Paternalism, Libertarianism, and the Nature of Disagreement [*]

Uliana Loginova[†] Petra Persson[‡]

July 2012

Abstract

Regulation to protect individuals from self-harm, such as euthanasia prohibitions and safety mandates, is widespread but controversial. Opponents and proponents are often believed to differ in their valuation of individual liberty. We model an authority's decision to constrain or inform a population of agents prone to self-harm and propose an alternative view: A benevolent politician's 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.

JEL Classification: D64, D78, D83, K32
Keywords: Paternalism, Libertarianism, Communication Difference of opinion, Coercion

[*] We are grateful to Navin Kartik for his generous support at all stages of this project. We also thank Patrick Bolton, Alessandra Casella, Yeon-Koo Che, Vouter Dessein, Samuel Lee, Andrea Prat, Rajiv Sethi, Joel Sobel, Giancarlo Spagnolo, and participants of the Columbia University Economic Theory Colloquium for helpful comments.

[†] U. Loginova, Department of Economics, Columbia University, 420 West 118th Street, MC 3308, New York, NY 10027; e-mail: ul2108@columbia.edu.

[‡] P. Persson, Department of Economics, Columbia University, 420 West 118th Street, MC 3308, New York, NY 10027; e-mail: pmp2116@columbia.edu.

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" (Harris (2001), p. 367)—constitutes criminal homicide in most jurisdictions, even if committed at the patient's request. Yet debates on the topic are ongoing in several countries. Some deem it the right of the incurably ill to end their own suffering, while others repudiate such requests, demanding that the government protect such requestors from themselves.

A restriction on individual freedom justified solely on the grounds that it makes a person better off represents an instance of paternalism (Dworkin (2010)). Paternalism has the ring of benevolence: Like a father (lat. *pater*) disciplining his child out of love, the government constrains the populace in the populace's best interest. But this raises a central question: Would not a better-informed government—a benevolent one, at least—just provide advice and 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, forbid swimming at public beaches when lifeguards are not present, prevent women from working heavy-duty jobs, require minors to have life-saving blood transfusions even when their religious beliefs forbid it, protect individuals against skin cancer by banning the use of tanning beds, or protect the health of sex film workers by requiring the use of condoms (Lovett (2012))?

One commonly held view is that advocates of a laissez faire approach—to provide information but let each individual make his own choice—simply place

[1] There was no living will. Dying was, however, affirmed as her wish in court after the testimony of 18 witnesses on her end-of-life wishes (Greer (2000)). The obituary, published in *St. Petersburg Times*, was written by Benham (2005).

a higher value on individual liberty than advocates of regulation. "I kind of believe in personal freedom," explains Dr. Steven Mings, a past president of the Idaho Dermatology Society, who opposes a bill to ban minors from using tanning beds (Yardley (2012)). In response to the same bill, the Freedom Foundation voiced concerns over Idaho becoming a "nanny state" (Zuckerman (2012)). Similarly, opponents of euthanasia regulation insist that choosing when to die is a fundamental individual liberty.

But the view that opponents and proponents of paternalistic regulation simply differ in their valuations of individual liberty is hard to reconcile with the fact that some are in favor of regulation that others oppose *and vice versa*. If, say, R simply 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 restrictions on individual liberty do not always satisfy this property: Consider Jeb Bush, former governor of Florida, and Bill Nelson, Senator of Florida. On the issue of Terri Schiavo, Bush fought to intervene and restore Schiavo's removed feeding tube, whereas Nelson opted not to co-sponsor a bill to intervene (The Washington Post (2005)).[2] On the issue of traffic safety, however, Bush vetoed a mandate he deemed "too intrusive," whereas Nelson recently voted in favor of tighter regulation (Lade (2012), OpenCongress (2012)). In the same vein, in both Washington and Oregon, where euthanasia is legal, seat belt laws are stricter than in many states where it is illegal.

So what, then, explains differing views on the regulation of activities that cause self-harm? We model an authority's decision to constrain or inform a population of agents prone to self-harm and propose an alternative answer: A politician's decision to regulate an activity is determined by (i) her benevolence and (ii) whether she deems the action a matter of preference or opinion. A benevolent politician safeguards individual liberty on decisions that, in her view, reflect preferences; however, she enacts paternalistic regulation to constrain decisions that, in her view, reflect opinions. This is consistent with some (politicians) favoring regulation that others oppose and vice versa. More importantly, our theory may be helpful in developing a better understanding of why and where paternalistic regulation emerges, since it yields a precise prediction about which issues a given politician wants to regulate.

