in the presence of lying cost c, but not when $c = 0$. We refer to this region as $T(\varphi, c)$. The dotted lines in Figure 3.7 plot the region $T(\varphi, c)$ for $\varphi = 0$ and $c = 0.2$. The shape of $T(\varphi, c)$ illustrates that a FRE always exists when the individual's prior is close to zero or one. This is because the benefit of lying is increasing in the distance between $a_I(s = m)$ and $a_I(s \neq m)$. This distance vanishes as the individual's prior approaches either 0 or 1; intuitively, when the individual has a strong belief about the state of the world ex ante, the new signal will have a negligible effect on her desired action. Hence, if lying entails a cost, the advisor strictly prefers to send a truthful message.

The dependence of $T(\varphi, c)$ on φ reflects that, when lying is costly, the advisor's altruism is crucial for the existence of a FRE. In particular, $T(\varphi, c)$ shrinks as φ increases; formally, when φ increases, ($\mathrm{TT}_{\mathrm{op}}$) converges to ($\mathrm{TT}'_{\mathrm{op}}$). This implies that, in all circumstances when the advisor's lying aversion can promote truthful communication, increasing the degree of altruism eventually destroys truthful communication. When $\varphi = 0$, the advisor's sole benefit from lying is that the lie can induce the individual to take an action that the advisor deems better, which benefits the advisor. Because the advisor and the individual have identical preferences, the advisor also believes that if she lies, this benefits the individual. However, in the absence of altruism, the advisor does not take this into account. When this sole benefit of lying is not large enough to outweigh the lying cost c, the advisor communicates truthfully. In contrast, in the presence of altruism, the advisor benefits from lying for two reasons: first, because she anticipates that the lie will increase her own ex-post utility; and second, because she believes that the lie will benefit the individual. As the altruism strengthens, the advisor internalizes more of the individual's benefit from her lie, so the advisor's expected benefit from lying increases. As the cost of lying remains constant, the FRE eventually breaks down as the level of altruism increases.

We make these notions precise in the following proposition:

Proposition 13. *For any lying cost $c > 0$, when $\varphi = 0$, a FRE exists if and only if $(\pi_A, \pi_I) \in TT' \bigcup T(0, c)$. For all $(\pi_A, \pi_I) \in TT'$ a FRE exists independently of the level of φ. For all $(\pi_A, \pi_I) \in T(0, c)$, there exists a threshold level of altruism $\overline{\varphi}(\pi_A, \pi_I, c)$ such that a FRE exists if and only if $\varphi \leq \overline{\varphi}(\pi_A, \pi_I, c)$. When no FRE exists, for $(\pi_A, \pi_I) \notin TT' \bigcup T(0, c)$, raising the level of altruism cannot improve communication.*

Paradoxically, when the advisor cares too much about the individual, communication can break down. If anything, altruism destroys communication because, when the parties have different opinions, the advisor believes that the individual will misinterpret a truthful report. Consequently, the altruistic advisor is inclined to lie in order to protect the individual from his own misinterpretation. Clearly, as the individual anticipates that the altruistic advisor lies, no informative communication can take place. This implies that the individual may prefer to hear the opinion of a disinterested advisor, whom he can trust to tell him the truth, over hearing the opinion of an altruistic advisor.

Corollary 2. *The threshold $\overline{\varphi}(\pi_A, \pi_I, c)$ increases with c.*

A higher cost of lying c makes truthful reporting more attractive for the advisor. When a FRE can be sustained for a larger range of altruism levels, the upper boundary for the levels of altruism supporting truth-telling, $\overline{\varphi}(\pi_A, \pi_I, c)$, increases.

3.11.3 Analysis: the altruistic authority

In this section we study interaction between the individual and the altruistic authority. The stages of the game are summarized in Figure 3.8 below. We analyze when the authority coerces the individual, and how this choice depends on the advisor's regard for the individual. Because this crucially depends on the nature of disagreement between the individual and his advisor, we consider each case in turn in order to isolate the mechanisms.

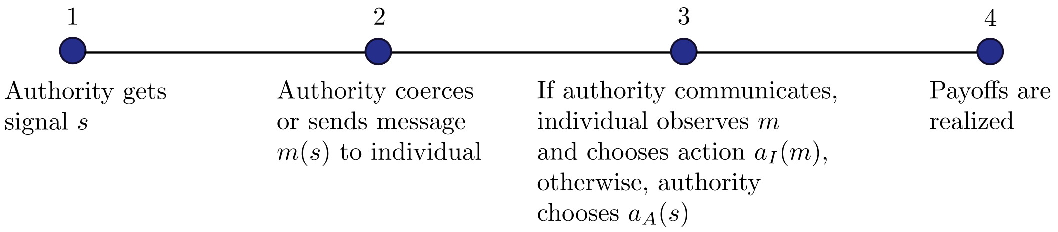

1
2
3
4
Authority gets signal s
Authority coerces or sends message $m(s)$ to individual
If authority communicates, individual observes m and chooses action $a_I(m)$, otherwise, authority chooses $a_A(s)$
Payoffs are realized

Figure 3.8: Stages of the setting with altruistic authority.

3.11.3.1 Conflicting preferences

The authority's expected payoff from coercion. Upon receiving the signal s, the authority forms her posterior about the state, $p(s) = \Pr(\theta = 1|s)$. Then, if the authority imposes an action on the individual through coercion, she optimally chooses the action $a_A(s) = \frac{1}{1+\varphi}p(s) + \frac{\varphi}{1+\varphi}(p(s) + b)$. Her expected utility from coercion is given by

$$\mathbb{E}(U_A(a_A(s), \theta)|s) = -q - (1+\varphi)p(s)(1-p(s)) - \frac{\varphi}{1+\varphi}b^2.$$

For a formal derivation of this expression, see the proof of Proposition 14.

The authority who truthfully advises. When does the altruistic authority act as a truth-telling advisor in equilibrium? First, revealing the true signal must be incentive compatible, given that the individual believes the message. From Proposition 12 in Section 3.11.2.1, we know that this occurs if and only if the level of altruism is sufficiently high; $\varphi \geq \overline{\varphi}(c=0,b)$. Second, when the advisor has the authority to force the individual's action, she must prefer communication over coercion. More precisely, for each signal s, the authority must prefer to communicate with the individual—and let him choose his preferred action conditional on the message m, $a_I(m) = p_I(m)+b$, where $p_I(m) = \Pr(\theta = 1|m)$—over incurring the cost q to impose her own preferred action, $a_A(s)$. This condition, $\mathbb{E}(U_A(a_I(m),\theta)|s) \geq \mathbb{E}(U_A(a_A(s),\theta)|s)$, can be expressed as

$$q(1+\varphi) \geq \left[(1+\varphi)(p_I(m)-p(s))+b\right]^2 . \qquad (\text{Communication}_{\text{pr}})$$

Under truth-telling, $m = s$ and $p_I(m) = p(s)$, implying that condition ($\text{Communication}_{\text{pr}}$) simplifies to $q(1+\varphi) \geq b^2$. Combining the two requirements yields that a FRE exists if and only if the authority cares sufficiently about the individual.

Proposition 14. *For any cost of coercion $q > 0$ and preference bias b, there exists a threshold $\varphi_{TT}(q,b)$, s.t. a FRE exists if and only if $\varphi \geq \varphi_{TT}(q,b)$.*

Proof. See Appendix. □

The more the advisor cares about the individual, the more she values that the individual gets to implement the action that he prefers. Hence, as φ increases, the authority's preferred action, $a_A(s)$, approaches the individual's preferred action, $a_I(s)$. This makes the benefit of coercion—the ability to implement $a_A(s)$ instead of the individual's choice, $a_I(s)$—decreasing in the level of altruism, φ. Consequently, for high levels of altruism, the benefit of coercion does not outweigh the (fixed) cost of coercion, q. Instead, the authority chooses to communicate truthfully with the individual, which ascertains that he makes as informed an action choice as possible. We refer to this behavior on the part of the individual as *libertarian*: after transferring all of her information to the individual, she gives the individual the liberty to choose the action that he wants.

The region of (q,φ) described in Proposition 14 is schematically illustrated in Figure 3.9. The threshold $\varphi_{TT}(q,b)$ has intuitive properties. First, it decreases in q, because a higher cost of coercion makes communication more attractive relative to coercion. Second, the threshold decreases with the bias b, because a greater level of altruism is required to mitigate the preference divergence, the higher is b.

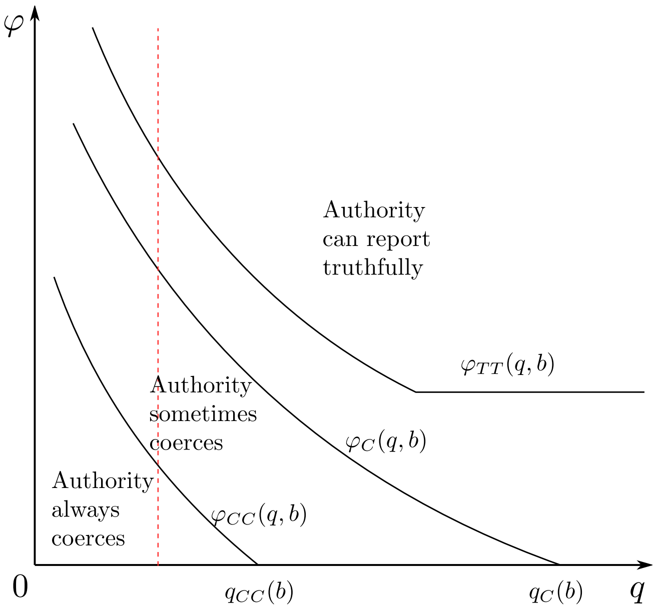


Figure 3.9: Equilibrium outcomes under conflicting preferences.

The authority who coerces. We now address the opposite question: under what circumstances does the authority exercise her power to choose which action to implement? We consider all (pure and mixed strategy) equilibria, and examine how the authority's coercion decision depends on her care for the individual. Because the individual, if he gets to implement an action, always takes the action which is equal to his (posterior) belief about the state of the world, he never plays a mixed strategy. There exist multiple equilibria in which the authority plays a mixed strategy, however: she may mix between coercion and communication, after one or both signals; or, conditional on communicating, she may mix between sending different messages. Given the individual's (pure) strategy—to take the action $a_I(m) = \Pr(\theta = 1|m) + b = p_I(m) + b$ upon receipt of the message m—the authority, after getting the signal s, prefers to coerce over sending the message m if and only if

$$q(1+\varphi) < \left[(1+\varphi)(p_I(m) - p(s)) + b\right]^2 . \qquad (\text{Coercion}_{\text{pr}})$$

Under truthful reporting ($m = s$ for $s \in \{0, 1\}$), we know from Proposition 14 that this condition is violated, for both $s = 1$ and $s = 0$, for large levels of altruism. Conversely, for low levels of altruism, when (Coercion$_{\text{pr}}$) is satisfied for both signals, the authority coerces regardless of the signal.

This intuition carries on beyond pure strategies, to the general case: as made precise in the below Lemma, coercion is more viable, the lower is the level of altruism. Hence, coercion is the most achievable when the authority does not care about the individual, $\varphi = 0$. We denote by $q_C(b)$ the cost of coercion below which the non-altruistic authority ($\varphi = 0$) coerces with strictly positive probability for at least one signal s, and by $q_{CC}(b)$ the cost below which coercion regardless of the signal is the unique equilibrium outcome.

Lemma 18. *For every b and $q < q_C(b)$ there exists a threshold $\varphi_C(q,b) > 0$, s.t. for any $\varphi < \varphi_C(q,b)$ every equilibrium outcome involves the authority imposing her preferred action with non-zero probability. Moreover, for every b and $q < q_{CC}(b)$ there exists a threshold $\varphi_{CC}(q,b) > 0$, s.t. for any $\varphi < \varphi_{CC}(q,b)$ there is a unique equilibrium, in which the authority coerces the individual for each signal $s \in \{0,1\}$.*

See Figure 3.9 for an approximate representation of these sets of (q,φ). Clearly, a greater cost q as well as lower preference bias b make coercion less attractive compared to communication, which implies that both thresholds, $\varphi_C(q,b)$ and $\varphi_{CC}(q,b)$, decrease with q and increase with b.

Together, Proposition 14 and Lemma 18 show that, under preference disagreement, a tyrannical authority (with $\varphi = 0$) may coerce the individual. A sufficiently benevolent authority, however, can communicate truthfully. In essence, the benevolent authority behaves in a *libertarian* fashion: she truthfully passes on her private information to the individual, thereby equipping him with the means to make as informed a choice as possible; then, she lets him choose his course of action.

3.11.3.2 Conflicting opinions

The authority's expected payoff from coercion. When the advisor gets the signal s, she forms her posterior belief $p_A(s) = P(\theta = 1|s)$. If the action a is implemented, her expected utility is given by

$$\mathbb{E}_A\left(U_A(a,\theta)|s\right) = -p_A(s)\left[(a-1)^2 + \varphi(a-1)^2\right] - (1-p_A(s))\left[a^2 + \varphi a^2\right] - q\mathbf{I}_{\{coercion\}}. \quad (\mathbb{E}U_A)$$

If the advisor executes her right to force the action choice, she implements her preferred action, $a = a_A(s) = p_A(s)$. Then, her expected utility simplifies to

$$\mathbb{E}_A\left(U_A(a_A(s), \theta)|s\right) = -p_A(s)(1 - p_A(s))(1 + \varphi) - q. \tag{$\mathbb{E}U_A$(Coercion)}$$

Intuitively, she incurs the cost of coercion, q; however, she benefits from picking exactly the action that she believes to be the best, $a_A(s)$, rather than any other action a implemented by the agent.

The authority who truthfully advises. When can the altruistic authority refrain from coercing, and instead communicate truthfully with the individual? From Section 3.11.2.2, we know that truthful communication can be sustained only if ($\mathrm{TT}'_{\mathrm{op}}$) is satisfied; that is, only if $(\pi_A, \pi_I) \in TT'$. When the advisor has the authority to force the individual's action, the existence of a FRE also requires that the authority prefers truthful communication over coercion. More precisely, for each signal s, the authority's expected utility from truthful communication must exceed her utility from coercion.

If the authority gets the signal s, her optimal action is $a_A(s) = p_A(s)$ and the individual's optimal action (in a FRE) is $a_I(s) = p_I(s) \neq a_A(s)$. The authority communicates the signal truthfully if and only if $\mathbb{E}_A\left(U_A(a_I(s), \theta)|s\right) \geq \mathbb{E}_A\left(U_A(a_A(s), \theta)|s\right)$. By ($\mathbb{E}U_A$) and ($\mathbb{E}U_A$(Coercion)), when $s = 1$, this simplifies to

$$\varphi \leq \frac{q}{(a_A(1) - a_I(1))^2} - 1, \tag{$\mathrm{TT1}_{\mathrm{op,C}}$}$$

(where the subscript "C" signifies coercion). An analogous calculation yields the incentive compatibility condition for truthful reporting of the signal $s = 0$, ($\mathrm{TT0}_{\mathrm{op,C}}$). Clearly, truthful communication can most easily be sustained in the absence of friendship, $\varphi = 0$, in which case ($\mathrm{TT1}_{\mathrm{op,C}}$) and ($\mathrm{TT0}_{\mathrm{op,C}}$) yield that a FRE exists if and only if the cost of coercion, q, exceeds the benefit of coercion for both signals, i.e., if $q \geq q_{TT} = \max\left\{(a_A(0) - a_I(0))^2, (a_A(1) - a_I(1))^2\right\}$. For all strictly higher costs of coercion q, a FRE can be sustained if and only if the authority does not care too much about the individual.

Lemma 19. *For any coercion cost $q \geq q_{TT}$ and for any $(\pi_A, \pi_I) \in TT'$, there exists a threshold $\varphi_{TT}(q, \pi_A, \pi_I)$ such that a FRE exists if and only if $\varphi \leq \varphi_{TT}(q, \pi_A, \pi_I)$.*

Intuitively, after observing signal s, the authority prefers truth-telling over coercion if and only if the cost of coercion, q, exceeds the benefit. The benefit from coercion is an improvement in action choice: the authority implements $a_A(s)$, whereas the individual would implement $a_I(s) \neq a_A(s)$. Because the authority believes that this benefit also accrues to the individual, she values the benefit less, the less she cares about the individual. When the level of altruism falls below some threshold, she therefore prefers to pass on signal s truthfully. A FRE exists when the authority prefers truthful communication over coercion after observing both signals; that is, when both ($\mathrm{TT1}_{\mathrm{op,C}}$) and ($\mathrm{TT0}_{\mathrm{op,C}}$) are satisfied (Figure 3.10 depicts this region of (q, φ)).