These insights emerge in a framework where an authority is faced with a continuum of subjects, each of which must choose an action, for example,

[2] The bill gave the federal court jurisdiction in the case in an effort to restore Schiavo's feeding tube (Allen (2005)).

whether to wear a seat belt. By assumption, the 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; say, the risks associated with inaction. Even if the authority transfers her private information to a subject, the two (still) disagree on the proper course of action because they have different preferences (Crawford and Sobel (1982)) or different opinions (priors) about the true state of the world (Che and Kartik (2009)). In the context of our example, given the same information about risks, (i) if the subject agrees with the authority on the risks but 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 population of subjects has either heterogenous preferences or heterogenous priors. The authority has two tools at her disposal: laissez faire and paternalism. Under laissez faire, she issues a non-verifiable (cheap talk) recommendation—in which case she may face a non-zero lying cost—and then gives each subject the liberty to choose his course of action. Under paternalism, the authority incurs a cost to overrule the subjects' own action choices by mandating a certain action. The authority's benevolence, or altruism, is modeled as the share of the individuals' material payoffs that she internalizes, $\varphi \geq 0$ (Becker (1974)).

After an illustrative example in Section 2, Section 3 develops our framework. We begin our general analysis in Section 4 by studying an *advisor*. Unlike an authority, an advisor has only one of the two tools at her disposal, namely, laissez faire (but not paternalism). We study how the advisor's decision to issue a truthful recommendation depends on her level of altruism. 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 lesser 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 private signal she observes about the state of the world, maximizes

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

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 actions they take when she lies. Lying then protects the individuals from misinterpreting a truthful report. When φ increases, the advisor internalizes more of the disutility that she expects each individual to suffer from his suboptimal (in her view) choice of action. Paradoxically, a sufficiently altruistic advisor may therefore lie.[4]

We then consider an *authority* that can either send a public message and then let each individual choose his action (laissez faire), *or* incur some cost to mandate one action for all individuals (paternalism). 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; then, if she allows the individuals to choose their actions, they will base their choices on less information than she has. Second, if the authority is sufficiently altruistic, she enacts a mandate even if truthful communication is 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.

Section 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 advi-

[4] This is particularly remarkable given that, before her private information is observed, the advisor's ex ante utility is higher in a truthful equilibrium. A strongly altruistic advisor is nevertheless non-credible, since she, ex interim, would prefer to lie—out of benevolence—given that individuals believe her report.

sor 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 becomes larger. 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. Otherwise, the advisor is more credible with private messages; she still issues truthful private advice to some individuals. 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.

Equipped with these results, in Section 6 we return to the central issues that we set out to answer: What determines whether a politician advocates laissez faire or paternalism and why do those who advocate restrictions in individual liberty deem such restrictions better than information provision? Our results offer a precise prediction: To understand what issues a benevolent authority regulates, it suffices to ask what issues she deems a matter of opinion. A benevolent politician favors euthanasia if she is convinced that an individual's request to die reflects his own true preference. She instead outlaws euthanasia if she is convinced that the requestor has an incorrect understanding of his own wish to die. This can arise, for example, if the politician believes that suicide is a sin that precludes the individual from afterlife benefits that the requestor, if aware of this, would not want to give up. Whenever the authority fears that the individuals' actions are 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.

Put differently, we find that the distinction between preferences and opinions is crucial for whether intervention is socially beneficial or harmful. This insight is illustrated in a recent set of papers that empirically estimate the welfare gain from a universal health insurance mandate. These papers depart from the same empirical observation—that insurance decisions cannot be explained by individual risk types alone—but make different assumptions about *what else* determines insurance decisions. Cohen and Einav (2007) and Einav et al. (2010) attribute the unexplained variation in the demand for insurance to *preferences*: Some uninsured simply have a preference for risk. Spinnewijn (2012) suggests instead that some uninsured have incorrect perceptions, or *opinions*, about their own true risk types and hence of their insurance needs. In the spirit of our general insights, Spinnewijn (2012) finds that the welfare gain from a universal mandate is higher the more insurance decisions are driven by erroneous risk perceptions.[5]

1.1 Related Literature

This paper is related to that of Che and Kartik (2009), who contrast differences in opinion with differences in preferences in a communication game. The authors show that only opinion differences between a principal and an agent can incentivize the agent to exert effort to persuade the principal. Van den Steen (2006), Van den Steen (2009), and Hirsch (2011) also analyze this mechanism. 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. Relatedly, Van den Steen (2009) shows that a principal may incur a cost to alter the agent's beliefs to boost the agent's effort. 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 may find it optimal to allow the agent to implement the agent's preferred policy. 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 of Che and Kartik (2009): The distinction between differing opinions and differing preferences may be crucial.

The key mechanism in our paper, instead, is altruism. This relates the paper to the emerging literature on communication and altruism. Carlin et al. (2010) show that an altruistic (unbiased) principal may share information with an uninformed set of agents to help improve their action choices, but that this may hamper the agents' individual incentives to acquire information. Lee and Persson (2011) analyze how friends transmit hard information to each other

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

when sharing information dilutes its value. We contribute to this literature by analyzing difference of opinion between the communicating parties. Moreover, we contrast communication with coercive measures to affect individuals' action choices. 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 does not constrain the individual's choice set—and (hard) *paternalism*.