The authority who coerces. An authority who cares more about the individual, however, exercises her power to force the individual's action choice. From the above discussion about truth-telling equilibria when $(\pi_A, \pi_I) \in TT'$, it follows immediately that partial coercion equilibria exist for the intermediate levels of φ for which one of the two conditions, ($\mathrm{TT0}_{\mathrm{op,C}}$) and ($\mathrm{TT1}_{\mathrm{op,C}}$), is violated. To see this, suppose w.l.o.g. that φ exceeds the threshold in ($\mathrm{TT0}_{\mathrm{op,C}}$), but not the threshold in ($\mathrm{TT1}_{\mathrm{op,C}}$). Then, an equilibrium exists in which the authority coerces after getting the signal $s = 0$, but refrains from coercion otherwise. Clearly, when the individual gets to implement an action, he realizes that the authority observed the signal $s = 1$; consequently, he chooses $a_I(1)$. If φ rises further, so that ($\mathrm{TT1}_{\mathrm{op,C}}$) is violated as well, an equilibrium exists in which the authority coerces after observing both signals. Intuitively, the partial coercion equilibrium breaks down because the authority prefers the action $a_A(1)$ over $a_I(1)$ after observing $s = 1$, and the threshold in ($\mathrm{TT1}_{\mathrm{op,C}}$) defines the level of altruism above which the authority's expected benefit from improving the action choice from $a_I(1)$ to $a_A(1)$ exceeds the cost of coercion.

In general, moving beyond pure strategy equilibria and considering all (π_A, π_I), there exist equilibria in which the authority mixes between (i) coercion and communication; and (ii) sending different messages when communicating. The above intuition carries on to the general case as well: as established formally in the following Proposition, for levels of altruism exceeding some threshold, some coercion necessarily occurs in equilibrium; and for levels of altruism exceeding some higher threshold, the authority always coerces (see Figure 3.10 for schematic illustration of these parameter sets).

Proposition 15. *For any coercion cost q and for any (π_A, π_I), there exist thresholds $\varphi_C(q, \pi_A, \pi_I) \leq \varphi_{CC}(q, \pi_A, \pi_I)$ such that, for $\varphi > \varphi_C(q, \pi_A, \pi_I)$, every equilibrium involves the authority coercing*

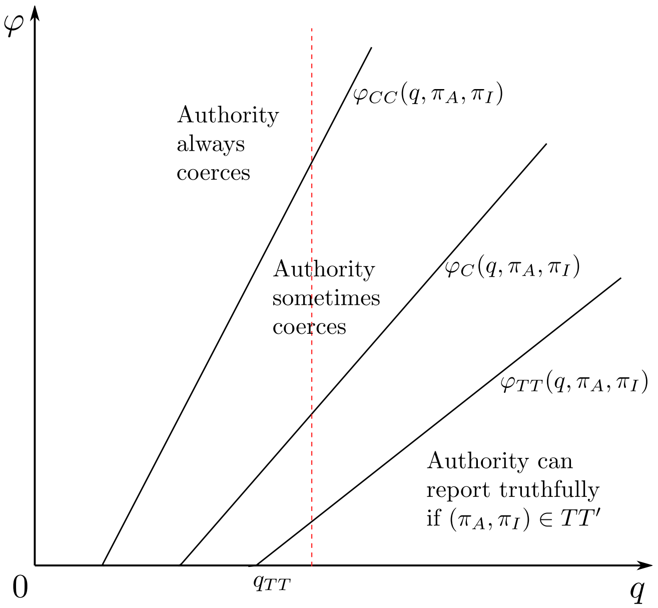


Figure 3.10: Equilibrium outcomes under conflicting opinions.

the individual with strictly positive probability. For $\varphi > \varphi_{CC}(q, \pi_A, \pi_I)$, there exists a unique equilibrium, in which the authority coerces the individual with probability one, regardless of the signal s she obtains.

Proof. See Appendix. □

To show that coercion after both signals eventually is the unique equilibrium, we proceed in two steps. First, we show that in every equilibrium such that the authority coerces with probability one after getting one signal, and communicates with some probability $p \in (0, 1]$ after getting the other signal, the agent takes his preferred action conditional on the signal for which the authority communicates. Because $a_A(s) \neq a_I(s)$, the authority can improve upon the individual's action choice; consequently, for a high enough level of altruism, she prefers to coerce for both signals. Second, we show that in every equilibrium where the authority communicates with strictly positive probability for both signals, the actions that the agent is induced to take cannot be $p_A(s)$ after the authority observes s for both signals. Thus, for a sufficiently high level of altruism, the authority will find it worthwhile to improve the action choice(s).

Together, Lemma 19 and Proposition 15 show that, under opinion disagreement, a tyrannical authority (with $\varphi = 0$) may communicate truthfully with the individual. A sufficiently benevolent authority, however, counteracts the will of this individual. Intuitively, the authority forces the individual to give up his preferred action choice, even though it is costly for her, because she believes that this, ultimately, is in the individual's best interest. In essence, she employs coercion to protect the individual from (what she believes to be) his own misguided view of the world. Thus, the authority behaves in a *paternalistic* fashion: she restricts the individual's freedom solely on the grounds that it makes him better off.

3.11.4 Granting authority

In Section 3.11.3, we considered an authority who can, at some cost $q > 0$, coerce the individual. In this section, we endogenize the coercion cost. In particular, we add a stage zero to the game analyzed in Section 3.11.3, in which the individual chooses some cost of coercion $q \in [0, \infty]$. If, given the chosen cost of coercion, multiple equilibria exist, we assume that the individual's preferred equilibrium is induced.

Intuitively, we can think of the individual as deciding how readily coercion should be available to the authority. Put differently, the individual chooses *how much* authority to concede. If he chooses $q = \infty$ (or a sufficiently large finite q), coercion is de facto made unavailable; hence, we say that the individual grants *no authority*. If he chooses $q = 0$ (or a sufficiently small $q > 0$), the authority is given the ability to coerce at almost no cost. In this case, (in the unique equilibrium) the authority chooses to coerce—regardless of the nature of disagreement—for both signals: coercion enables her to implement exactly her preferred choice of action, whereas the individual would choose a different action after at least one signal realization. Hence, we refer to this case as the individual granting *full authority*. If the individual chooses some intermediate cost of coercion that triggers partial coercion in equilibrium, we say that he grants *intermediate authority*.

[Dessein, 2002] analyzes the problem of an uninformed party (in our setting, the individual) who chooses between the two extreme options—granting no authority and granting full authority—in a setting where the parties' have different preferences. When the uninformed party grants the informed party full authority, [Dessein, 2002] refers to it as *delegation*.[45] We generalize the analysis

[45] In [Dessein, 2002], the uninformed party first chooses whether or not to delegate. Then, the informed party observes the signal, after which she always picks the action; hence, after observing the signal, she does not face any choice between coercion and communication. In our context, if the individual sets $q = 0$ (or sufficiently small $q > 0$), then the authority always chooses to coerce. Hence, if the individual sets $q = 0$, it does not matter whether the informed party has the right to choose whether to coerce or not. Thus, if the individual sets $q = 0$, he delegates the

in [Dessein, 2002] in the sense that we consider the full range of possible costs of coercion $q \in [0, \infty]$; moreover, we consider both types of disagreement. [46]

Proposition 16. *Consider either type of disagreement between the individual and the authority. If a FRE exits when $q = \infty$, then the individual grants no authority. If no FRE exists when $q = \infty$, then the individual chooses to grant intermediate authority if the divergence in preferences or opinions is limited, and to grant no authority under severe disagreement; the individual never finds it strictly optimal to grant full authority.*

When a FRE exists if the individual makes coercion infinitely costly, the individual optimally concedes no authority. Intuitively, he has no incentive to give up his decision-making powers when his most preferred alternative is available: he gets access to the authority's information, and he then implements his desired action. When no fully revealing equilibrium exists, however, the individual does not make coercion infinitely costly.

The decision to grant some authority. Why does the individual grant some authority? That is, why does he refrain from choosing $q = \infty$? When informative communication cannot be sustained, the individual faces a trade-off between loss of information (if he keeps the decision-making powers) and loss of control (if he concedes them). This logic is demonstrated by [Dessein, 2002], who shows that the individual may prefer to grant full authority over granting no authority. Proposition 16 shows that this trade-off between loss of control and loss of information carries on to the case of disagreement in priors: indeed, when no FRE exists and the divergence in opinion is limited, the individual does not make coercion infinitely costly.

The decision not to grant full authority. What keeps the individual from granting full authority? That is, why does he refrain from choosing $q = 0$? When the individual can choose any cost of coercion $q \in [0, \infty]$, he chooses an intermediate cost, thereby granting partial authority. The mechanism that drives this result is novel, as it arises only when we allow the individual to choose intermediate costs of coercion.

decision in the precise sense of [Dessein, 2002].

[46] Dessein assumes that, if multiple equilibria exist, then the equilibrium preferred by the two players is chosen. We assume that, if multiple equilibria exist, the equilibrium preferred by the individual is chosen; this differs from the assumption of Dessein whenever, in our setting, multiple equilibria exist and the two players' preferred equilibria do not coincide.

For intermediate costs of coercion, *partial coercion equilibria* exist. In a partial coercion equilibrium, the authority coerces after getting one signal, while communicating with positive probability after getting the other signal. Without loss of generality, suppose that she coerces after getting the signal $s = 0$. Clearly, when the individual gets to implement an action, this perfectly reveals that the authority observed the signal $s = 1$; consequently, he chooses his preferred action $a_I(1)$. From the individual's perspective, this equilibrium is strictly preferred to the equilibrium in which the authority coerces with certainty after getting both signals: then, the authority implements $a_A(1) \neq a_I(1)$ after getting the signal $s = 1$. Yet, if the individual grants full authority, the unique equilibrium is the equilibrium in which the authority coerces after both signals. Consequently, the individual always prefers to set the coercion cost high enough to induce some partial coercion equilibrium.[47]

Self-paternalism. Under conflicting opinions, the result that the individual may voluntarily concede some authority can be interpreted as an instance of *self-paternalism* (reference): The individual is aware of the fact that, if he grants a benevolent leader authority, the leader engages in coercion (for the signal $s = 0$). On the contrary, if the individual does not grant the benevolent leader authority, no such coercion will take place. Ex ante, he may prefer to grant partial authority over retaining all decision-powers. Ex post, however, he may regret having granted this authority. More precisely, if the authority obtains the signal $s = 0$, she implements her preferred action $a_A(0)$, which may be strictly worse for the individual than the action that she would have taken, for example, in a babbling equilibrium, $a = \pi_I$.

Such self-paternalism is, in fact, consistent with the *harm principle*. To see this, recall that the harm principle holds that individual freedom should take priority over benevolent legislative limits on our liberty. Thus, an individual that, ex ante, asks, of free will, to be coerced for his own good in the future, de facto paternalizes himself. Because he explicitly consented to the regimen in which others paternalize him, such self-paternalism is consistent with the harm principle.

[47] We conjecture that the result that the uninformed party would choose an intermediate cost of coercion would arise in [Dessein, 2002] as well if the uninformed party were allowed to choose an intermediate cost of coercion. As [Dessein, 2002] considers a more general state space, partial coercion would not perfectly reveal which signal the informed party observed; however, it would be partially informative.

Moderate regulatory powers. Most governments have some powers to enact regulation. At the same time, governments are typically not empowered with unconstrained powers to coerce. (In fact, if a government has such extensive powers, our framework suggests that they have been assumed; not granted!) Such moderate regulatory powers is precisely what our framework suggests that a populace would consent to yielding to their (benevolent) government. Ex ante, the populace may find it optimal to write a constitution that does not prohibit its leaders from coercion; nevertheless, to ensure that coercion does not occur more often than necessary, coercion entails some cost. For example, the cost of coercion can be thought of as the trouble that the government has to go through to enact a new law: debates, several rounds of acceptance, etc.

3.11.5 Extensions

3.11.5.1 General conflict

We now allow the parties to have both conflicting preferences and conflicting opinions. In this setting, conditional on the signal s, the individual's preferred action is $a_I(s) = p_I(s) + b$ and the advisor's preferred action is $a_A(s) = p_A(s) + \frac{\varphi}{1+\varphi}b$.

The altruistic advisor. Stronger altruism improves communication under (pure) conflicting preferences, but impedes it under (pure) conflicting priors. We now ask how (stronger) altruism affects communication under two-dimensional conflict. To simplify the exposition, we assume a zero lying cost, $c = 0$.

The advisor reports $s = 1$ truthfully if and only if inducing $a_I(1)$ yields a higher expected utility than inducing $a_I(0)$. This can be written

$$(\varphi + 1)[2p_A(1) - (p_I(1) + p_I(0))] \geq 2b. \tag{$TT1_{\text{general}}$}$$

Similarly, the incentive compatibility constraint for truthful reporting of $s = 0$ is

$$(\varphi + 1)[-2p_A(0) + (p_I(1) + p_I(0))] \geq -2b. \tag{$TT0_{\text{general}}$}$$

Clearly, whenever the coefficients on $(\varphi + 1)$ in ($TT1_{\text{general}}$) and ($TT0_{\text{general}}$) exceed zero, both constraints are satisfied for sufficiently high φ. This condition coincides with (TT'_{op}); thus, whenever the priors (π_A, π_I) are in the region TT' depicted in Figure 3.7, raising the level of altruism, φ, eventually leads to the existence of a FRE. On the other hand, if at least one of the coefficients on $(\varphi + 1)$ in $(TT1_{\text{general}})$ or $(TT0_{\text{general}})$ is strictly below 0, then the corresponding constraint is violated for a sufficiently large φ; hence, no FRE exists. This corresponds to (π_A, π_I) outside of the region TT'.

Proposition 17. *For any bias b and for any priors π_A, π_I, we have that: (i) if $(\pi_A, \pi_I) \in TT'$, there exists a level of altruism above which a FRE exists; (ii) if $(\pi_A, \pi_I) \notin TT'$, there exists a level of altruism above which a FRE cannot exist.*

Remark. The set of parameter values for which a FRE can be achieved are independent of b. In particular, they are equivalent to the set of parameter values for which a FRE exists when $b = 0$.

Thus, when the conflict is two-dimensional, the effect of altruism crucially depends on how severe the opinions conflict is. If, on the one hand, the opinions conflict is moderate, so that truthful communication possible when $b = 0$ (i.e., if $(\pi_A, \pi_I) \in TT'$), then any negative effect on communication arising from conflicting preferences can be offset by an increase in the level of altruism. This reflects the intuition in the pure conflicting preferences case.

If, on the other hand, the opinions conflict is severe, so that a FRE does not exist when $b = 0$ (i.e., if $(\pi_A, \pi_I) \notin TT'$), then too high a level of altruism destroys communication. To see why this is the case, consider some $(\pi_A, \pi_I) \notin TT'$ such that the advisor is considerably more convinced that the state of the world is 1 than is the individual. Then, increasing altruism will eventually lead to truth-telling when she gets the signal $s = 1$. However, increasing the level of altruism will also, eventually, lead to the advisor withholding information from the individual when the advisor gets the signal $s = 0$: when altruism is strong enough, the preference divergence is mitigated, so the advisor starts to lie (just as she would do if there were no preference conflict).

Intuitively, when the conflict is two-dimensional, a high level of altruism mitigates the preference divergence, which eventually brings us back to the case of (pure) conflicting opinions.

Non-monotonicity of truthful communication. Interestingly, if $(\pi_A, \pi_I) \notin TT'$, the level of communication need not be a monotonic function of the level of altruism. In particular, a FRE may exist only for some intermediate levels of altruism. For example, if $\pi_A = 0.9,\ \pi_I = 0.6,\ b = 0.4,$

and $\gamma = 0.6$, then a FRE exists only for $\varphi \in [0.194, 0.533]$. Intuitively, for $\varphi < 0.194$, the preference conflict destroys truthful communication. In particular, the advisor's IC constraint for truthful reporting of $s = 1$ is violated: she prefers to misreport, to induce a lower action. When the level of altruism increases, the preferences of the advisor and the individual are gradually aligned, and at $\varphi = 0.194$, advisor finds it optimal to report $s = 1$ truthfully (as well as $s = 0$). Once the preference conflict is mitigated, however, raising the level of altruism further has the same effect as in the pure conflicting opinions setting: for high enough level of altruism (in this case, for $\varphi > 0.533$), the advisor cares so much about the individual that she withholds information to protect him. In this particular case, the advisor believes that the individual underestimates the probability that $\theta = 1$. Consequently, she prefers to misreport the signal $s = 0$, in order to alleviate the individual's disutility from an (in the advisor's view) incorrect action choice.

The altruistic authority. When the advisor is endowed with the ability to coerce, the same results obtain. In particular, whether increasing altruism eventually leads to coercion solely depends on the difference in priors: (i) for $(\pi_A, \pi_I) \in TT'$, there exists a level of altruism above which a FRE exists; (ii) if $(\pi_A, \pi_I) \notin TT'$, there exists a level of altruism above which the authority necessarily coerces after both signals.

3.11.5.2 Mutual altruism

A natural extension of the model is to consider a setting in which the individual holds the same regard for the advisor/authority as she does for him. Then, his utility function is given by $U_I(a|\theta) = -(a - \theta - b(\theta))^2 - \varphi(a - \theta)^2$. In what follows we show that our main results hold in this setting.

Conflicting preferences. First, we consider communication between the individual and the advisor. Conditional on the signal s, the individual and the advisor hold the same posterior beliefs $p(s) = \Pr(\theta = 1|s)$. Because both players take each others' preferences into account, their optimal actions are weighted sums of the preferred actions of a non-altruistic advisor, $p(s)$, and a non-altruistic individual, $p(s) + b$:

$$\begin{aligned} a_I(s) &= \frac{1}{1+\varphi}(p(s)+b) + \frac{\varphi}{1+\varphi}p(s), \\ a_A(s) &= \frac{\varphi}{1+\varphi}(p(s)+b) + \frac{1}{1+\varphi}p(s). \end{aligned}$$

Naturally, as φ increases, the players' preferred actions converge.