Our analysis of targeted advice (but not mandates) relates to the work 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 and preference disagreement. When we shut down altruism in our model, we replicate their result that the relative credibility of private versus 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 that of Hanson (2003), who models a regulator who is empowered to ban an activity or to warn the public about it. The author shows that when a government is concerned about some market imperfection, cheap talk may not be credible, so the government may resort to prohibition. Our mechanism is distinct in two ways. First, we study a setting without any externalities, to rule out any motives for regulation other 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, however, a benevolent government would regulate under differences of opinion 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.

2 Example

Before introducing the general model, we present a simple setting with one individual that illustrates our key results and their underlying mechanisms. 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 cares about the action choice not only 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 (PBE).

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, $a_I(0)$.[6] A truthful equilibrium thus exists iff $a_I(1) - a_A(1) \leq a_A(1) - a_I(0)$, which can be written

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

where $\tau \equiv a_I(1) - a_I(0) > 0$ 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, (TT$_{c\,=\,0}$) reduces to

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

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

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

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, since 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 $K \equiv a_I(1) - a_A(1) > 0$, we can thus write (TT$_{c\,=\,0}$) as

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

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 $\pi_I \leq 0.604$ (we recall that $\pi_A = 0.5$). When $c > 0$, a truthful equilibrium exists iff

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

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

The lying cost matters only if the advisor prefers to lie when $c = 0$, that is, 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 suboptimal (in her view) 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 as $\varphi \leq 0.188$; when she cares more about the individual, she lies.

The second part of the paper 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, $a_A(s)$.[7] 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*, since she wants to inform the individual and then let him choose his own 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* since she constrains the individual's liberty in, as we shall see, his supposed

[7] In most applications discussed in Section 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 (online) Appendix D.

self-interest.

The logic driving this result is as follows. After obtaining the signal s, the advisor prefers to send some message m, which induces $a_I(m)$, over imposing $a_A(s)$ iff

$$\begin{aligned} & \mathbb{E}_A\left\{\left(u_A(a_A(s),\theta)+\varphi u_I(a_A(s),\theta)\right)-\left(u_A(a_I(m),\theta)+\varphi u_I(a_I(m),\theta)\right)\right\} \\ < \quad & q-c\mathbf{I}_{\{m\neq s\}}. \end{aligned} \tag{1}$$

A truthful equilibrium exists if it exists in the advisor game (above) and if truthful communication is preferred to coercion, that is, if (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 $m=s$, (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 that 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 (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 $m=s$, (1) reduces to

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

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 $a_A(s)$ since she believes that this protects the individual from his erroneous (in her view) action choice and thus ultimately benefits him. Since 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 $\varphi \geq 0.6$.

The remainder of the paper studies a more general model where the advisor or authority is faced with a population, that is, a continuum of individuals with heterogeneous preferences or beliefs. We show that the above insights remain applicable and we develop additional results. Proofs are given in Appendix A and (online) Appendix B. Appendix C demonstrates that the main results, driven by intuitively analogous mechanisms, continue to apply in a setting that is closely related in spirit but where all results arise in a dominance-solvable setting. Appendix D formally shows that all of the results discussed in the example above (also) obtain in a richer setting and with general (mixed and pure) strategies; further, it shows that the main insights arise in the presence of both preference and opinion disagreement.

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 gives him a 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 $g(\pi)$.[8]

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 a 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 a 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, that is, $\int_{b\neq 0} f(b)db > 0$; 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

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

denotes the set $\{a_i\}_{i\in[0,1]}$. Her material benefit is thus maximized when each individual's action a_i matches the state of the world, θ. We allow the advisor to be altruistic, that is, to internalize a share φ of the individuals' payoffs (Becker (1974)). Here φ is the marginal rate of substitution of the advisor's material payoff for the material payoffs of the individuals.[9] Her utility is thus given by

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

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 $g(\pi)$, that is, $\int_0^{0.5} g(\pi)d\pi \geq 0.5$ 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 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.[10] The advisor's material payoff is given by $u_A(a,\theta) = -c \cdot \mathbf{I}(m \neq s)$. 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

$$U_A(a,\theta) = u_A - \varphi \int_0^1 (a_i - \theta)^2 g(\pi_i) d\pi_i. \tag{3}$$

Note that the setting remains essentially equivalent if the material payoff of the advisor is defined similarly to the case of preference disagreement, that is, $u_A(a,\theta) = -\int_0^1 (a_i-\theta)^2 g(\pi_i)d\pi_i - c\cdot\mathbf{I}(m \neq s)$. Indeed, in this case the utility of the advisor is $U_A(a,\theta) = -c\cdot\mathbf{I}(m\neq s) - (1+\varphi)\int_0^1 (a_i-\theta)^2 g(\pi_i)d\pi_i$, which is equivalent to (3) if altruism under opinion disagreement is redefined as $\widetilde{\varphi} = 1 + \varphi$.

[9] Section 7 discusses an 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.

[10] Note that in the setting with preference disagreement we do not assume that the advisor is representative of the median preference. Since 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.

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, the action $a_i(m)$ that he takes, $a_i : \{0,1\} \to \mathbb{R}$. We solve for 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 the possibility that the advisor deems another action optimal for an individual than the individual does for himself. If the advisor instead evaluates an individual's expected payoff using *the individual's* prior, she will derive utility from him choosing an action that he believes is optimal even though she is convinced that he is making a mistake he later will come to regret. This distinction is essential and speaks to the notion of altruism that we apply: We say that a more altruistic advisor cares more about *her own valuation of* individuals' payoffs. We discuss this further in Section 7.

3.2 The Altruistic Authority

After observing the signal s, the authority (A, she) chooses between sending a public message m, after which individuals choose their actions, and engaging in *coercion*, whereby the authority mandates one action for all individuals, $a_A \in \mathbb{R}$. To capture the fact that coercion may be costly, we let the advisor's cost of coercion be given by q, where $q \geq c$.[11] For simplicity, we assume $c = 0$.

Under preference disagreement the authority's material (non-altruistic) payoff is given by $u_A(a, \theta) = -\int_{-\infty}^{+\infty}(a_i - \theta)^2 f(b_i)db_i - q \cdot \mathbf{I}(\text{coercion})$, and under opinion disagreement by $u_A = -q \cdot \mathbf{I}(\text{coercion})$.[12] Individuals' material

[11] Depending on the application, this cost may reflect the instrumental cost of active intervention or the authority's intrinsic aversion against removing the individuals' liberty to choose. In most applications discussed in Section 6, constraining the individual's liberty may be costlier than withholding information. To reflect this, we let $q > c$.

[12] Under opinion disagreement, all results remain under the (essentially equivalent) assumption that the authority's material payoff is given by $u_A(a, \theta) = -\int_0^1 (a_i - \theta)^2 g(\pi_i) d\pi_i -$

payoff functions, under both preference and opinion disagreement, 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, 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. Individuals' beliefs and strategies remain as specified above. We solve for PBE.

4 Libertarianism and Paternalism

4.1 Altruism and Truthful Advice

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

Proposition 1. *The impact of stronger altruism on the advisor's incentives to report truthfully depends on the nature of disagreement:*

- *Under preference disagreement, for any bias distribution $f(b)$ and lying cost c, there exists a threshold $\overline{\varphi}(c, \bar{b})$ such that an 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 an FRE exists if and only if $\varphi \leq \overline{\varphi}(c, g)$. For any $g(\pi) \notin \mathcal{G}$ an 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.

$q \cdot \mathbf{I}(\text{coercion})$.

Preference disagreement. From the setting with a single individual in Section 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 the advisor's altruism strengthens, her disagreement with each individual in the population lessens. 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 an FRE when the advisor is faced with a single individual with bias $\bar{b}$, the population's average preference bias. The proof of Proposition 1 establishes this formally; to see the logic of this result, w.l.o.g. suppose that $\bar{b} > 0$. Then the advisor always reveals the signal $s = 0$ to individual $\bar{b}$ but 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 $\bar{b}$.[13] Our result thus follows immediately from the discussion in Section 2: For strong enough altruism, truthful reporting of the signal $s = 1$ is incentive compatible and an FRE exists.

Opinion Disagreement. Since 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 $\pi_1 < \pi_A = 0.5 < \pi_2$, respectively. 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 a 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) it brings t_2. Figure 1 illustrates that this induces the advisor to misreport $s = 1$ in two regions of the (π_1, π_2) space, denoted $L1_{\text{left}}$ and $L1_{\text{right}}$, when the

[13] 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.

signal precision is $\gamma = 0.6$. An analogous argument shows that the advisor misreports $s = 0$ in the regions $L0_{\text{up}}$ and $L0_{\text{low}}$.[14]

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 $a_2(0)$, that is, when ($\text{TT}_{c=0}$) is satisfied. This occurs when opinion disagreement is minor; in particular, as noted in our example in Section 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) 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 Proposition 1 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 (the $L1_{\text{left}}$ region) and, potentially, when π_1 is close to 0.5 (the $L1_{\text{right}}$ region).

The non-monotonicity of the loss from lying to t_1 is due to two opposing effects. As π_1 increases, t_1 becomes 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 $s \in \{0, 1\}$. Since 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 $L1_{\text{left}}$ 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 $L1_{\text{right}}$, 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 $L0 \in \{L0_{\text{up}}, L0_{\text{down}}\}$.[15] Since the regions $L1_{\text{left}}$ and $L0_{\text{up}}$ are non-empty for all γ, the set of such distributions $\mathcal{G}$ is uncountable.

[14] 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, an FRE exists for any $\varphi \geq 0$ and $c \geq 0$.

[15] This is discussed in more detail in the proof of Proposition 1.