To find conditions for truthful reporting, we assume that the individual believes that the advisor reports truthfully. Then the advisor reports the signal $s = 1$ truthfully if and only if $\mathbb{E}[U_A(a_I(1), \theta)|s = 1] \geq \mathbb{E}[U_A(a_I(0), \theta)|s = 1]$, which can be written

$$\varphi(p(1) - p(0) + 2b) \geq -(p(1) - p(0)) + 2b - \frac{c}{p(1) - p(0)}. \qquad (TT1_{\text{pr,mutual}})$$

Analogously, the advisor reports $s = 0$ truthfully if and only if

$$\varphi(p(1) - p(0) - 2b) \geq -(p(1) - p(0)) - 2b - \frac{c}{p(1) - p(0)}. \qquad (TT0_{\text{pr,mutual}})$$

Truth-telling conditions ($TT1_{\text{pr,mutual}}$) and ($TT0_{\text{pr,mutual}}$) imply that a higher level of altruism improves communication. More precisely, there exists a threshold $\overline{\varphi}(c, b) \in [0, 1)$, s.t a FRE exists if and only if $\varphi \geq \overline{\varphi}(c, b)$.

Next, we consider the individual's interaction with the authority. The result that a FRE can be sustained for high enough levels of altruism remains in this setting, because for a sufficiently large φ, (i) truth-telling is incentive compatible, and (ii) the preferred actions of the individual and the advisor converge, which reduces the benefit from coercion. Clearly, the results regarding coercion for low levels of altruism remain valid as well.

Conflicting opinions. Because the utility function of the advisor/authority and the preferred action of the individual given the signal s, $a_I(s) = p_I(s)$, are the same as in the previously considered setting with a non-altruistic individual, the analysis—and hence all of the results—is identical in this alternative case.

3.11.6 Appendix

Proof of Proposition 14. We prove a more general version of Proposition 14 for state dependent preferences: For any coercion cost $q > 0$ and any preference biases $b_0 = b(0)$, $b_1 = b(1)$, there exists a threshold $\varphi_{TT}(q, b_0, b_1)$, s.t. a FRE exists if and only if $\varphi \geq \varphi_{TT}(q, b_0, b_1)$.

If the authority who obtained signal s decides to coerce the individual, she optimally chooses action

$$a_A(s) = \frac{1}{1+\varphi} p(s) + \frac{\varphi}{1+\varphi}(p(s) + \mathbb{E}(b(\theta)|s)) = p(s) + \frac{\varphi \mathbb{E}(b(\theta)|s)}{1+\varphi}.$$

To simplify the notational exposition, denote $p = p(s)$, $a = a_A(s)$, $\bar{b} = \mathbb{E}(b(\theta)|s)$ and $\overline{b^2} = \mathbb{E}(b(\theta)^2|s)$. Then the expected utility of the authority is

$$\begin{aligned}
\mathbb{E}(U_A(a,\theta)|s) &= -q - p\left[(a-1)^2 + \varphi(a-1-b_1)^2\right] - (1-p)\left[a^2 + \varphi(a-b_0)^2\right] \\
&= -q - p\left[(1+\varphi)(a-1)^2 - 2\varphi(a-1)b_1 + \varphi b_1^2\right] - (1-p)\left[(1+\varphi)a^2 - 2a\varphi b_0 + \varphi b_0^2\right] \\
&= -q - (1+\varphi)\left[(a-p)^2 + p(1-p)\right] + 2\varphi[a\bar{b} - pb_1] - \varphi\overline{b^2} \\
&= -q - (1+\varphi)p(1-p) + \frac{\varphi^2 \bar{b}^2}{(1+\varphi)} - 2\varphi p(1-p)(b_1 - b_0) - \varphi\overline{b^2}.
\end{aligned}$$

Assume that the individual's beliefs are such that the individual takes action a_1 when he receives $m = 1$ and a_0 when $m = 0$. The authority who obtained the signal s prefers communicating over coercing and imposing her preferred action iff

$$\max_{\tilde{a} \in \{a_0, a_1\}} \mathbb{E}(U_A(\tilde{a},\theta)|s) \geq \mathbb{E}(U_A(a_A(s),\theta)|s).$$

Suppose that the maximum of the left-hand side is achieved at $\tilde{a}$ that generates the expected utility of

$$\mathbb{E}(U_A(\tilde{a},\theta)|s) = -(1+\varphi)\left[(\tilde{a}-p)^2 + p(1-p)\right] + 2\varphi[\tilde{a}\bar{b} - pb_1] - \varphi\overline{b^2}.$$

State independent biases. When $b_0 = b_1$, the expected utility of the authority after imposing her preferred action simplifies to

$$\mathbb{E}(U_A(a_A(s),\theta)|s) = -q - (1+\varphi)p(1-p) - \frac{\varphi}{(1+\varphi)} b^2.$$

The authority's expected payoff from inducing action $\tilde{a} = \tilde{p} + b$, where $\tilde{p}$ is the individual's posterior, can be expressed as

$$\begin{aligned} \mathbb{E}\left(U_A(\tilde{a}, \theta)|s\right) &= -(1+\varphi)\left[(\tilde{p}-p)^2 + 2b(\tilde{p}-p) + b^2 + p(1-p)\right] + 2\varphi[b^2 + b(\tilde{p}-p)] - \varphi b^2 \\ &= -(1+\varphi)(\tilde{p}-p)^2 - (1+\varphi)p(1-p) - 2b(\tilde{p}-p) - b^2. \end{aligned}$$

Communication and inducing action $\tilde{a}$ generates greater expected payoff to the authority than coercion if

$$-(1+\varphi)(\tilde{p}-p)^2 - 2b(\tilde{p}-p) - b^2 \geq -q - \frac{\varphi}{(1+\varphi)}b^2,$$

which can be simplified to

$$q(1+\varphi) \geq \left[(1+\varphi)(\tilde{a}-p) + b\right]^2.$$

In particular case of truthful advising this boils down to

$$q(1+\varphi) \geq b^2.$$

Hence, the threshold for the level of altruism is $\varphi_{TT}(q, b) = \max\{-1 + \frac{b^2}{q}\ , -1 + \frac{2|b|}{p(1)-p(0)}\}$, where $\overline{\varphi}(c = 0, b) = -1 + \frac{2|b|}{p(1)-p(0)} =$ is the threshold for truthful reporting in case of advising.

State dependent biases. Under truthful communication the individual chooses the action $a_I(s) = p(s) + \bar{b}$, i.e., $\tilde{a} = p + \bar{b}$. This implies

$$\mathbb{E}(U_A(\tilde{a}, \theta)|s) = -(1+\varphi)\left[\bar{b}^2 + p(1-p)\right] + 2\varphi[\bar{b}^2 - p(1-p)(b_1 - b_0)] - \varphi\overline{b^2}.$$

The authority prefers truthful reporting of the signal s to coercion whenever

$$\begin{aligned} & -(1+\varphi)\left[\bar{b}^2 + p(1-p)\right] + 2\varphi[\bar{b}^2 - p(1-p)(b_1 - b_0)] - \varphi\overline{b^2} \\ \geq\ & -q - (1+\varphi)p(1-p) + \frac{\varphi^2\bar{b}^2}{(1+\varphi)} - 2\varphi p(1-p)(b_1 - b_0) - \varphi\overline{b^2}, \end{aligned}$$

that simplifies to

$$\begin{aligned} -(1+\varphi)\bar{b}^2 + 2\varphi\bar{b}^2 &\geq -q + \frac{\varphi^2\bar{b}^2}{(1+\varphi)}, \\ q(1+\varphi) &\geq \bar{b}^2. \end{aligned}$$

As a result, truthful communication generates a greater expected payoff for the authority if $q(1+\varphi) \geq [\mathbb{E}(b(\theta)|s)]^2$, for every $s \in \{0,1\}$. Hence, the threshold for the level of altruism is

$$\varphi_{TT}(q, b_0, b_1) = \max\{-1 + \frac{1}{q}\max_{s\in\{0,1\}} [\mathbb{E}(b(\theta)|s)]^2 \ , \overline{\varphi}(c = 0, b(0), b(1))\},$$

where $\overline{\varphi}(c = 0, b(0), b(1))$ is the threshold for truthful reporting in case of advising. **QED.**

Proof of Lemma 18. As regards the first part of the Lemma, we show that sufficiently low cost q and altruism φ preclude the existence of communication equilibria. We study two different types of communication equilibria: (i) only one action $\hat{a}$ is induced, and (ii) two different actions $\hat{a}_0 < \hat{a}_1$ are induced.

In case (i), the authority induces the same action independently of the underlying signal, i.e., no informative communication takes place and the induced posterior belief necessarily coincides with the prior, meaning that $\hat{a} = \pi + b$. Such a communication equilibrium fails to exist when, after some signal s, the authority would rather exercise her power to choose her preferred action, i.e., when

$$q(1+\varphi) < [(1+\varphi)(\pi - p(s)) + b]^2 .$$

Assume that the authority does not incorporate the individual's payoff into her utility: $\varphi = 0$. Then, for a sufficiently low cost $q < \max_s[\pi - p(s) + b]^2$, the authority coerces the individual for at least one signal s. Clearly, for any such q and for a sufficiently low level of altruism, the authority chooses to coerce as well.

Now consider case (ii), in which the induced actions are different, $\hat{a}_0 < \hat{a}_1$. The authority cannot be indifferent between the actions $\hat{a}_0$ and $\hat{a}_1$ after both signal realizations. Thus, for these actions to be induced in equilibrium, it must be the case that the authority mixes between messages for no more than one signal. It implies that $\hat{a}_s = p(s) + b$ for at least one signal s. Hence, a non-altruistic authority who obtained s will choose to impose her preferred action whenever $q < b^2$. Clearly, she will still do so for sufficiently low levels of altruism.

Together, cases (i) and (ii) imply that for a sufficiently low coercion cost, $q < q_C(b)$, and altruism level, $\varphi < \varphi_C(q,b)$, each equilibrium involves coercion with a non-zero probability.

As regards the second part of the Proposition, we argue that for sufficiently low q and φ no communication is sustainable in equilibrium. Below we consider two possibilities of how communication can arise.

First, assume that the authority communicates with positive probability for both signal realizations $s = 0, 1$. If only one action $\hat{a} = \hat{p} + b$, where $\hat{p} \in [p(0), p(1)]$, is induced in equilibrium, then a non-altruistic authority would strictly prefer to coerce for at least one signal when

$$q < q(\hat{p}, b) = \max_{s \in \{0,1\}} [\hat{p} - p(s) + b]^2. \qquad (\text{Coercion}_{\text{pr,1 action}})$$

Clearly, $q(\hat{p}, b) > 0$ for every $\hat{p}$. Because $q(\hat{p}, b)$ is a continuous function of $\hat{p}$, there exists $0 < \bar{q}(b) = \min_{\hat{p} \in [p(0),p(1)]} q(\hat{p}, p)$. The inequality ($\text{Coercion}_{\text{pr,1 action}}$) is strict for any $q < \bar{q}(b)$ and $\hat{p} \in [p(0), p(1)]$; hence, the authority still prefers to coerce whenever the level of altruism is below the threshold $\varphi(q, \hat{p}, b) > 0$.[48] Because $\varphi(q, \hat{p}, b)$ is continuous in $\hat{p} \in [p(0), p(1)]$, there exists

$$0 < \overline{\varphi}(q, b) = \min_{\hat{p} \in [p(0),p(1)]} \varphi(q, \hat{p}, b).$$

As a result, for $q < \bar{q}(b)$ and $\varphi < \overline{\varphi}(q, b)$, there exists no equilibrium with non-zero communication after both signals and only one induced action. Next, if two different actions $\hat{a}_0 < \hat{a}_1$ are induced in equilibrium, then the authority strictly prefers inducing one action over the other for at least one signal, meaning that $\hat{a}_s = p(s) + b$ for some $s \in \{0, 1\}$. Clearly, such an equilibrium cannot exist for q lower than b^2 and sufficiently low φ.

Second, suppose that the authority communicates with non-zero probability for one signal and coerces with certainty for the other signal. For example, assume that the authority sometimes communicates after $s = 1$ and always coerces after $s = 0$. In this case the mere fact that he was not coerced indicates to the individual that the signal is $s = 1$. As a result, the induced action $\hat{a}$ necessarily equals $p(1) + b$. Such an equilibrium fails to exist when the authority is better off

[48] Assuming that the maximum in ($\text{Coercion}_{\text{pr,1 action}}$) is achieved at $\hat{s}$, the threshold $\varphi(q, \hat{p}, b) > 0$ is a closest to 0 root of $q(1 + \varphi) = [(1 + \varphi)(\hat{p} - p(\hat{s})) + b]^2$.

by imposing the action herself than by inducing $\hat{a} = p(1) + b$, i.e., $q(1+\varphi) < b^2$. The case when the authority sometimes communicates the low signal $s = 0$ and always coerces when $s = 1$ is analogous. It follows that for every $q < b^2$, communicating after only one signal cannot be an equilibrium if φ is sufficiently low.

Combining the results of the two cases, when the coercion cost q and altruism level φ are below the respective bounds $q_{CC}(b_0, b_1)$ and $\varphi_{CC}(q, b_0, b_1)$, the unique equilibrium is the one where the authority always coerces the individual. **QED.**

Proof of Proposition 15. To prove the first part of the Proposition, we argue that all equilibria with communication after both signal realizations break down for a sufficiently high level of altruism φ. To show this, we distinguish between (i) communication equilibria where only one action $\hat{a}$ is induced, and (ii) communication equilibria where two different actions $\hat{a}_0 < \hat{a}_1$ are induced.

In case (i), the communicating authority induces the same action for each signal, which necessarily implies uninformative communication, and hence, the individual chooses actions equal to his prior belief, $\hat{a} = \pi_I$. Because the authority's ideal action choice depends on the signal she observes, $a_A(0) \neq a_A(1)$, the individual's implemented action π_I cannot correspond to the authority's desired action for both signal realizations. Hence, coercion enables the authority to improve the action choice after observing at least one signal s. Comparing the expected payoffs, the authority prefers to coerce after observing signal s if $\varphi > \frac{q}{(a_A(s)-\pi_I)^2} - 1$. Consequently, for a large enough degree of altruism the authority coerces the individual after at least one signal realization, so such a communication equilibrium breaks down.

In case (ii), the communicating authority induces the individual to take different actions $\hat{a}_0 < \hat{a}_1$. Clearly, rational behavior on the part of the individual ensures that $\hat{a}_0$, $\hat{a}_1 \in [p_I(0), p_I(1)]$. Because the authority cannot be indifferent between these actions after both signal realizations $s = 0$ and $s = 1$, it must be the case that the authority mixes between the messages for at most one signal. Thus, $\hat{a}_s = p_I(s)$ for at least one $s \in \{0, 1\}$. Because the authority can improve on this action choice, she prefers coercion over communication when her concern for the individual is sufficiently strong, $\varphi > \frac{q}{(p_I(s)-p_A(s))^2} - 1$.

Combining the results of the two cases, we obtain that for a sufficiently high level of altruism, $\varphi > \varphi_C(q, \pi_A, \pi_I)$ (where the subscript C denotes coercion), the authority cannot communicate with probability one in equilibrium.

Regarding the second part of the Proposition, to show that for a sufficiently large φ no communication is sustainable in equilibrium, we consider two manners in which information transmission can arise: (*i*) the authority communicates with positive probability after both signal realizations, and (*ii*) the authority communicates with non-zero probability after one signal realization and coerces with certainty after the other.

In case (*i*), we consider two possibilities: either one action $\hat{a}$ is induced in equilibrium, or two different actions $\hat{a}_0 < \hat{a}_1$ are induced in equilibrium. First, if only one action $\hat{a}$ is induced, then the authority would prefer to coerce for at least one signal when

$$\varphi > \varphi(q, \pi_A, \pi_I, \hat{a}) = \frac{q}{\max\{(\hat{a} - p_A(1))^2, (\hat{a} - p_A(0))^2\}} - 1.$$

Because $\varphi(q, \pi_A, \pi_I, \hat{a})$ is continuous in $\hat{a} \in [p_I(0), p_I(1)]$, there exists a maximum

$$0 < \overline{\varphi}(q, \pi_A, \pi_I) = \max_{\hat{a} \in [p_I(0), p_I(1)]} \varphi(q, \pi_A, \pi_I, \hat{a}).$$

As a result, for $\varphi > \overline{\varphi}(q, \pi_A, \pi_I)$, there exists no equilibrium with non-zero communication for both signals and only one induced action $\hat{a}$. Second, if the induced actions are different, $\hat{a}_0 < \hat{a}_1$, then the authority strictly prefers inducing one action over the other for at least one signal, meaning that $\hat{a}_s = p_I(s)$ for some $s \in \{0, 1\}$. This implies that such a communication equilibrium cannot exist for $\varphi > \frac{q}{(p_A(s) - p_I(s))^2} - 1$.

In case (*ii*), suppose, for example, that the authority coerces for sure after getting the signal $s = 0$. In this equilibrium, the authority's decision to communicate perfectly reveals that she observed the signal $s = 1$, so the individual implements $a_I(1) = p_I(1)$. Clearly, for a high enough level of altruism, $\varphi > \frac{q}{(p_A(1) - p_I(1))^2} - 1$, the authority would choose to improve upon this action; consequently, such an equilibrium breaks down.

Combining the results of (*i*) and (*ii*), for sufficiently high $\varphi > \varphi_{CC}(q, \pi_A, \pi_I)$ the unique equilibrium is the one where the authority always coerces the individual. **QED.**

Chapter 4

Strategic communication with guilt aversion

Uliana Loginova

Abstract

I study a model of strategic communication between an informed Sender and an uninformed Receiver, who needs to take an action affecting both players. The Sender has a preference bias and is guilt averse to letting down the Receiver's payoff expectations, meaning that the Sender incurs some cost if the Receiver's actual payoff is lower than he expected. I show that no separating equilibrium exists; rather, in case of uniform state of the world and quadratic utilities, I demonstrate that there exist partition equilibria, in which the Sender effectively reports only the interval where the state lies. An increase in the guilt aversion intensity has similar equilibrium effects as a decrease in the preference divergence: holding the number of elements in the partition fixed, greater guilt aversion intensity results in more balanced intervals, and higher guilt aversion intensity allows for more intervals in the equilibrium partition. Finally, I show that both, the Sender and the Receiver, prefer an equilibrium with more elements in the partition, and if players coordinate on the number of steps in the partition, the Receiver prefers a more guilt averse Sender.

4.1 Introduction

In many cases, decision makers rely on the experts' reports about the state of the world. However, the expert's preferences over the actions are often different from the preferences of the decision maker, which gives rise to strategic communication of the expert's private information. The general framework of this situation is the following. First, the expert (Sender) obtains private information about the state of the world (also called the type of the Sender). Second, the Sender communicates his private information to the decision maker (Receiver). Finally, the Receiver takes an action. The most popular and fundamental ways of modeling the communication process in this setting are "verifiable information transmission" and "cheap talk". Under verifiable information transmission, introduced by [Grossman, 1981] and [Milgrom, 1981], the Sender may withhold some parts of the information, but not lie. On the contrary, under cheap talk, pioneered by [Crawford and Sobel, 1982] and [Green and Stokey, 2007], the information is unverifiable and the Sender can misreport at no cost.

Framing these two approaches in terms of lying, verifiable information transmission means that the Sender faces *infinite* costs of lying, while cheap talk implies *zero* costs of lying for the Sender. However, a more realistic approach would be to assume that the Sender faces *some* costs of lying. Clearly, the nature of the lying costs can be quite different. For example, there may be time and effort costs of creating a false message and misreporting the numbers. Another example is a possible punishment or a monetary fine once the falsification is detected. Finally, there might be an intrinsic aversion to lying, which itself has several reasons. One is that people might have preference to be honest and not to lie per se. Another reason is that people might not like to feel shame, i.e., to appear in a situation in which others learn about lying. Finally, people might dislike letting down others' expectations, which will be referred to as "guilt aversion".[1]

Potential auditing penalties, technological costs of manipulating information, or psychological costs stemming from the preference not to lie per se, refer to the case of "literal" costs, in which the lying cost depends only on the state of the world and the message sent. Mathematically, the cost function, $c(t', t)$, depends on the actual state t and the reported state t'. This type of costs is considered in [Kartik, 2009]. He shows that introducing the literal cost of lying in the basic model

[1] [Tadelis, 2011] studies the difference between shame and guilt aversion.

of [Crawford and Sobel, 1982] (hereafter CS) results in the incomplete separation with an inflated language for sufficiently low types and some pooling on the highest messages. [Kartik *et al.*, 2007] also consider the case of the literal lying cost, but in a setting, in which the state space has an unbounded support. They find that a fully-separating equilibrium exists even for an arbitrary small intensity of the lying cost.

In this chapter, I study strategic communication setting of CS with the guilt aversion type of the psychological cost on part of the Sender, $k \cdot c(x, e)$, where k is the cost intensity parameter. I follow [Battigalli and Dufwenberg, 2007] in considering guilt aversion as a disutility from letting down the Receiver. To be more precise, I say that the Sender lets down the Receiver, if the Receiver's actual payoff x after the message m appears to be lower than the Receiver's expected payoff e (conditional on the message m). It is straightforward to see, that the psychological cost of lying arising from guilt aversion is *not* a specific case of the literal cost approach. Indeed, in the guilt aversion case, the Sender suffers from the cost when the Receiver gets a lower payoff than the Receiver expected. Hence, the level of the cost depends on how the Receiver interprets the message and forms his payoff expectation, and not on the literal difference between the actual and the reported states.

Recall, that the outcome of the classic CS model is a partition equilibrium, in which the Sender indicates only the interval where the state of the world lies. As already mentioned, introducing literal cost of lying ([Kartik, 2009]) results in the ability of the lower type senders to separate themselves through an inflated language. On the contrary, I show that introducing quilt aversion cost of lying precludes the existence of equilibria with separating intervals of types. Instead, under the assumptions of uniform state distribution and quadratic payoff functions, there exist partition equilibria like in CS.

Assume that the Sender has a persistent positive preference bias, i.e., the Sender wants a higher action than the Receiver, given any state of the world. I demonstrate that as in CS model, in any partition equilibrium, the higher type senders transmit less information than the lower type senders: the intervals in the equilibrium partition expand as the Sender's type increases. Increasing the level of guilt aversion (or decreasing the preference divergence), while holding the number of partition elements fixed, does not change this pattern, but makes the intervals more balanced. Further, an increase in the cost intensity (or a decrease in the preference bias) allows for a greater number of intervals in the equilibrium partition. Hence, guilt aversion considerations can explain the over-communication result documented in [Cai and Wang, 2006]. Finally, I show that both, the Sender and the Receiver, prefer an equilibrium with more intervals in the partition; and if players coordinate on the number of steps in the partition, the Receiver prefers a more guilt averse (and less biased) Sender.

The chapter is organized as follows. The model and the solution concept are presented in Section 4.2. Section 4.3 provides the general result on the impossibility of separation, and under the assumptions of uniform state distribution and quadratic payoff functions, characterizes partition equilibria, and discusses welfare implications. Section 4.4 extends the characterization of partition equilibria to more general cases. Section 4.5 concludes the chapter. Appendix 4.6 contains all the proofs.

4.1.1 Related literature

The fact that people care about each others' payoffs and suffer from psychological costs of lying has a lot of experimental evidence. For example, consider a setting of cheap talk, i.e., lying does not impose any direct cost on the Sender. A purely self-interested Sender would deceive the Receiver whenever lying generates a greater payoff for the Sender—even when the gain is marginal and the loss to the Receiver is significant. However, in experiments people do not behave like that. [Gneezy, 2005] shows that in the cheap talk setting, people care nor only about the gains from lying, but also about the extent to which lying harms the other party. In particular, if lying is associated with a marginal gain in the Sender's payoff and a significant loss to the Receiver, then the average person prefers not to lie. However, [Hurkens and Kartik, 2009] show that results of [Gneezy, 2005] do not contradict the simple hypothesis that some people are "extremely fair" and never lie, while others lie as soon as it is (at least marginally) beneficial to them. [Andreoni and Bernheim, 2009] consider a fairness hypothesis with the additional assumption that people "like to be perceived as fair". In addition, people appear to be sensitive to the size of the lie. [Lundquist *et al.*, 2009] present an experimental evidence that individuals have an aversion towards lying and this aversion increases with the size of the lie (as well as with the strength of the promise).

While the existence of psychological costs of lying is quiet well established, the precise nature of these costs is still under question. That is, whether people do not like lying per se (literal cost of lying), or they do not like letting down the others' expectations (guilt aversion), or whether they suffer from shame. [Tadelis, 2011] discusses the difference between "guilt" and "shame" and designs an experiment to distinguish between them. However, in many settings including the one considered in this chapter, shame and guilt are indistinguishable. Therefore, to avoid ambiguity, I will utilize only the term of "guilt aversion". Regarding guilt aversion and the preference to not to lie per se, the experimental evidence is not sufficient: there are papers slightly favoring one explanation or the other, but I am not aware of any results that clearly distinguish between the two. For example, [Charness and Dufwenberg, 2006] and [Dufwenberg and Gneezy, 2000] present an evidence that individuals' behavior can be explained by guilt aversion. Somewhat related is the case of promise-giving. In this setting, the analogue of the "preference to not to lie per se" is the individual's "preference to keep their word per se"; while "guilt aversion" corresponds to the "disutility of letting down others' payoff expectations" by not keeping their promise. [Vanberg, 2008] attempts to distinguish between the two reasons for promise keeping, and finds that the experimental data are better explained by the "preference to keep their word per se".

Also related to this chapter is the experimental literature studying the extent of communication in CS setting. [Cai and Wang, 2006] present an over-communication result: messages are more informative and receivers rely more on the senders' messages relative to the most informative equilibrium predicted by CS model. There are several reasons for this outcome including bounded rationality and existence of "fair" individuals, who have preferences for truth-telling, while others follow only material incentives ([Sánchez Pagés and Vorsatz, 2007]). This chapter gives another possible explanation to the over-communication result—guilt aversion considerations.

4.2 Model

The model is a modification of CS cheap-talk model. As in CS, there are two players: a Sender (S) and a Receiver (R). First, the Sender observes his private information about the state of the world (also referred to as the type of the Sender), $t \in [0,1]$, which is drawn from a distribution $F(t)$ with density $f(t) > 0$ for all $t \in [0,1]$. Next, the Sender sends a message $m \in M$ to the Receiver. The set M is assumed to be sufficiently large; no other restrictions on the message interpretation are imposed. Finally, the Receiver takes an action $a \in \mathbb{R}$ and the payoffs are realized.

The payoff of the Receiver is $u^R(a,t)$, where $u_1^R(a,t) = 0$ for some a, $u_{11}^R(a,t) < 0$, $u_{12}^R(a,t) > 0$.[2] These assumptions ensure that for any state of the world t there exist a unique action $a^R(t)$ that maximizes the payoff of the Receiver, and this action increases with t. Indeed, $\frac{da^R(t)}{dt} = -\frac{u_{12}^R}{u_{11}^R} > 0$.

The utility of the Sender is $u^S(a,t,b) - k \cdot c(x,e)$, where b represents the preference bias, $c(x,e)$ is a cost function, $k \geq 0$ parameterizes its intensity, $x = u^R(a,t)$ is the realized payoff of R, and $e = E_\beta(u^R(a,t))$ is the expected payoff of R. Assumptions about $u^S(a,t,b)$ are the same as in the case of the Receiver: $u_1^S(a,t,b) = 0$ for some a, $u_{11}^S(a,t,b) < 0$, $u_{12}^S(a,t,b) > 0$. Thus, there is a unique action $a^S(t,b)$, that maximizes $u^S(a,t,b)$ for any given state t, and this action is increasing in t. I assume that there is a conflict of interests such that $a^S(t,b) > a^R(t)$ for all $t \in [0,1]$. That is, ignoring the costs, the Sender prefers a higher action than the Receiver.[3]

Guilt aversion. The cost function, $c(x,e)$, depends on the actual payoff, $x = u^R(a,t)$, and the expected payoff of the Receiver who has belief β about the state, $e = E_\beta(u^R(a,t))$. I make the following assumptions about the cost function:

(i) For all $x \geq e$: $c(x,e) = 0$.

(ii) For all $e > x$: $c_1(x,e) < 0 < c_{11}(x,e)$ and $c_{12}(x,e) < 0$.

(iii) For all e: $c_1(e,e) = c_2(e,e) = 0$.

Such formulation of the cost function posits that given some payoff expectation for R, say e, the Sender suffers a guilt disutility if and only if S induces an action that yields R a lower payoff than e (condition (i)). Moreover, the disutility is increasing and convex in how much R is let down (condition (ii)). The negative cross-partial derivative means that the marginal increase in the cost from hurting the Receiver by a little bit is higher the more S already lets R down. Finally, condition (iii) says that small changes in either R's true payoff of expected payoff starting at a point where his expectation is fulfilled has negligible effect on S's guilt.

Solution concept. I study the model using the concept of pure strategies Perfect Bayesian Equilibrium (PBE). The Sender's pure strategy is a function $\mu : [0,1] \to M$, i.e., S of type t sends

[2] The subscripts denote the corresponding derivatives.

[3] Note that this setup coincides with CS model when the cost intensity $k = 0$.

a message $\mu(t)$ to R. The Receiver's posterior belief about the state after getting a message m is described by a probability distribution function $\beta(t|m)$, which will be referred to as $\beta(m)$ when it doesn't cause any confusion. The Receiver's pure strategy is a function $\alpha : M \to \mathbb{R}$, i.e., R takes the action $\alpha(m)$ after receiving the message m.

Definition. A triple (α, β, μ) is an equilibrium if and only if the following conditions are satisfied

(i) For any $m \in M$, R chooses an action $\alpha(m)$ to maximize his expected utility:

$$\alpha(m) \in \arg\max_{a} \mathbb{E}_{\beta(m)} \left[u^R(a,t)\right].$$

(ii) Given the action strategy $\alpha(m)$ and the belief $\beta(m)$, for any $t \in [0,1]$, S chooses a message $\mu(t)$ to maximize his expected utility:

$$\mu(t) \in \arg\max_{m} \left[u^S(\alpha(m),t,b) - kc\left(u^R(\alpha(m),t),e(m)\right)\right], \tag{4.1}$$

where $e(m) := \mathbb{E}_{\beta(m)} \left[u^R(\alpha(m),t)\right]$.

(iii) Belief $\beta(t|m)$ satisfies Bayes rule whenever possible.

Importantly, R's payoff expectation given the message m, $e(m)$, is computed using the equilibrium belief $\beta(m)$ and R's optimal action $\alpha(m)$. In general, $e(m)$ is different for different messages. Further, when S chooses amongst messages, he cares not only about his direct payoff from the change in R's action (captured by $\alpha(m)$ entering the first argument of u^S in (4.1)), but he also takes into account *both* the different expectation it causes for R (captured by $e(m)$ entering the second argument of c) *and* the different level of the actual payoff for R (captured by $\alpha(m)$ entering indirectly the first argument of c).

4.3 Analysis

In this section, I first demonstrate that in a general setting no equilibrium features separation of a non-zero interval of types. Instead, under the assumptions of uniform types distribution and quadratic payoff and cost functions, I show that there exist partition equilibria as in CS.

Finally, I study welfare implications of changing the number of intervals in the equilibrium partition, decreasing preference divergence and increasing guilt aversion intensity.

4.3.1 No separation in equilibrium

As in CS framework, there always exists a completely uninformative (babbling) equilibrium. Regarding other equilibria, the following result asserts that independently of the cost intensity k there can be no separating non-zero interval of types. This contrasts [Kartik, 2009] who got separating intervals (in the bounded state space) and [Kartik *et al.*, 2007] who got full separation (in the unbounded state space) for any, even arbitrary small lying cost intensity.

Theorem 18. *Independently of $k > 0$, there is no equilibrium, in which some non-zero interval of types separate themselves.*

Proof. See appendix 4.6. □

The intuition behind this statement is the following: suppose there is a non-zero interval of separating types. Then some interior type can deviate and report to be a slightly higher type, which results in a first-order gain from inducing a higher action and a second-order loss from guilt aversion of letting down R's payoff expectation.[4]

4.3.2 Partition equilibria

From now on, I let the state of the world t be uniformly distributed on $[0, 1]$ and assume quadratic functional forms for the payoffs. Specifically, R's payoff is a quadratic-loss utility function that depends solely on the difference between the actual action a and the state t:

$$u^R(a, t) = -(a - t)^2.$$

Next, S's utility is $u^S(a, t, b) - kc(\varphi(a - t), e)$, where

$$u^S(a, t, b) = -(a - t - b)^2,\ b > 0,\ \text{and}\ c(x, e) = (\max\{0, e - x\})^2.$$

[4] Note that this result might not hold if $c_2(e, e) > 0$ and $u_2^R(a^R(t), t) > 0$. For explanation, see the proof of Theorem 18.

In this case, the actions that maximize $u^R(a,t)$ and $u^S(a,t,b)$ are $a^R(t) = t$ and $a^S(t,b) = t + b$, respectively. Condition $b > 0$ insures that there is a persistent conflict between the players, $a^S(t,b) = a^R(t) + b$ for all $t \in [0,1]$.

Given the beliefs of the Receiver, $\beta(m)$, and his action strategy under these beliefs, I can say that sending some message m *induces* a pair (a,e), where $a = a(m)$ and $e = E_{\beta(m)}(u^R(a,t))$. Thus, when communicating, the Sender chooses which pair (a,e) to induce, given $\beta(m)$.

Definition. A PBE with a set of different pairs $\{(a_i,e_i)\}_{i=1}^N$ is called a *partition* PBE with the partition $0 = t_0 < t_1 < ... < t_N = 1$, if the types in (t_{i-1},t_i) induce the pair (a_i,e_i), $i = 1,...,N$.

In CS model (the guilt aversion intensity k is 0), any equilibrium is essentially equivalent to some partition PBE, with a finite number of intervals (N) and induced actions $a_1 < a_2 < ... < a_N$. Because $b > 0$, the partition equilibrium features the property that the intervals expand for higher types t; the precise formula for the interval lengths is: $\Delta_{i+1} = \Delta_i + 4b$. Thus, less information is revealed when the type of the Sender is higher, which implies lower expected payoff for the Receiver: $e_1 > e_2 > ... > e_N$.

An interesting question to ask is whether this pattern for actions and expectations can be reversed, if the guilt aversion intensity k is large enough? The answer is: No, not in a partition PBE. The following theorem demonstrates this point and provides characterization of partition equilibria.