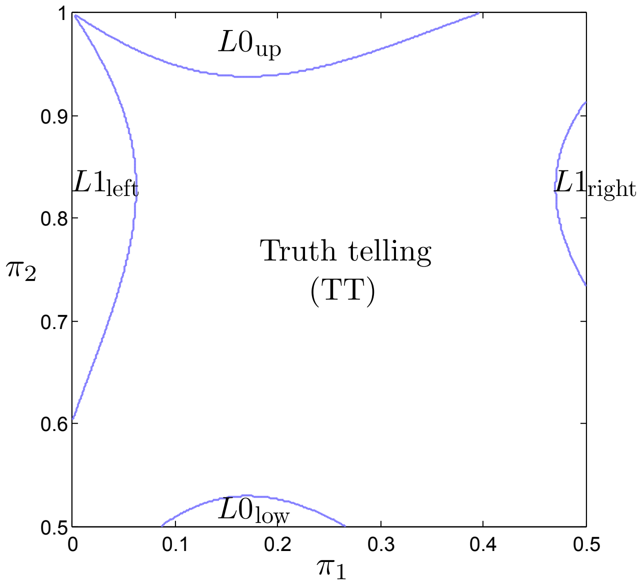

Figure 1: Incentives to misreport the signal, $\gamma = 0.6$.

For distributions $g \in \mathcal{G}$, an FRE fails to exist 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 the advisor 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, since lying protects the individuals from their own informed (in her view erroneous) choices. Since the population anticipates that the altruistic advisor lies, no informative communication can take place. Paradoxically, the population may thus prefer a disinterested advisor who can be trusted to tell the truth.

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 2. *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. An 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 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 $\varphi_{TT}(q, g)$, such that 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.[16] 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 each individual getting 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 an action other than an individual's own choice $a_i(s)$—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

[16] When $\bar{b}^2 = \overline{b^2}$, the preference distribution is degenerate; this case is thus equivalent to the single individual setting analyzed in Appendix D (and in the example in Section 2).

individuals, $a_{i,A}(s) = p(s) + \frac{\varphi}{1+\varphi} b_i$. When enacting a uniform mandate, the authority 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 level 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 1 shows that an FRE exists iff the authority's altruism is sufficiently strong. In the FRE, the individuals choose actions that the authority 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 all individuals the means to make as informed a choice as possible, and to give them the liberty to choose the actions they want.

Coercion becomes more viable, however, the weaker the 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 2 gives an approximate representation of these sets of (q, φ) and illustrates that coercion is less viable the higher q is.[17]

Opinion disagreement. Since 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 $a_i \neq a_A(s)$. Since the benefit accrues to the population, her valuation of this benefit increases with her altruism, φ, which makes coercion increasingly viable. Indeed, for $\varphi > \varphi_C(q, g(\pi))$, every equilibrium involves coercion with positive probability after at least one signal s. When altruism strengthens further, for $\varphi > \varphi_{CC}(q, g(\pi))$, coercion after both signals is the unique equilibrium; this is true even if an 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 the advisor is willing to bear the cost q to

[17] 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.

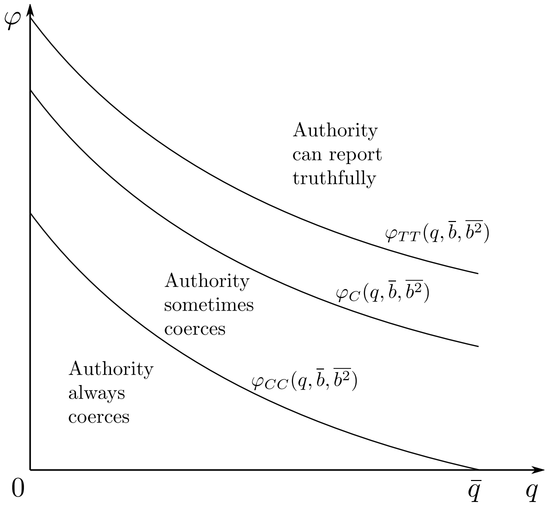

Figure 2: Equilibrium outcomes under preference disagreement, where $\bar{b}^2 \neq \overline{b^2}$.

mandate these actions, thereby protecting the individuals who would otherwise make erroneous (in her view) action choices. We say that the altruistic authority is *paternalistic*, since she constrains the individuals' liberty *out of affection*; that is, 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 an FRE exists for opinion distributions $g(\pi) \notin \mathcal{G}$. Figure 3 gives an approximate representation of these sets of (q, φ) and illustrates that coercion is less viable the higher q is.[18]

5 Targeted Advise and Targeted Mandates

Above we consider 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 in targeted communication, whereby she sends different messages to different individuals, and the authority to also enact targeted mandates.

[18] 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 uses coercion.