Theorem 19. *There exists a positive integer $N(b,k)$, such that for any N, $1 \leq N \leq N(b,k)$, there exists exactly one partition PBE (α,β,μ), in which, effectively, N messages $m \in \{1,\dots,N\}$, are used and N pairs (a_i,e_i), $i = 1,\dots,N$, are induced. The belief $\beta(t|m)$ is is uniform on $[t_{i-1},t_i]$ if $m = i$, the interval partition $0 = t_0 < t_1 < ... < t_N = 1$, is determined from*

$$\begin{aligned} & -(a_i - t_i - b)^2 - k(\max\{0, e_i + (a_i - t_i)^2\})^2 \\ = \quad & -(a_{i+1} - t_i - b)^2 - k(\max\{0, e_{i+1} + (a_{i+1} - t_i)^2\})^2, \ i = 1,...,N-1. \end{aligned} \tag{4.2}$$

The Sender's message strategy is

$$\mu(t) = i, \ \textit{if } t \in (t_{i-1},t_i), \quad \textit{and } \mu(t_i) \in \{i, i+1\}, \tag{4.3}$$

the Receiver's actions are

$$a_i = \frac{t_{i-1} + t_i}{2},\ i = 1, ..., N, \tag{4.4}$$

and the Receiver's expected payoffs are

$$e_i = -\frac{(t_i - t_{i-1})^2}{12},\ i = 1, ..., N. \tag{4.5}$$

Moreover, $t_{i+1} - t_i = \Delta_{i+1} > \Delta_i = t_i - t_{i-1}$ for every i. Further, there is no partition equilibrium with the number of intervals greater than $N(b,k)$.

Proof. See appendix 4.6. □

As is shown in the proof, the difference in the neighboring intervals when $k > 0$ is smaller compared to the CS setting ($k = 0$): $\Delta_{i+1} < \Delta_i + 4b$, which possibly explains the over-communication result of [Cai and Wang, 2006]. The following Lemma asserts that an increase in the intensity of guilt aversion has similar effects on the equilibrium partition as a decrease in the preference divergence.

Lemma 20. *If the cost intensity k increases (or bias b decreases) then*

(i) Holding fixed the number of the intervals in the equilibrium partition, N, the intervals become more even, i.e., $(\Delta_{i+1} - \Delta_i)$ decreases. If $k \to \infty$ (or $b \to 0$) then $\Delta_{i+1} \to \Delta_i$.

(ii) The maximal number of intervals in equilibrium partition, $N(b,k)$, increases.

Proof. See appendix 4.6. □

Illustrative example. To illustrate the effects of the cost intensity k and the preference bias b, I consider 2-interval partition PBE, which exists if $b < \frac{1+k/9}{4}$. In this equilibrium, the types $[0, t_1]$ induce one pair, (a_1, e_1), and the types $(t_1, 1]$ induce another pair, (a_2, e_2).

First, note that the largest b for which the 2-interval partition PBE exists is $\bar{b} = \frac{1}{4}\left(1 + \frac{k}{9}\right)$ (which is greater than in the CS model, $\frac{1}{4}$). The greater the guilt aversion intensity k is, the higher $\bar{b}$ is; that is, guilt aversion "makes up" for the preference divergence.

Second, Figure 4.1 illustrates that, given the bias b, the threshold type t_1 increases in the guilt aversion intensity k, making the two intervals more even. Finally, holding the cost intensity k

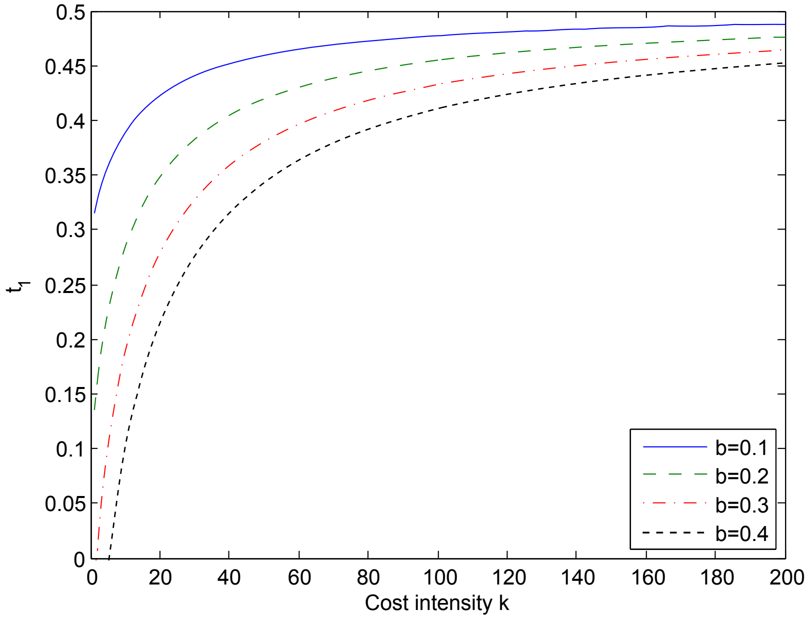


Figure 4.1: Threshold t_1 in 2-interval partition PBE as a function of k.

fixed, higher bias b results in lower value of t_1, making the interval less even and the equilibrium less balanced.

Welfare implications. Similar to CS, I show that both, the Sender and the Receiver, ex-ante prefer greater information transmission, i.e., equilibrium partitions with more steps.

Theorem 20. *Holding the preference bias b and the cost intensity k fixed, R and S ex-ante prefer equilibrium partitions with more intervals.*

Proof. See appendix 4.6. □

Remark. It is worth noting, that the greater number of intervals in the equilibrium partition increases S's expected utility through two channels: first, by increasing the expectation of the direct payoff, $u^S(a, t, b)$, and second, by decreasing the expected disutility from guilt aversion, $c(x, e)$.[5]

CS show that, holding the number of steps in the partition equilibrium fixed, R's expected utility increases when agents' preferences become more similar. This happens, because as b decreases, the

[5] This point is demonstrated in the proof of Theorem 20.

intervals in the equilibrium partition become more balanced. As Lemma 20 demonstrated, an increase in k has similar effects as a decrease in b. Thus, it is natural to expect that R's expected utility rises not only when b decreases, but also when S's cost intensity k increases. The following theorem formalizes this point.

Theorem 21. *Holding the number of intervals N in the partition equilibrium fixed, R would ex-ante prefer an equilibrium partition associated with lower b and higher k.*

Proof. See appendix 4.6. □

4.4 Extensions

In this section, I maintain the assumption that the state of the world t is uniformly distributed on $[0, 1]$, and extend the characterization of partition PBE to a more general case of payoff and cost functions. In particular, I assume that R's payoff depends solely on the difference between the actual action a and the state t in a symmetric way: $u^R(a, t) = \varphi(a - t)$, where $\varphi(x) = \varphi(-x)$, $\varphi''(x) < 0 \ \forall x \in \mathbb{R}$. These conditions imply that $\varphi(x)$ achieves its maximum at 0 and $\varphi'(0) = 0$, i.e., $a^R(t) = t$. Next, S's payoff is $u^S(a, t, b) - kc(\varphi(a - t), e)$, where I maintain the general assumptions about $u^S(a, t, b)$ and continue to assume that the Sender's preferred action $a^S(t, b)$ is strictly greater than that of the Receiver, $a^R(t) = t$.[6] Regarding the cost function, I assume that it depends only on the difference between the actual and expected utilities of the Receiver. That is, $c(\varphi(a - t), e) = \psi(\varphi(a - t) - e) \geq 0$, where $\psi''(x) > 0$ and $\psi'(x) < 0 \ \forall x < 0$; $\psi(x) = 0 \ \forall x > 0$; $\psi(0) = 0$ and $\psi'(0) = 0$.

As in quadratic case, I find that in any partition equilibrium higher types t induce greater actions ($a_1 < ... < a_N$) and lower expectations($e_1 > ... > e_N$), and correspond to greater partition intervals ($\Delta_1 < ... < \Delta_N$). The following theorem provides characterization of partition equilibria.[7]

Theorem 22. *Any partition PBE with N intervals satisfies the following conditions. The belief*

[6] Note that, while R's payoff is assumed to depend only on the difference between the action and the state, S's payoff might also depend on the state.

[7] Here I don't prove the existence of partition equilibria.

$\beta(t|m)$ is uniform on $[t_{i-1}, t_i]$, if $m = i$, the interval partition $\{t_i\}$, $i = 1, \dots, N$, solves

$$u^S(a_i, t_i, b) - k\psi(\varphi(a_i - t_i) - e_i) = u^S(a_{i+1}, t_i, b) - k\psi(\varphi(a_{i+1} - t_i) - e_{i+1}), \tag{4.6}$$

where $t_0 = 0$ and $t_N = 1$. The Sender's message strategy is

$$\mu(t) = i, \text{ if } t \in (t_{i-1}, t_i), \ \mu(t_i) \in \{i, i+1\}, \tag{4.7}$$

the Receiver's actions are

$$a_i = \frac{t_{i-1} + t_i}{2}, \ i \in \{1, \dots, N\}, \tag{4.8}$$

and the Receiver's expected payoffs are

$$e_i = \frac{1}{\Delta_i} \int_0^{\Delta_i} \varphi\left(\frac{\Delta_i}{2} - s\right) ds, \ i \in \{1, \dots, N\}, \tag{4.9}$$

where $\Delta_i = t_i - t_{i-1}$ and $\Delta_i < \Delta_{i+1}$ for all $i = 1, ..., N-1$.

Proof. See appendix 4.6. □

4.5 Conclusion

In this chapter, I study a model of strategic communication between an informed Sender and an uninformed Receiver, who needs to take an action affecting both players. The Sender's preferences are upward biased relative to those of the Receiver, and the Sender is assumed to be guilt averse to letting down the Receiver's payoff expectations. I show that no separating or partially separating equilibrium exists for general state distribution and payoff specifications.

Under the assumptions of uniform state distribution and quadratic utilities, I demonstrate that there exist partition equilibria, in which the Sender effectively reports only the interval where the state lies. In any equilibrium partition, the intervals expand as the Sender's type gets higher. Regarding comparative statics, I show that an increase in the guilt aversion intensity has a similar effects to a decrease in the preference bias. Namely, holding the number of elements in the equilibrium partition fixed, a higher level of guilt aversion leads to more balanced intervals, and greater guilt aversion intensity allows for more elements in the equilibrium partition. Regarding the welfare implications: before the Sender observes the state of the world, both, the Sender and the Receiver prefer equilibrium with more elements in the partition. In addition, if players coordinate on the number of steps in the partition, the Receiver prefers a more guilt averse (and a less biased) Sender.

For future analysis, it would be interesting to study other types of equilibria (apart from partition), or prove that they do not exist. Further, an intuitive extension of the model would be to incorporate some sort of disappointment aversion on part of the Receiver. Another modification for potential study would be to consider less smooth guilt aversion costs, and check whether this allows for separation in equilibrium.

4.6 Appendix: Proofs

Proof of Theorem 18. Suppose the opposite: there is an equilibrium, in which the types in the interval (t_1, t_2) perfectly reveal themselves. That means that for any $t \in (t_1, t_2)$ the posterior $\beta(\mu(t))$ is a degenerate point-mass distribution on the type t. Therefore, for any $t \in (t_1, t_2)$ the action is $\alpha(\mu(t)) = a^R(t)$.

Now consider some type $t \in (t_1, t_2)$. In what follows I show that for sufficiently small $\varepsilon > 0$ this type t sender would prefer to deviate and send the message $\mu(t+\varepsilon)$. Indeed, consider $\varepsilon > 0$ such that $t + \varepsilon < t_2$. Upon receiving $\mu(t+\varepsilon)$, R thinks that the state is $t+\varepsilon$ for sure, thus, he picks the action $a^R(t+\varepsilon)$ and expects to receive the payoff of $u^R\left(a^R\left(t+\varepsilon\right), t+\varepsilon\right)$. Hence, the expected utility of the type t sender from deviating to $\mu(t+\varepsilon)$ is

$$u^S\left(a^R\left(t+\varepsilon\right), t, b\right) - kc\left(u^R\left(a^R\left(t+\varepsilon\right), t\right), u^R\left(a^R\left(t+\varepsilon\right), t+\varepsilon\right)\right).$$

Differentiating with respect to ε and evaluating at $\varepsilon = 0$,

$$\begin{aligned}
& u_1^S\left(a^R\left(t\right), t, b\right) a_1^R(t) - kc_1\left(u^R\left(a^R\left(t\right), t\right), u^R\left(a^R\left(t\right), t\right)\right) u_1^R\left(a^R\left(t\right), t\right) a_1^R(t) \\
- \;& kc_2\left(u^R\left(a^R\left(t\right), t\right), u^R\left(a^R\left(t\right), t\right)\right)\left[u_1^R\left(a^R\left(t\right), t\right) a_1^R(t) + u_2^R\left(a^R\left(t\right), t\right)\right] \\
= \;& u_1^S\left(a^R\left(t\right), t, b\right) a_1^R(t) - kc_2\left(u^R\left(a^R\left(t\right), t\right), u^R\left(a^R\left(t\right), t\right)\right) u_2^R\left(a^R\left(t\right), t\right) \\
= \;& u_1^S\left(a^R\left(t\right), t, b\right) a_1^R(t) > 0,
\end{aligned}$$

where $a_1^R(t) = \frac{da^R}{dt}(t)$. Here the first equality uses the fact that $u_1^R\left(a^R(t), t\right) = 0$ (by the definition of a^R), the second equality uses the assumption that $c_2(e,e) = 0$ for any e, and the last inequality holds because $a^S(t,b) > a^R(t)$ and $a_1^R(t) > 0$.

Even if $c_2(e,e) \neq 0$ for all e, the result would still go through if $u_2^R\left(a^R(t), t\right) = 0$, as is the case with quadratic preferences. However, if $c_2(e,e) > 0$ and $u_2^R\left(a^R(t), t\right) > 0$, then the derivative might be negative and the type t sender would prefer to report truthfully. **QED.**

Proof of Theorem 19. Step 1: Equilibrium conditions. First, I show that the belief $\beta(t|m)$ and equations (4.2), (4.3), (4.4) and (4.5) form equilibrium conditions. If the types in the interval $[t_{i-1}, t_i]$ (and only they) send the same message $m = i$, then the posterior belief $\beta(t|m)$ is uniform on $[t_{i-1}, t_i]$, and the optimal action for the Receiver is

$$a_i = \mathbb{E}_\beta(t) = \int_{t_{i-1}}^{t_i} t dt = \frac{t_{i-1} + t_i}{2}.$$

In this case, R's expected payoff is

$$e_i - \int_{t_{i-1}}^{t_i} \left(\frac{t_{i-1} + t_i}{2} - t\right)^2 \frac{1}{t_i - t_{i-1}} dt = -\frac{(t_i - t_{i-1})^2}{12}.$$

Thus, indeed, (4.4) gives the Receiver's best response given the belief $\beta(m)$ and the Sender's strategy.

Now I will show that, given the belief $\beta(m)$ and R's action strategy presented in (4.4), the best response of S is described by (4.3). Note that upon sending the message i and inducing (a_i, e_i), the types in $[a_i - \sqrt{-e_i}, a_i + \sqrt{-e_i}] \subset (t_{i-1}, t_i)$ won't suffer from guilt aversion, and the types in $(t_{i-1}, a_i - \sqrt{-e_i}) \cup (a_i + \sqrt{-e_i}, t_i)$ - will. In particular, the types t_{i-1} and t_i always suffer from some costs when inducing (a_i, e_i).