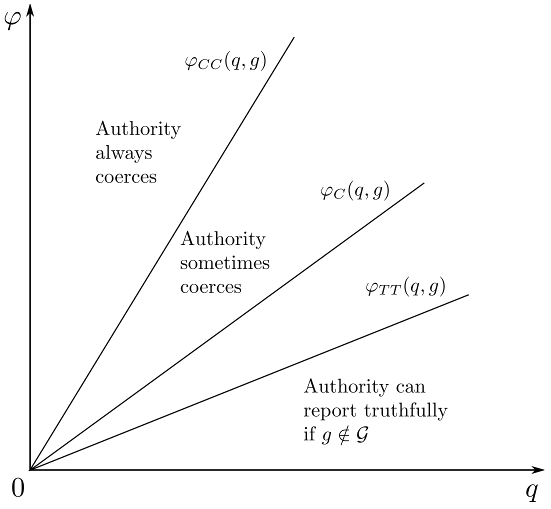

Figure 3: Equilibrium outcomes under opinion disagreement.

5.1 Altruism and Targeted versus Public Advice

Under targeted communication, the advisor privately observes a 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 the fact that lying can 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. 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 3. *The impact of stronger altruism on the relative credibility of public and private messages depends 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 shows that the advisor's message to individual i is credible iff ($\mathrm{TT}_{\mathrm{pr},\, c\,>\,0}$) 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, the advisor 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 1 shows that she can be credible whenever she would send a truthful message to an individual with bias $\bar{b}$, that is, when $\varphi \geq \overline{\varphi}(c, \bar{b})$. Consider $\overline{\varphi}(c, \bar{b}) > 0$.[19] For $\varphi < \overline{\varphi}(c, \bar{b})$, the advisor is more credible under private messages, since 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 does 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 biases more extreme 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 these authors' result that the relative credibility of private versus public messages depends on the preference

[19] If, instead, $\overline{\varphi}(c, \bar{b}) = 0$, public communication is (more) credible than targeted communication for all φ.

distribution, $f(b)$.[20] In this sense, we extend the authors' 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.[21] Lastly, we note that considering combined communication—where the advisor can send both public and private messages—does not improve advisor 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 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 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 ($L0_{\pi_1}L1_{\pi_2}$).[22]

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 $T_{\pi_1}L1_{\pi_2}/L1_{\text{right}}$, since she can provide credible private ad-

[20] More precisely, all the cases 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 $f(b)$.

[21] 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 $\varphi \geq \overline{\varphi}(c, \bar{b})$, for types $b_i \notin (\underline{b}(\varphi), \bar{b}(\varphi))$ the advisor is *disciplined* by the presence of others.

[22] $L1_{\pi_i}$ and $L0_{\pi_i}$ denote the incentives to misreport the signals 1 and 0 to individual i.

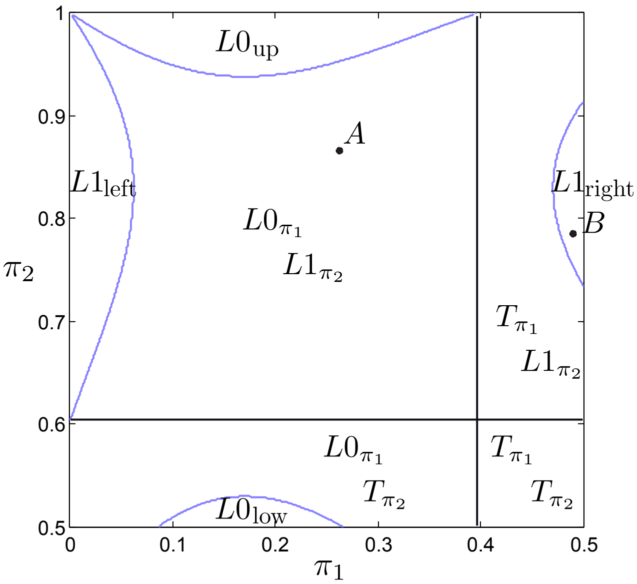

Figure 4: Incentives to misreport the signal under opinion disagreement, $\gamma = 0.6$.

vice to only half of the population. The converse can also arise, however: Consider $B \in L1_{\mathrm{right}}$. 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 $B \in L1_{\mathrm{right}}$.[23] Regardless of the strength of altruism, the advisor provides credible private advice to t_1 since 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 1 shows that the advisor is credible (to the entire 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

[23] Figure 4 plots all regions for $c = 0$. When $c > 0$, the size of the regions where the advisor misreports some signal(s) shrinks; however, the regions' relative positions largely remain as in Figure 4.

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 ($\varphi = 0$), in our setting with opinion conflict we obtain 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, $g(\pi)$.[24] 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 communication is private.[25] 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 can thus lead to 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 opinion distributions $g(\pi) \in \mathcal{G}$.[26]

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 the fact 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 $q \geq c = 0$; otherwise, all utility functions remain as specified above. A pure

[24] In this sense, we extend their results to the case of opinion conflict (and allow for populations with more than two individuals).

[25] 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 does not arise.