Consider 2 pairs, $(a_i, e_i) = (\frac{t_{i-1}+t_i}{2}, -\frac{(t_i - t_{i-1})^2}{12})$ and $(a_{i+1}, e_{i+1}) = (\frac{t_i + t_{i+1}}{2}, -\frac{(t_{i+1} - t_i)^2}{12})$, $0 \leq t_{i-1} < t_i < t_{i+1} \leq 1$, which are supposed to be induced by the neighboring intervals (in equilibrium), (t_{i-1}, t_i) and (t_i, t_{i+1}), respectively. Type t_i is indifferent between the two pairs, (a_i, e_i) and

(a_{i+1}, e_{i+1}), if and only if

$$\begin{aligned} & -\left(\frac{t_{i-1}+t_i}{2}-t_i-b\right)^2-k\left(-\frac{(t_i-t_{i-1})^2}{12}+\left(\frac{t_{i-1}+t_i}{2}-t_i\right)^2\right)^2 \\ = & -\left(\frac{t_i+t_{i+1}}{2}-t_i-b\right)^2-k\left(-\frac{(t_{i+1}-t_i)^2}{12}+\left(\frac{t_i+t_{i+1}}{2}-t_i\right)^2\right)^2 . \end{aligned}$$

Using the notation $t_i - t_{i-1} = \Delta_i$ and $t_{i+1} - t_i = \Delta_{i+1}$, this equality becomes:

$$-\left(-\frac{\Delta_i}{2}-b\right)^2-k\left(\frac{\Delta_i^2}{6}\right)^2=-\left(\frac{\Delta_{i+1}}{2}-b\right)^2-k\left(\frac{\Delta_{i+1}^2}{6}\right)^2$$

$$\left(\frac{\Delta_{i+1}}{2}-b\right)^2-\left(-\frac{\Delta_i}{2}-b\right)^2=\frac{k}{36}\left(\Delta_i^4-\Delta_{i+1}^4\right),$$

$$\frac{1}{4}\left(\Delta_{i+1}-\Delta_i-4b\right)\left(\Delta_{i+1}+\Delta_i\right)=\frac{k}{36}(\Delta_i-\Delta_{i+1})(\Delta_i+\Delta_{i+1})(\Delta_i^2+\Delta_{i+1}^2),$$

$$\underbrace{\Delta_{i+1}-\Delta_i-4b}_{LHS}=\underbrace{\frac{k}{9}(\Delta_i-\Delta_{i+1})(\Delta_i^2+\Delta_{i+1}^2)}_{RHS}. \tag{4.10}$$

Clearly, LHS is increasing in Δ_{i+1}. RHS is decreasing in Δ_{i+1} because

$$\frac{k}{9}\left(2\Delta_i\Delta_{i+1}-\Delta_i^2-3\Delta_{i+1}^2\right)=\frac{k}{9}\left(-(\Delta_{i+1}-\Delta_i)^2-2\Delta_{i+1}^2\right)<0.$$

If $\Delta_{i+1} = 0$, $LHS = -\Delta_i - 4b < 0$, $RHS = \frac{k}{9}\Delta_i^3 > 0$. If Δ_{i+1} increases, LHS will eventually become positive, and RHS will become negative. Thus, there always exists a unique solution for Δ_{i+1} in terms of Δ_i: $\Delta_{i+1} = \Delta(\Delta_i, b, k)$. Differentiating (4.10) w.r.t. Δ_i yields:

$$\frac{d\Delta_{i+1}}{d\Delta_i}\left[1+\frac{k}{9}\left((\Delta_{i+1}-\Delta_i)^2+2\Delta_{i+1}^2\right)\right]=1+\frac{k}{9}\left((\Delta_{i+1}-\Delta_i)^2+2\Delta_i^2\right)>0.$$

Thus, $\Delta(\Delta_i, b, k)$ increases in Δ_i.

Note that Δ_{i+1} can not be smaller than Δ_i, because in that case $LHS < 0$ and $RHS \geq 0$, meaning that type t_i would not be indifferent and would strictly prefer inducing the pair (a_{i+1}, e_{i+1}). Thus, it must be that $\Delta_{i+1} > \Delta_i$. In this case, $RHS < 0$, and for LHS to be negative it must be that $\Delta_{i+1} < \Delta_i + 4b$. I.e., the length of Δ_{i+1} is smaller than it would be in the standard CS framework ($k = 0$). It can be easily seen that $\Delta(\Delta_i, b, k)$ decreases in k and increases in $b > 0$; and $\Delta_{i+1} \to \Delta_i$, as $k \to \infty$ (or $b \to 0$).

The condition (4.10) states that the type t_i is indifferent between (a_i, e_i) and (a_{i+1}, e_{i+1}). Now I will show that the type $s \neq t_i$ prefers (a_i, e_i) to (a_{i+1}, e_{i+1}) if and only if $s < t_i$. The condition that the pair (a_i, e_i) is better than (a_{i+1}, e_{i+1}) for the type s is

$$\underbrace{(a_{i+1} - s - b)^2 - (a_i - s - b)^2}_{LHS(s)} > \underbrace{k\left[(\max\{0, e_i + (a_i - s)^2\})^2 - (\max\{0, e_{i+1} + (a_{i+1} - s)^2\})^2\right]}_{RHS(s)}$$

Because $LHS(t_i) = RHS(t_i)$, it is sufficient to show that $LHS(s)$ is strictly decreasing in s and $RHS(s)$ is strictly increasing in s.

$LHS(s)$ can be rewritten as $LHS(s) = (a_{i+1} + a_i - 2s - 2b)(a_{i+1} - a_i)$, hence, LHS is strictly decreasing in s.

To show that $RHS'(s) > 0$, I separately consider cases 5 intervals to which [0,1] interval is split by the points $0 < a_i - \sqrt{-e_i} < a_i + \sqrt{-e_i} < a_{i+1} - \sqrt{-e_{i+1}} < a_{i+1} + \sqrt{-e_{i+1}} < 1$.

Case 1. $s \in [a_i - \sqrt{-e_i}, a_i + \sqrt{-e_i}]$. For these types, The cost from inducing the pair (a_i, e_i) is 0, hence, $RHS'(s) = 4(e_{i+1} + (a_{i+1} - s)^2)(a_{i+1} - s) > 0$ because $a_{i+1} > a_i + \sqrt{-e_i} \geq s$.

Case 2. $s \in [a_{i+1} - \sqrt{-e_{i+1}}, a_{i+1} + \sqrt{-e_{i+1}}]$. For these types, the cost from inducing the pair (a_{i+1}, e_{i+1}) is 0, hence, $RHS'(s) = -4(e_i + (a_i - s)^2)(a_i - s) > 0$ because $a_i < a_{i+1} - \sqrt{-e_{i+1}} \leq s$.

Case 3. $s \in (a_i + \sqrt{-e_i}, a_{i+1} - \sqrt{-e_{i+1}})$. For these types, the costs from inducing both pairs are non-zero, hence, $RHS'(s) = -4(e_i + (a_i - s)^2)(a_i - s) + 4(e_{i+1} + (a_{i+1} - s)^2)(a_{i+1} - s) > 0$ because $a_i < s < a_{i+1}$.

Case 4. $s \in [0, a_i - \sqrt{-e_i})$. For these types, the costs from inducing both pairs are non-zero, hence, $RHS'(s) = -4(e_i + (a_i - s)^2)(a_i - s) + 4(e_{i+1} + (a_{i+1} - s)^2)(a_{i+1} - s)$. Note that $a_{i+1} - s > a_i - s > 0$. To prove that $RHS'(s) > 0$, it is sufficient to show that $e_{i+1} + (a_{i+1} - s)^2 > e_i + (a_i - s)^2 > 0$, i.e.,

$$(a_{i+1} - s)^2 - (a_i - s)^2 > e_i - e_{i+1}.$$

The left-hand side can be expressed as $(a_{i+1} - a_i)(a_{i+1} + a_i - 2s)$. Thus, it decreases in s and

$$(a_{i+1} - s)^2 - (a_i - s)^2 > (a_{i+1} - s)^2 - (a_i - a_i + \sqrt{-e_i})^2 = (a_{i+1} - s)^2 + e_i > -e_{i+1} + e_i.$$

Case 5. $s \in [a_{i+1} + \sqrt{-e_{i+1}}, 1)$. For these types, the costs from inducing both pairs are non-zero, hence, $RHS'(s) = -4(e_i + (a_i - s)^2)(a_i - s) + 4(e_{i+1} + (a_{i+1} - s)^2)(a_{i+1} - s)$. Note that $a_i - s < a_{i+1} - s < 0$. To prove that $RHS'(s) > 0$, it is sufficient to show that $0 < e_{i+1} + (a_{i+1} - s)^2 < e_i + (a_i - s)^2$, i.e.,

$$(a_{i+1} - s)^2 - (a_i - s)^2 < e_i - e_{i+1}.$$

The left-hand side can be expressed as $(a_{i+1} - a_i)(a_{i+1} + a_i - 2s)$. Thus, it decreases in s and

$$(a_{i+1} - s)^2 - (a_i - s)^2 < (a_{i+1} - a_{i+1} + \sqrt{-e_{i+1}})^2 - (a_i - s)^2 = -e_{i+1} - (a_i - s)^2 < -e_{i+1} + e_i.$$

This proves that the best response of S takes the form described in Theorem 19.

Step 2: Finiteness of $N(b, k)$ and existence of partition equilibria. To complete the proof, I will show that the maximum number of intervals in the partition PBE, $N(b, k)$, is finite and that for any N, $1 \le N \le N(b, k)$, there exists exactly one partition $\{t_i\}$ of $[0, 1]$ in N intervals, satisfying condition (4.10).

Because $\Delta_{i+1} = \Delta(\Delta_i, b, k)$ strictly increases with Δ_i and the length of $\Delta_i > 0$, then $\Delta(\Delta_i, b, k) \ge \Delta(0, b, k) > 0$. Hence, the number of intervals in the partition must not exceed $(1/(\Delta(0, b, k)) + 1)$. Next, following CS analysis, introduce a function

$$M(t, b, k) = \max\{i : \text{ there exists a partial partition } 0 < t < t_2 < \ldots < t_i \le 1 \text{ satisfying (4.10)}\}.$$

For any $t \in [0, 1]$, $M(t, b, k) < 1/(\Delta(0, b, k)) + 1$. Thus $M(t, b, k)$ is finite, uniformly bounded and well-defined as a function of t. That means that there is $N(b, k) = \sup_{0 \le t \le 1} M(t, b, k) < 1/(\Delta(0, b, k)) + 1 < \infty$. Now I will argue that for every N, $1 \le N \le N(b, k)$ there exists exactly one partition of $[0, 1]$ into N intervals, satisfying (4.10). Let $t^{M(t,b,k)}$ be the partial partition with $M(t, b, k)$ points, such that $t_1^{M(t,b,k)} = t$ and the corresponding lengths satisfy (4.10). Solutions to (4.10) vary continuously w.r.t. initial conditions, that is why if the last point in the partial partition $t_{M(t,b,k)}^{M(t,b,k)}$ is less than 1, then $M(\cdot, b, k)$ is continuous and locally constant at t. Because equation (4.10) determines an increasing function $\Delta(\Delta_i, b, k)$ of Δ_i, $M(t, b, k)$ is a non-increasing function w.r.t. t and can change by at most 1 at discontinuity and $M(1, b, k) = 1$. Thus, $M(t, b, k)$ takes on all integer values from 1 to $N(b, k)$. If $M(t_1, b, k) = N$ and $M(t, b, k)$ is discontinuous at

$t = t_1$, then the partial partition $t^{Mt_1,b,k}$ is the full partition of $[0,1]$, i.e., $t_0 = 0$, $t_N = 1$, (4.10) is satisfied for every $i = 1, \dots, N-1$. Because $\Delta(\Delta_i, b, k)$ is the unique solution of (4.10) and strictly increases in Δ_i, for every N, $1 \leq N \leq N(b,k)$, there exists exactly one partition $\{t_i\}$ of $[0,1]$ in N intervals, satisfying condition (4.10). **QED.**

Proof of Lemma 20. From the equation (4.10) it can be easily seen that the function $\Delta_{i+1} = \Delta(\Delta_i, b, k)$ decreases in k and increases in b. Because $\Delta(\Delta_i, b, k)$ increases in Δ_i, greater k or smaller b lead to more even intervals, when holding fixed the number of the intervals in the partition equilibrium.

Consider $b_1 < b_2$. Because $\Delta(\Delta_i, b, k)$ increases in b, it is immediate that $M(t, b_1, k) \geq M(t, b_2, k)$ for any $t \in [0,1]$. Thus, $N(b_1, k) = \sup_{0 \leq t \leq 1} M(t, b_1, k) \geq \sup_{0 \leq t \leq 1} M(t, b_2, k) = N(b_2, k)$. Similarly, consider $k_1 < k_2$. Because $\Delta(\Delta_i, b, k)$ decreases in k, it is immediate that $M(t, b, k_1) \leq M(t, b, k_2)$ for any $t \in [0,1]$. Thus, $N(b, k_1) \leq N(b, k_2)$. **QED.**

Proof of Theorem 20. Call the sequence $\{t_0, \dots, t_N\}$ a *forward (backward) solution* to the indifference condition (4.2) if the condition holds for all $1 \leq i < N$ and $t_0 < t_1$ ($t_0 > t_1$). Because the function $\Delta_{i+1} = \Delta(\Delta_i, b, k)$ (solution of (4.10)) increases in Δ_i, the monotonicity conditions (M) and (M') of CS hold here:

(M) Given b and k, if $\hat{t}$ and $\tilde{t}$ are two forward solutions of (4.2) with $\hat{t}_0 = \tilde{t}_0$ and $\hat{t}_1 > \tilde{t}_1$, then $\hat{t}_i > \tilde{t}_i$ for all $i \geq 2$.

(M') Given b and k, if $\hat{t}$ and $\tilde{t}$ are two backward solutions of (4.2) with $\hat{t}_0 = \tilde{t}_0$ and $\hat{t}_1 > \tilde{t}_1$, then $\hat{t}_i > \tilde{t}_i$ for all $i \geq 2$.

Welfare of the Receiver. First, I prove the welfare statement for R. The proof has the same structure as the one in Section 5 of [Crawford and Sobel, 1982]. Consider two equilibrium partitions $t(N) = (t_0(N), ..., t_N(N))$ and $t(N+1) = (t_0(N+1), ..., t_{N+1}(N+1))$ with N and $N+1$ intervals, respectively, $N < N(b,k)$. As in CS, I will argue that there exists a continuous deformation of $t(N)$ into $t(N+1)$, along which the expected utility of R increases.

Denote by $t^x = (t^x_0, t^x_1, \dots, t^x_{N+1})$ a partition that satisfies the indifference condition (4.2) for $i = 2, \dots, N$ with $t^x_0 = 0$, $t^x_N = x$, and $t^x_{N+1} = 1$. If $x = t_{N-1}(N)$, then $t^x_1 = 0$; if $x = t_N(N+1)$, then $t^x = t(N+1)$. Note that the condition (M') implies that $t_{N-1}(N) < t_N(N+1)$. Moreover, the lengths of the intervals in the partition t^x satisfy $t_i^x - t_{i-1}^x = \Delta_i^x < \Delta_{i+1}^x = t_{i+1}^x - t_i^x$ for all $i = 1, ..., N$. This is evident for $i = 2, ..., N$ (because $\Delta_{i+1} = \Delta(\Delta_i, b, k)$ increases in Δ_i). To see that $\Delta_1^x < \Delta_2^x$, note that $t_1^x \leq t_1(N+1)$ (by the condition (M')) and $t_2^x - t_1^x \geq t_2(N+1) - t_1(N+1) > t_1(N+1)$ (because $\Delta_{i+1} = \Delta(\Delta_i, b, k)$ increases in Δ_i).

I will show that as x changes from $t_{N-1}(N)$ to $t_N(N+1)$, the expected payoff of R increases. The expected utility of R depending on x is

$$\mathbb{E}u^R(x) = \sum_{i=1}^{N+1} \int_{t_{i-1}^x}^{t_i^x} u^R(a_i^x, t)dt,$$

where $a_i^x = \frac{t_{i-1}^x + t_i^x}{2}$ is the optimal action, given that the state is in the interval (t_{i-1}^x, t_i^x). Differentiating with respect to x and using the Envelope Theorem:

$$\frac{d\mathbb{E}u^R(x)}{dx} = \sum_{i=1}^{N} \frac{dt_i^x}{dx} \left(u^R(a_i^x, t_i^x) - u^R(a_{i+1}^x, t_i^x) \right).$$

By the monotonicity condition (M'): $\frac{dt_i^x}{dx} > 0$, $i = 1, \ldots, N$. Next, for each $i = 1, \ldots, N$:

$$\begin{aligned} & u^R(a_i^x, t_i^x) - u^R(a_{i+1}^x, t_i^x) = -\left(a_i^x - t_i^x\right)^2 + \left(a_{i+1}^x - t_i^x\right)^2 \\ = \quad & -\left(\frac{t_{i-1}^x + t_i^x}{2} - t_i^x\right)^2 + \left(\frac{t_i^x + t_{i+1}^x}{2} - t_i^x\right)^2 = -\frac{1}{4}\left(\Delta_i^x\right)^2 + \frac{1}{4}\left(\Delta_{i+1}^x\right)^2 > 0. \end{aligned}$$

Here the last inequality evidently holds because $\Delta_{i+1}^x > \Delta_i^x$ for $i = 1, \ldots, N$ and $x \in [t_{N-1}(N), t_N(N+1)]$.

Welfare of the Sender. Second, I prove the welfare statement for S. Consider two equilibrium partitions: $t(N) = (0 = t_0(N), t_1(N), ..., t_N(N) = 1)$ and $t(N+1) = (0 = t_0(N+1), t_1(N+1), ..., t_{N+1}(N+1) = 1)$. Denote the lengths of the intervals in these partitions as $\{\Delta_i(N)\}_{i=1}^N$ and $\{\Delta_i(N+1)\}_{i=1}^{N+1}$, respectively. Let $\{a_i(N)\}_{i=1}^N$, where $a_i(N) = \frac{t_i(N) + t_{i-1}(N)}{2}$, be the set of equilibrium actions corresponding to $t(N)$, and $\{a_i(N+1)\}_{i=1}^{N+1}$, where $a_i(N+1) = \frac{t_i(N+1) + t_{i-1}(N+1)}{2}$, be the set of equilibrium actions corresponding to $t(N+1)$. I will show that S is ex-ante better off under the partition $t(N+1)$ than under $t(N)$.