[26] More precisely, all cases discussed in these papers can arise, depending on the distribution $g(\pi)$. In the two-individual case depicted in Figure 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) occurs in $L1_{\text{left}}$ or $L0_{\text{up}}$, (iii) subversion of private communication occurs in $L1_{\text{right}}$ and $L0_{\text{low}}$, (iv) one-sided discipline occurs in $L0_{\pi_1}T_{\pi_2}/L0_{\text{low}}$ and $T_{\pi_1}L1_{\pi_2}/L1_{\text{right}}$, and (v) mutual discipline occurs in $LL/\{L1_{\text{left}}, L0_{\text{up}}\}$.

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 4. *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 decreases with the advisor's altruism.*
- *Under opinion disagreement, a single mandate is applied to all individuals whose action choices are restricted. The share of the population subjected to the (single) mandate increases with the advisor's altruism.*

When the authority can target her advice and mandates, she treats each individual i as she would if faced with a single individual with preference b_i or prior π_i. Our analysis in Section 2 (and Appendix D) 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, $q = 0$, the authority 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 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 toward 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 the authority mandates $a_A(s)$ for only a subset of the population, her mandate therefore applies to those 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. For strong enough altruism, she mandates the action $a_A(s)$ for all individuals with $\pi_i \neq \pi_A$.

6 Discussion

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. What determines whether a politician advocates one stance or the other? One commonly held view is that politicians differ in their valuations of personal liberty. Our results offer an alternative view.

Prediction (regulation versus laissez faire). *A benevolent politician wants to regulate issues that she deems matters of opinion but to allow individuals to make their own choices on issues that she deems matters of preference.*

Three examples illustrate that this view can be useful in understanding where and why paternalistic regulation emerges.

Example: Protection from physical self-harm. A benevolent authority opposes euthanasia if she fears that an individual who requests it is incapable of making decisions that are in his own long-term self-interest. Such differences in opinion (between the authority and individual) would arise, for example, if the authority believes that (i) the individual experiences pain which he, if denied euthanasia, will learn to endure over time and that (ii) he will then be happy that his request was denied. In contrast, a benevolent authority allows euthanasia if she believes that an individual who requests it is fully conscious of the consequences of his decision, and that he simply prefers to die.

In the few places where assisted suicide is legal—Belgium, Luxembourg, the Netherlands, Switzerland, and the U.S. states of Oregon, Washington, and Montana—the legal provisions are precise and 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 was doing, was capable of making the decision, and had requested death several times (Whiting (2002)).[27]

Example: 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, believing that the individual would change his mind if he were of a sound opinion.[28] This highlights the tenuous nature of paternalism: How can a government know better than a single individual what he 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, 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 for (only) certain individuals is restrictions on minors. Many governments require minors to have blood transfusions even when their religious beliefs forbid it while no such restriction is imposed on adults. Similarly, there is a legal drinking age, driving age, and voting age, an age of criminal responsibility, and so forth. Our framework suggests that these distinctions are justified by the view that an adult knows what he is doing but a minor may not; the government may, therefore, protect a minor from his own misjudgment.

This also relates to how opinions are formed. Over time, individuals may update their beliefs about, for example, the dangers of indoor tanning. If learning brings the individual's opinion closer to that of the authority—which

[27] 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).

[28] 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.

can occur when there exists an "objective truth," such as the actual risk of melanoma, which may correspond to π_A—then an individual's own decisions will improve over time. Consequently, the authority would want to target only those who, at a given point in time, hold (the most) erroneous beliefs; this is accomplished through minority regulation, since the law no longer applies when the individual becomes an adult.

Example: 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 (and, in her view, correct) belief over time. This can 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 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—such that the individual would be better off resisting and would otherwise come to regret his behavior in the future—criminalization can be a benevolent act to improve citizens' well-being (in this life or in a potential 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 the consumption of (adult) pornography (Time, 1969) and certain sexual acts between consenting adults, for example, of the same gender. Externalities arising from these activities are arguably small but, more to the point, regulation often explicitly relies 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. Support for legislation that prohibits same-sex marriage is indeed consistent with benevolence under the (morally paternalistic) belief that homosexuality is a disease that can be cured through heterosexual marriage, in sharp contrast to when an individual's choice of (the gender of) spouse is deemed a matter of preference.

7 Discussion of Assumptions

Altruism. We use the simplest formulation for our purpose. A few choices are noteworthy. First, the advisor (authority) evaluates the individuals' expected utilities using *her own prior*, π_A. This is essential. If she, instead, uses the individuals' priors, the results under opinion disagreement will be similar to those under preference disagreement. This is not a qualification of our results; on the contrary, it underscores our message. If the advisor evaluates an individual's expected payoff using his prior, then she will want the individual to take the action that he deems best for himself, even though she is convinced that the individual will later 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 advisor cares more about *her own valuation of* the individuals' payoffs.

Second, the advisor places the same weight on each individual. This is merely a simplification; suitable versions of our results are obtained as long as disagreement is non-negligible between the advisor and the individuals she cares (more) about.