The expected utility of S in $t(N)$ partition equilibrium is

$$\begin{aligned}
&\sum_{i=1}^{N}\int_{t_{i-1}(N)}^{t_i(N)}\left(-(a_i(N)-t-b)^2-kc(u^R(a_i(N),t),e_i(N))\right)dt\\
=\;&\sum_{i=1}^{N}\int_{t_{i-1}(N)}^{t_i(N)}\left(-(a_i(N)-t)^2+2b(a_i(N)-t)-b^2\right)dt-\sum_{i=1}^{N}\int_{t_{i-1}(N)}^{t_i(N)}kc(u^R(a_i(N),t),e_i(N))dt\\
=\;&\mathbb{E}u^R(t(N))-b^2-\sum_{i=1}^{N}\int_{t_{i-1}(N)}^{t_i(N)}kc(u^R(a_i(N),t),e_i(N))dt,
\end{aligned}$$

where $\mathbb{E}u^R(t(N))$ is R's expected payoff in $t(N)$ partition equilibrium. Similarly, the expected utility of S in $t(N+1)$ partition equilibrium is

$$\mathbb{E}u^R(t(N+1))-b^2-\sum_{i=1}^{N+1}\int_{t_{i-1}(N+1)}^{t_i(N+1)}kc(u^R(a_i(N+1),t),e_i(N+1))dt.$$

From the welfare result for the Receiver, I already know that $\mathbb{E}u^R(t(N))<\mathbb{E}u^R(t(N+1))$. Thus, it is sufficient to show that

$$\sum_{i=1}^{N}\int_{t_{i-1}(N)}^{t_i(N)}c(u^R(a_i(N),t),e_i(N))dt\geq\sum_{i=1}^{N+1}\int_{t_{i-1}(N+1)}^{t_i(N+1)}c(u^R(a_i(N+1),t),e_i(N+1))dt. \tag{4.11}$$

Recall that $e_i(N)=-\frac{(t_i(N)-t_{i-1}(N))^2}{12}=-\frac{\Delta_i(N)^2}{12}$ and that $c(u^R(a_i(N),t),e_i(N))$ is 0 for all $t\in(a_i(N)-\sqrt{-e_i(N)},a_i(N)+\sqrt{-e_i(N)})$. Now calculate the expected cost in the interval Δ_i:

$$\int_{t_{i-1}(N)}^{t_i(N)}c(u^R(a_i(N),t),e_i(N))dt=2\int_{a_i(N)+\sqrt{-e_i(N)}}^{t_i(N)}(e_i(N)-u^R(a_i(N),t))^2dt.$$

Express t in terms of a new variable s: $t=a_i(N)+\sqrt{-e_i(N)}+s$, where s changes from 0 to $\frac{\Delta_i(N)}{2}-\sqrt{-e_i(N)}=\Delta_i(N)\frac{\sqrt{3}-1}{2\sqrt{3}}$. Thus,

$$(e_i(N)-u^R(a_i(N),t))^2=\left(-e_i(N)+(a_i(N)-a_i(N)-\sqrt{-e_i(N)}-s))^2\right)^2=\left(s^2+\frac{s\Delta_i(N)}{\sqrt{3}}\right)^2.$$

Then the integral $\int_{a_i(N)+\sqrt{-e_i(N)}}^{t_i(N)}(e_i(N)-u^R(a_i(N),t))^2dt$ becomes

$$\int_0^{\Delta_i(N)\frac{\sqrt{3}-1}{2\sqrt{3}}}\left(s^2+\frac{s\Delta_i(N)}{\sqrt{3}}\right)^2 ds = A\cdot(\Delta_i(N))^5,$$

where $A>0$ is some constant. Hence, proving the inequality (4.11) is equivalent to showing that

$$\sum_{i=1}^{N}(\Delta_i(N))^5 \geq \sum_{i=1}^{N+1}(\Delta_i(N+1))^5. \tag{4.12}$$

As in the previous step, consider $t^x=(0=t_0^x,t_1^x,\dots,t_N^x=x,t_{N+1}^x=1)$ where the indifference condition (4.2) is satisfied for $i=2,\dots,N$. If $x=t_{N-1}(N)$, then $t_1^x=0$ and $\sum_{i=1}^{N+1}(\Delta_i^x)^5=\sum_{i=1}^{N}(\Delta_i(N))^5$; if $x=t_N(N+1)$, then $t^x=t(N+1)$ and $\sum_{i=1}^{N+1}(\Delta_i^x)^5=\sum_{i=1}^{N+1}(\Delta_i(N+1))^5$.

I will show that as x changes from $t_{N-1}(N)$ to $t_N(N+1)$, $\sum_{i=1}^{N+1}(\Delta_i^x)^5$ decreases. Indeed, differentiating w.r.t. x yields:

$$\frac{d}{dx}\left(\sum_{i=1}^{N+1}(\Delta_i^x)^5\right)=5\sum_{i=1}^{N}\frac{dt_i^x}{dx}((\Delta_i^x)^4-(\Delta_{i+1}^x)^4)<0$$

because $\frac{dt_i^x}{dx}>0$ and $\Delta_i^x<\Delta_{i+1}^x$ for all i. Thus, (4.12) holds, i.e., the Sender's expected cost is greater in $t(N)$-partition equilibrium. Hence, S ex-ante prefers a partition equilibrium with more intervals. **QED.**

Proof of Theorem 21. The proof has the same logic as argumentation in Section 5 of [Crawford and Sobel, 1982]. Here I consider the case of an increase in k: $k_1<k_2$ (the case of a decrease in bias b is proved in the same way). Fix the number of intervals in the equilibrium partition, N. Denote the equilibrium partition corresponding to k_i as t^{k_i}. As before, I will argue that there exists a continuous deformation of t^{k_1} to t^{k_2} along which the expected utility of R increases.

Denote by $t^k=(t_0^k,t_1^k,\dots,t_N^k)$ a partition that satisfies the indifference condition (4.2) with cost intensity k for $i=1,\dots,N-1$, $t_0^k=0$, and $t_N^k=1$. Note that if $k=k_1$, then $t^k=t^{k_1}$; if $k=k_2$, then $t^k=t^{k_2}$. In what follows, I show that as k changes from k_1 to k_2, the expected utility of R increases.

First, I prove the following lemma (an analogue of Lemma 4 of [Crawford and Sobel, 1982]):

Lemma 21. *If t^{k_1} and t^{k_2} are two partial partitions of length n satisfying the indifference condition (4.10) with $k_1 < k_2$, and $t_0^{k_1} = t_0^{k_2}$, then $t_n^{k_1} = t_n^{k_2}$ implies that $t_i^{k_1} < t_i^{k_i}$ for all $i = 1, ..., n-1$.*

Proof. The proof is by induction on n. Clearly, the statement holds for $n = 1$. Consider $n = 2$. Because the solution of (4.10), $\Delta_{i+1} = \Delta(\Delta_i, b, k)$, decreases in k, it must be the case that $t_1^{k_2} > t_1^{k_1}$. Consider $n = 3$. By the same logic, it must be that $t_1^{k_2} > t_1^{k_1}$ and $t_2^{k_2} > t_2^{k_1}$. Now consider $n \geq 4$. Clearly, $t_1^{k_2} > t_1^{k_1}$ and $t_{n-1}^{k_2} > t_{n-1}^{k_1}$. Assume that the statement does not hold for all i, and take the largest $j \in \{2, ..., N-2\}$, such that $t_j^{k_2} \leq t_j^{k_1}$. Let $t^x = (0 = t_0^x, x = t_1^x, ..., t_j^x)$ be a partition, such that (4.10) is satisfied with the cost intensity of k_2 for $i = 1, ..., j-1$. One can find $x \geq t_1^{k_2}$ such that $t_j^x = t_j^{k_1}$. Then, by the induction hypothesis, $t_i^x > t_i^{k_1}$ for all $i = 1, ..., j-1$. In particular, $t_{j-1}^x > t_{j-1}^{k_1}$, and if one continues this partition t^x (satisfying (4.10) with the cost intensity k_2) up to the n'th interval, one would get that $t_n^x < t_n^{k_1}$ (because $\Delta_{i+1} = \Delta(\Delta_i, b, k)$ decreases in k). But this leads to a contradiction that $t_n^{k_1} = t_n^{k_2}$ initially (because $x \geq t_1^{k_2}$). This proves the statement of the lemma. □

Lemma 21 immediately implies that t_i^k increases in k for all $i = 1, ..., N-1$. The expected payoff of R depending on k is

$$\mathbb{E}u^R(k) = \sum_{i=1}^{N} \int_{t_{i-1}^k}^{t_i^k} u^R(a_i^k, t)dt,$$

where $a_i^k = a(t_{i-1}^k, t_i^k) = \frac{t_{i-1}^k + t_i^k}{2}$ is the optimal action given the state is in the interval (t_{i-1}^k, t_i^k). Differentiating with respect to k and using the Envelope Theorem yields:

$$\frac{d\mathbb{E}u^R(k)}{dk} = \sum_{i=1}^{N-1} \frac{dt_i^k}{dk} \left(u^R(a_i^k, t_i^k) - U^R(a_{i+1}^k, t_i^k) \right).$$

Because t_i^k increases in k, $\frac{dt_i^x}{dx} > 0$, $i = 1, \ldots, N$. Now consider the expression in brackets for the term i in the sum:

$$u^R(a_i^k, t_i^k) - u^R(a_{i+1}^k, t_i^k) = -\frac{1}{4}\left(\Delta_i^k\right)^2 + \frac{1}{4}\left(\Delta_{i+1}^k\right)^2 > 0$$

because $\Delta_{i+1}^k > \Delta_i^k$ for $i = 1, \ldots, N-1$, and $k \in [k_1, k_2]$. **QED.**

Proof of Theorem 22. Step 1: Equilibrium conditions. First, I show that the belief $\beta(t|m)$ and equations (4.6), (4.7), (4.8) and (4.9) form equilibrium conditions. If the types in the interval (t_{i-1}, t_i) send the same message $m = i$, then the preferred action of the Receiver is determined from $\max_{a\in\mathbb{R}} \int_{t_{i-1}}^{t_i} \varphi(a-t)dt$. The best action a_i satisfies the following first-order condition:

$$\int_{t_{i-1}}^{t_i} \varphi'(a_i - t)dt = 0,$$

$$\varphi(a_i - t_i) - \varphi(a_i - t_{i-1}) = 0,$$

which, by symmetry of $\varphi(\cdot)$, implies that $a_i = \frac{t_{i-1}+t_i}{2}$. Thus, (4.8) provides R's best response, given the uniform belief β and S's strategy (4.7). R's expected utility in this case is given by:

$$e_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \varphi\left(\frac{t_{i-1}+t_i}{2} - t\right) dt = \frac{1}{\Delta_i} \int_0^{\Delta_i} \varphi\left(\frac{\Delta_i}{2} - s\right) ds,$$

which is exactly (4.9).

Now I will show that given the belief $\beta(m)$ (uniform on the intervals of the partition (4.6)) and R's action strategy presented in (4.8), the best response of S is described by (4.7). Note that if S sends the message i, then the induced payoff expectation is strictly greater than the actual utility at the ends of the interval:

$$\frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \varphi(a_i - t)dt > \frac{1}{t_i - t_{i-1}}(t_i - t_{i-1})\varphi(a_i - t_{i-1}) = \varphi(a_i - t_{i-1}) = \varphi(a_i - t_i).$$

This fact, accompanied by the strict concavity of φ, means that for any interval (t_{i-1}, t_i) there exist c_i^l and c_i^r, such that $t_{i-1} < c_i^l < \frac{t_{i-1}+t_i}{2} = a_i < c_i^r < t_i$ and $\psi(\varphi(a_i - t) - e_i) = 0$ for $t \in (c_i^l, c_i^r)$, and $\psi(\varphi(a_i - t) - e_i) > 0$, otherwise. Note that the intervals (c_i^l, c_i^r) and (c_j^l, c_j^r) do not intersect for $i \neq j$.

Consider two pairs (a_i, e_i) and (a_{i+1}, e_{i+1}) that correspond to the intervals (t_{i-1}, t_i) and (t_i, t_{i+1}), respectively. Type t_i sender is indifferent between the pairs if and only if

$$\underbrace{u^S(a_i, t_i, b) - u^S(a_{i+1}, t_i, b)}_{LHS(t_i)} = \underbrace{k\left[\psi\left(\varphi(a_i - t_i) - e_i\right) - \psi\left(\varphi(a_{i+1} - t_i) - e_{i+1}\right)\right]}_{RHS(t_i)}.$$

Now I will argue that the type s sender prefers the pair (a_i, e_i) to (a_{i+1}, e_{i+1}) if and only if $s \leq t_i$. The condition that the pair (a_i, e_i) is better for sender s than (a_{i+1}, e_{i+1}) is

$$\underbrace{u^S(a_i, s, b) - u^S(a_{i+1}, s, b)}_{LHS(s)} > k\,\underbrace{(\psi(\varphi(a_i - s) - e_i) - \psi(\varphi(a_{i+1} - s) - e_{i+1}))}_{RHS(s)}.$$

Because $LHS(t_i) = RHS(t_i)$, it is sufficient to show that $LHS(s)$ is strictly decreasing in s and $RHS(s)$ is increasing in s.

$LHS(s)$ can be rewritten as

$$LHS(s) = -\int_{a_i}^{a_{i+1}} u_1^S(a, s, b)da.$$

Differentiation with respect to s gives

$$LHS'(s) = -\int_{a_i}^{a_{i+1}} u_{12}^S(a, s, b)da < 0,$$

because $u_{12}^S(a, s, b) > 0$ for all s. Thus, LHS strictly decreases with s.

Differentiating $RHS(s)$ gives

$$RHS'(s) = -\psi'(\varphi(a_i - s) - e_i)\varphi'(a_i - s) + \psi'(\varphi(a_{i+1} - s) - e_{i+1})\varphi'(a_{i+1} - s).$$

To show that $RHS'(s) > 0$, I separately consider 5 intervals to which [0,1] interval is split by the points $0 < c_i^l < c_i^r < c_{i+1}^l < c_{i+1}^r < 1$.

Case 1. $s \in [c_i^l, c_i^r]$. For these types, the cost from inducing pair (a_i, e_i) is 0, hence

$$RHS'(s) = \underbrace{\psi'(\varphi(a_{i+1} - s) - e_{i+1})}_{<0}\underbrace{\varphi'(a_{i+1} - s)}_{<0} > 0,$$

where $\psi'(\varphi(a_{i+1} - s) - e_{i+1}) < 0$ because $\varphi(a_{i+1} - s) - e_{i+1} < 0$, and $\varphi'(a_{i+1} - s) < 0$ because $a_{i+1} - s > 0$.

Case 2. $s \in (c_i^r, c_{i+1}^l)$. For these types, the costs from inducing both pairs are nonzero,

$$RHS'(s) = -\underbrace{\psi'(\varphi(a_i - s) - e_i)}_{<0}\underbrace{\varphi'(a_i - s)}_{>0} + \underbrace{\psi'(\varphi(a_{i+1} - s) - e_{i+1})}_{<0}\underbrace{\varphi'(a_{i+1} - s)}_{<0} > 0,$$

where $\psi'(\varphi(a_i - s) - e_i) < 0$ because $\varphi(a_i - s) - e_i < 0$; $\varphi'(a_i - s) > 0$ because $a_i - s < 0$; $\psi'(\varphi(a_{i+1} - s) - e_{i+1}) < 0$ because $\varphi(a_{i+1} - s) - e_{i+1} < 0$; and $\varphi'(a_{i+1} - s) < 0$ because $a_{i+1} - s > 0$.

Case 3. $s \in [c^l_{i+1}, c^r_{i+1}]$. For these types, the cost from inducing pair (a_{i+1}, e_{i+1}) is 0, hence

$$RHS'(s) = -\underbrace{\psi'(\varphi(a_i - s) - e_i)}_{<0}\underbrace{\varphi'(a_i - s)}_{>0} > 0,$$

where $\psi'(\varphi(a_i - s) - e_i) < 0$ because $\varphi(a_i - s) - e_i < 0$, and $\varphi'(a_i - s) > 0$ because $a_i - s < 0$.

Case 4. $s \in [0, c^l_i)$. For these types, the cost from inducing any pair is non-zero and

$$RHS'(s) = -\underbrace{\psi'(\varphi(a_i - s) - e_i)}_{<0}\underbrace{\varphi'(a_i - s)}_{<0} + \underbrace{\psi'(\varphi(a_{i+1} - s) - e_{i+1})}_{<0}\underbrace{\varphi'(a_{i+1} - s)}_{<0}.$$

Note that $|\varphi'(a_i - s)| < |\varphi'(a_{i+1} - s)|$. To prove that $RHS'(s) > 0$, it is sufficient to show that $|\psi'(\varphi(a_i - s) - e_i)| < |\psi'(\varphi(a_{i+1} - s) - e_{i+1})|$, i.e.,

$$0 > \varphi(a_i - s) - e_i > \varphi(a_{i+1} - s) - e_{i+1},$$

$$\varphi(a_i - s) - \varphi(a_{i+1} - s) > e_i - e_{i+1}.$$

Differentiating the left-hand side yields

$$-\varphi'(a_i - s) + \varphi'(a_{i+1} - s) < 0,$$

meaning that it is decreasing in s. Hence,

$$\varphi(a_i - s) - \varphi(a_{i+1} - s) > \varphi(a_i - c^l_i) - \varphi(a_{i+1} - c^l_i) = e_i - \varphi(a_{i+1} - c^l_i) > e_i - e_{i+1}$$

because $\varphi(a_i - c^l_i) = e_i$ (by the definition of c^l_i) and $\varphi(a_{i+1} - c^l_i) < e_{i+1}$. Hence, $|\psi'(\varphi(a_i - s) - e_i)| < |\psi'(\varphi(a_{i+1} - s) - e_{i+1})|$ and $RHS'(s) > 0$.

Case 5. $s \in (c^r_i, 1]$. For these types, the cost from inducing any pair is non-zero and

$$RHS'(s) = -\underbrace{\psi'(\varphi(a_i - s) - e_i)}_{<0}\underbrace{\varphi'(a_i - s)}_{>0} + \underbrace{\psi'(\varphi(a_{i+1} - s) - e_{i+1})}_{<0}\underbrace{\varphi'(a_{i+1} - s)}_{>0}.$$

Note that $|\varphi'(a_i - s)| > |\varphi'(a_{i+1} - s)|$. To prove that $RHS'(s) > 0$, it is sufficient to show that $|\psi'(\varphi(a_i - s) - e_i)| > |\psi'(\varphi(a_{i+1} - s) - e_{i+1})|$, i.e.,

$$0 > \varphi(a_{i+1} - s) - e_{i+1} > \varphi(a_i - s) - e_i,$$

$$\varphi(a_i - s) - \varphi(a_{i+1} - s) < e_i - e_{i+1}.$$

Differentiating the left-hand side yields

$$-\varphi'(a_i - s) + \varphi'(a_{i+1} - s) < 0,$$

meaning that it is decreasing in s. Hence,

$$\varphi(a_i - s) - \varphi(a_{i+1} - s) < \varphi(a_i - c^r_{i+1}) - \varphi(a_{i+1} - c^r_{i+1}) = \varphi(a_i - c^r_{i+1}) - e_{i+1} < e_i - e_{i+1}.$$

Hence, $|\psi'(\varphi(a_i - s) - e_i)| > |\psi'(\varphi(a_{i+1} - s) - e_{i+1})|$ and $RHS'(s) > 0$.

This proves that the best response of S takes the form described in Theorem 22.

Step 2: Monotonicity of Δ_i. Now I will argue that $\Delta_i < \Delta_{i+1}$ for all $i = 1, ..., N-1$. Type t_i sender is indifferent between pair (a_i, e_i) and (a_{i+1}, e_{i+1}) if and only if

$$\underbrace{u^S\left(t_i - \frac{\Delta_i}{2}, t, b\right) - u^S\left(t_i + \frac{\Delta_{i+1}}{2}, t, b\right)}_{LHS} = \underbrace{k\left[\psi\left(\varphi\left(\frac{\Delta_i}{2}\right) - e_i\right) - \psi\left(\varphi\left(\frac{\Delta_{i+1}}{2}\right) - e_{i+1}\right)\right]}_{RHS}. \tag{4.13}$$

First, I show that $\Delta_i \neq \Delta_{i+1}$. Suppose the opposite: $\Delta_i = \Delta_{i+1}$. Then $e_i = e_{i+1}$, which leads to $RHS = 0$, while $LHS < 0$ because $u^S(\cdot, t_i, b)$ is symmetric in the first argument around $a^S(t_i, b) > t_i$. Thus, it cannot be that $\Delta_i = \Delta_{i+1}$.

Second, suppose $\Delta_i > \Delta_{i+1}$. Then $LHS < 0$ by the same symmetry reason. Now I show that $RHS > 0$. Consider a function

$$g(\Delta) = \varphi\left(\frac{\Delta}{2}\right) - e(\Delta) = \varphi\left(\frac{\Delta}{2}\right) - \frac{1}{\Delta}\int_0^{\Delta} \varphi\left(\frac{\Delta}{2} - s\right) ds.$$

Differentiating $g(\cdot)$ w.r.t. Δ gives

$$\begin{aligned} g'(\Delta) &= \frac{1}{2}\varphi'\left(\frac{\Delta}{2}\right) + \frac{1}{\Delta^2}\int_0^{\Delta} \varphi\left(\frac{\Delta}{2} - s\right) ds - \frac{1}{\Delta}\varphi\left(\frac{\Delta}{2}\right) \\ &= \frac{1}{2}\varphi'\left(\frac{\Delta}{2}\right) + \frac{1}{\Delta}\left[\varphi\left(\tilde{s}\right) - \varphi\left(\frac{\Delta}{2}\right)\right] \\ &= \frac{1}{2}\varphi'\left(\frac{\Delta}{2}\right) - \frac{1}{\Delta}\varphi'(\hat{s})\left(\frac{\Delta}{2} - \tilde{s}\right) \\ &= \frac{1}{2}\varphi'\left(\frac{\Delta}{2}\right) - \varphi'(\hat{s})\left(\frac{1}{2} - \frac{\tilde{s}}{\Delta}\right) < 0. \end{aligned}$$

Here, the first equality is an application of Leibniz's rule; the second equality follows from the mean value theorem for integration with $\tilde{s} \in (0, \Delta/2)$. The third equality follows from the mean value theorem for differentiation with $\hat{s} \in (\tilde{s}, \Delta/2)$. The last expression is strictly negative because $\left(\frac{1}{2} - \frac{\tilde{s}}{\Delta}\right) < \frac{1}{2}$ and, by concavity of φ, $0 < -\varphi'(\hat{s}) < -\varphi'\left(\frac{\Delta}{2}\right)$. The negative derivative of g means that if $\Delta_i > \Delta_{i+1}$, then $g(\Delta_i) < g(\Delta_{i+1}) < 0$. This, by convexity of ψ, implies that $\psi(g(\Delta_i)) - \psi(g(\Delta_{i+1})) > 0$, i.e., $RHS > 0$. Contradiction. As a result, $\Delta_i < \Delta_{i+1}$ is the only case possible. **QED.**

Bibliography

[Abaluck and Gruber, 2011] Jason Abaluck and Jonathan Gruber. Choice inconsistencies among the elderly: Evidence from plan choice in the medicare part d program. *American Economic Review*, 101(4):11801210, Jun 2011.

[Acemoglu *et al.*, 2007] Daron Acemoglu, Victor Chernozhukov, and Muhamet Yildiz. Learning and disagreement un an uncertain world. 2007.

[Alonso *et al.*, 2008] Ricardo Alonso, Wouter Dessein, and Niko Matouschek. When does coordination require centralization? *American Economic Review*, 98(1):145–179, Mar 2008.

[Andreoni and Bernheim, 2009] James Andreoni and B. Douglas Bernheim. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636, Sep 2009.

[Aumann, 1976] Robert J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[Bala and Goyal, 2000] Venkatesh Bala and Sanjeev Goyal. A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229, Sep 2000.

[Battigalli and Dufwenberg, 2007] Pierpaolo Battigalli and Martin Dufwenberg. Guilt in games. *American Economic Review*, 97(2):170–176, May 2007.

[Becker, 1974] Gary S. Becker. Theory of social interactions. *Journal of Political Economy*, 82:1063–1093, 1974.

[Benham, 2005] Kelley Benham. From ordinary girl to international icon. *St. Petersburg Times*, Mar 2005.

[Bolton and Dewatripont, 1994] Patrick Bolton and Mathias Dewatripont. The firm as a communication network. *Quarterly Journal of Economics*, 109(4):809–839, Nov 1994.

[Cai and Wang, 2006] Hongbin Cai and Joseph Tao-Yi Wang. Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36, Jul 2006.

[Caillaud and Tirole, 2007] Bernard Caillaud and Jean Tirole. Consensus building: How to persuade a group. *American Economic Review*, 97(5):1877–1900, Dec 2007.

[Calvó-Armengol and de Martí, 2007] Antoni Calvó-Armengol and Joan de Martí. Communication networks: Knowledge and decisions. *American Economic Review*, 97(2):86–91, May 2007.

[Calvó-Armengol and de Martí, 2009] Antoni Calvó-Armengol and Joan de Martí. Information gathering in organizations: equilibrium, welfare, and optimal network structure. *Journal of the European Economic Association*, 7:116–161, Mar 2009.

[Calvó-Armengol *et al.*, 2011] Antoni Calvó-Armengol, Joan de Martí, and Andrea Prat. Communication and influence. 2011.

[Camerer, 2003] Colin Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press, 2003.

[Carlin *et al.*, 2010] Bruce Carlin, Simon Gervais, and Gustavo Manso. Libertarian paternalism, information sharing, and financial decision-making. 2010.

[Charness and Dufwenberg, 2006] Gary Charness and Martin Dufwenberg. Promises and partnership. *Econometrica*, 74(6):1579–1601, Nov 2006.

[Che and Kartik, 2009] Yeon-Koo Che and Navin Kartik. Opinions as incentives. *Journal of Political Economy*, 117(5):815–860, Oct 2009.

[Chwe, 2000] Michael Suk-Young Chwe. Communication and coordination in social networks. *Review of Economic Studies*, 67(1):1–16, Jan 2000.

[Cohen and Einav, 2007] Alma Cohen and Liran Einav. Estimating risk preferences from deductible choice. *American Economic Review*, 97(3):745788, Jun 2007.

[Crawford and Sobel, 1982] Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, Nov 1982.

[Currarini *et al.*, 2009] Sergio Currarini, Matthew O. Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, Jul 2009.

[Cutler and Zeckhauser, 2004] David M. Cutler and Richard Zeckhauser. Extending the theory to meet the practice of insurance. *Brookings-Wharton Papers on Financial Services*, pages 1–53, 2004.

[Dessein and Santos, 2006] Wouter Dessein and Tano Santos. Adaptive organizations. *Journal of Political Economy*, 114(5):956–995, Oct 2006.

[Dessein, 2002] Wouter Dessein. Authority and communication in organizations. *Review of Economic Studies*, 69(4):811–838, Oct 2002.

[Dessein, 2007] Wouter Dessein. Why a group needs a leader: Decision-making and debate in committees. 2007.

[Dewatripont and Tirole, 2005] Mathias Dewatripont and Jean Tirole. Modes of communication. *Journal of Political Economy*, 113(6):1217–1238, Dec 2005.

[Dufwenberg and Gneezy, 2000] Martin Dufwenberg and Uri Gneezy. Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30(2):163–182, Feb 2000.

[Dworkin, 2010] Gerald Dworkin. Paternalism. *The Stanford Encyclopedia of Philosophy*, Jun 2010.

[Einav *et al.*, 2010] Liran Einav, Amy Finkelstein, and Mark R. Cullen. Estimating welfare in insurance markets using variation in prices. *Quarterly Journal of Economics*, 125(3):877–921, Aug 2010.

[Eső and Galambos, 2008] Peter Eső and Ádám Galambos. Disagreement and evidence production in pure communication games. 2008.

[Fang *et al.*, 2008] Hanming Fang, Michael P. Keane, and Dan Silverman. Sources of advantageous selection: Evidence from the medigap insurance market. *Journal of Political Economy*, 116(2):303–350, Apr 2008.

[Farrell and Gibbons, 1989] Joseph Farrell and Robert Gibbons. Cheap talk with two audiences. *American Economic Review*, 79(5):1214–1223, Dec 1989.

[Galeotti *et al.*, 2011] Andrea Galeotti, Christian Ghiglino, and Francesco Squintani. Strategic information transmission in networks. 2011.

[Gneezy, 2005] Uri Gneezy. Deception: The role of consequences. *American Economic Review*, 95(1):384–394, Mar 2005.

[Goltsman and Pavlov, 2011] Maria Goltsman and Gregory Pavlov. How to talk to multiple audiences. *Games and Economic Behavior*, 72(1):100–122, May 2011.

[Goyal, 2007] Sanjeev Goyal. *Connections: An introduction to the economics of networks*. Princeton University Press, 2007.

[Green and Stokey, 2007] Jerry R. Green and Nancy L. Stokey. A two-person game of information transmission. *Journal of Economic Theory*, 135(1):90–104, Jul 2007.

[Greer, 2000] George W Greer. In re: the guardianship of Theresa Marie Schiavo, Incapacitated. *Florida Sixth Judicial Circuit. Retrieved 2006-01-08.*, Feb 2000.

[Grossman, 1981] Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law & Economics*, 24(3):461–483, Dec 1981.

[Hagenbach and Koessler, 2010] Jeanne Hagenbach and Frederic Koessler. Strategic communication networks. *Review of Economic Studies*, 77(3):1072–1099, Jul 2010.

[Hanson, 2003] Robin Hanson. Warning labels as cheap-talk: why regulators ban drugs. *Journal of Public Economics*, 87(9-10):2013–2029, Sep 2003.

[Harris, 2001] Nonie M. Harris. The euthanasia debate. *J R Army Med Corps.*, 147:367–370, 2001.

[Hirsch, 2011] Alexander V. Hirsch. Experimentation and persuasion in political organizations. 2011.

[Hurkens and Kartik, 2009] Sjaak Hurkens and Navin Kartik. Would i lie to you? on social preferences and lying aversion. *Experimental Economics*, 12(2):180–192, Jun 2009.

[Jackson and Wolinsky, 1996] Matthew O. Jackson and Asher Wolinsky. A strategic model of social and economic networks. *Journal of Economic Theory*, 71(1):44–74, Oct 1996.

[Jackson, 2008] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[Kartik *et al.*, 2007] Navin Kartik, Marco Ottaviani, and Francesco Squintani. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1):93–116, May 2007.

[Kartik, 2009] Navin Kartik. Strategic communication with lying costs. *Review of Economic Studies*, 76(4):1359–1395, Oct 2009.

[Lazarsfeld and Merton, 1954] Paul F. Lazarsfeld and Robert K. Merton. *Friendship as a Social Process: A Substantive and Methodological Analysis*. New York: Van Nostrand Company, Inc., 1954.

[Lee and Persson, 2011] Samuel Lee and Petra Persson. Circles of trust. 2011.

[Locke, 1689] John Locke. *Two Treatises of Government*. 1689.

[Loginova and Persson, 2012] Uliana Loginova and Petra Persson. Paternalism, libertarianism, and the nature of disagreement. 2012.

[Loginova, 2012a] Uliana Loginova. Strategic communication in networks: Preference versus opinion conflict. 2012.

[Loginova, 2012b] Uliana Loginova. Strategic communication in networks: The choice between soft and hard information transmission. 2012.

[Lovett, 2012] Ian Lovett. Law on condoms threatens tie between sex films and their home. *The New York Times*, Mar 2012.

[Lundquist *et al.*, 2009] Tobias Lundquist, Tore Ellingsen, Erik Gribbe, and Magnus Johannesson. The aversion to lying. *Journal of Economic Behavior & Organization*, 70(1-2):81–92, May 2009.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[Milgrom, 1981] Paul R. Milgrom. Good news and bad news - representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391, 1981.

[Mill, 1859] John Stuart Mill. *On liberty*. 1859.

[Moody, 2001] James Moody. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, 107(3):679–716, Nov 2001.

[Morgan and Stocken, 2008] John Morgan and Phillip C. Stocken. Information aggregation in polls. *American Economic Review*, 98(3):864–896, Jun 2008.

[Morris, 1995] Stephen Morris. The common prior assumption in economic theory. *Economics and Philosophy*, 11(2):227–253, Oct 1995.

[Persson, 2011] Petra Persson. Information overload, obfuscation, and distraction. 2011.

[Radner, 1992] Roy Radner. Hierarchy - the economics of managing. *Journal of Economic Literature*, 30(3):1382–1415, Sep 1992.

[Radner, 1993] Roy Radner. The organization of decentralized information-processing. *Econometrica*, 61(5):1109–1146, Sep 1993.

[Rantakari, 2008] Heikki Rantakari. Governing adaptation. *Review of Economic Studies*, 5(4):1257–1285, Oct 2008.

[Sah and Stiglitz, 1986] Raaj Kumar Sah and Joseph E. Stiglitz. The architecture of economic-systems - hierarchies and polyarchies. *American Economic Review*, 76(4):716–727, Sep 1986.

[Sánchez Pagés and Vorsatz, 2007] Santiago Sánchez Pagés and Marc Vorsatz. An experimental study of truth-telling in a sender-receiver game. *Games and Economic Behavior*, 61(1):86–112, Oct 2007.

[Sethi and Yildiz, 2009] Rajiv Sethi and Muhamet Yildiz. Public disagreement. 2009.

[Spinnewijn, 2012] Johannes Spinnewijn. Heterogeneity, demand for insurance and adverse selection. 2012.

[Suber, 1999] P Suber. Paternalism. *Philosophy of Law: An Encyclopedia*, 2:632635, 1999.

[Tadelis, 2011] Steve Tadelis. The power of shame and the rationality of trust. Mar 2011.

[Thaler and Sunstein, 2003] Richard H. Thaler and Cass R. Sunstein. Libertarian paternalism. *American Economic Review*, 93:175–179, May 2003.

[Thaler and Sunstein, 2008] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press, 2008.

[Tversky and Kahneman, 1974] Amos Tversky and Daniel Kahneman. Judgment under uncertainty - heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[Van den Steen, 2006] Eric Van den Steen. The limits of authority: motivation versus coordination. 2006.

[Van den Steen, 2009] Eric Van den Steen. Authority versus persuasion. *American Economic Review*, 99:448–453, May 2009.

[Van Zandt and Radner, 2001] Timothy Van Zandt and Roy Radner. Real-time decentralized information processing and returns to scale. *Economic Theory*, 17(3):545–575, May 2001.

[Vanberg, 2008] Christoph Vanberg. Why do people keep their promises? an experimental test of two explanations. *Econometrica*, 76(6):1467–1480, Nov 2008.

[Whiting, 2002] Raymond Whiting. *A Natural Right to Die: Twenty-Three Centuries of Debate.* Westport: Greenwood Press, 2002.

[Yardley, 2012] William Yardley. Big sky, bright sun and melanoma. *The New York Times*, Mar 2012.