Third, an alternative formulation would be to let the advisor place the weight $(1-\varphi)$ on her own non-altruistic payoff (instead of a weight of 1). Our main insights would continue to apply. Indeed, under opinion disagreement, the key feature driving our results—that an increase in φ raises the advisor's valuation of the individuals' payoffs relative to her own non-altruistic payoff, and in particular relative to the cost of lying (or coercion)—holds in both formulations. Then, when the advisor 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, the advisor'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 paternalistic under opinion disagreement.

The representativeness of the advisor. In the setting with preference disagreement, we consider general distributions, hence, all our results hold for cases when the advisor is representative of the population's median preference, that is, when $f(b)$ has a median equal to zero. In the setting with opinion disagreement, we assume that the advisor is representative of the median opinion. This assumption makes lying and coercion more unattractive than if the advisor (authority) can hold an extreme, unrepresentative opinion. Indeed, fix some opinion distribution $g(\pi)$ and consider an advisor who obtains $s = 1$. 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 greater loss from lying when sending a public message than an advisor with an extremely pessimistic opinion.

Similarly, a pessimistic authority may derive greater benefit from mandating her preferred action, because she deems the individuals' beliefs skewed in the optimistic direction and thus that intervention considerably improves 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 use coercion. Trivially, all our main results remain the same, with lower thresholds for lying and coercion.

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 has the authority to force the individuals' actions, she is *libertarian* under preference disagreement: She communicates truthfully and gives the individuals the liberty to choose. In contrast, under differences of opinion, the altruistic authority is *paternalistic*: Believing that she acts in the individuals' best interest, she forces an action other than the actions individuals would choose for themselves.

Whether coercion is perceived as justifiable and deemed socially desirable is thus intimately linked to whether a conflict is, consciously or subconsciously, framed in terms of preferences or opinions: The same prohibition is viewed as discrimination by some—who deem the individual decision as one governed by (non-aligned) preferences—and as protection by others—who deem it as one governed by (erroneous) opinions. Debates on paternalistic regulation are literally clashes of the ideas of what drives individual choices.

While a rigorous empirical test of the theory is beyond the scope of this paper, we believe that our theory, coupled with assumptions that are appropriate in the specific empirical context, can yield precise and testable empirical predictions. Suppose, for example, that we are willing to assume that Democrats are more likely than Republicans to perceive individual decisions that are of a sensitive religious nature—such as whether to marry a person of the same sex—as governed by preferences and that Republicans, in contrast, are more likely to view these decisions as governed by (erroneous) beliefs, or opinions. If both Democrats and Republicans are equally benevolent, the model predicts that Republicans should place more restrictions than Democrats on the individual freedom to make such decisions. Republican states should have fewer liberal gay marriage laws, for example. More generally, whenever the context of study suggests a plausible assumption on which politicians deem a given decision as governed by preference and which politicians deem it as governed by opinion, our theory offers a sharp empirical prediction of who advocates regulation.

A substantial issue that we hope will be analyzed in future work is externalities. In the current paper, one individual's action choice does not influence the outcomes of others. While this is the natural starting point when analyzing non-paternalistic regulation, interactions between paternalistic motives and the protection of third parties are left for future work.

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *American Economic Review*, 101(4): 1180–1210.

Allen, Mike. 2005. "Counsel to GOP Senator Wrote Memo On Schiavo Martinez Aide Who Cited Upside For Party Resigns." *Washington Post*, A01.

Becker, Gary S. 1974. "Theory of social interactions." *Journal of Political Economy*, 82: 1063–1093.

Benham, Kelley. 2005. “From ordinary girl to international icon.” *St. Petersburg Times.*

Carlin, Bruce, Simon Gervais, and Gustavo Manso. 2010. “Libertarian Paternalism, Information Sharing, and Financial Decision-Making.”

Che, Yeon-Koo, and Navin Kartik. 2009. “Opinions as Incentives.” *Journal of Political Economy*, 117(5): 815–860.

Cohen, Alma, and Liran Einav. 2007. “Estimating Risk Preferences from Deductible Choice.” *American Economic Review*, 97(3): 745–788.

Crawford, Vincent P., and Joel Sobel. 1982. “Strategic Information Transmission.” *Econometrica*, 50(6): 1431–1451.

Cutler, David M., and Richard Zeckhauser. 2004. “Extending the Theory to Meet the Practice of Insurance.” *Brookings-Wharton Papers on Financial Services*, 1–53.

Dessein, Wouter. 2002. “Authority and communication in organizations.” *Review of Economic Studies*, 69(4): 811–838.

Dworkin, Gerald. 2010. “Paternalism.” *The Stanford Encyclopedia of Philosophy.*

Einav, Liran, Amy Finkelstein, and Mark R. Cullen. 2010. “Estimating Welfare in Insurance Markets Using Variation in Prices.” *Quarterly Journal of Economics*, 125(3): 877–921.

Fang, Hanming, Michael P. Keane, and Dan Silverman. 2008. “Sources of Advantageous Selection: Evidence from the Medigap Insurance Market.” *Journal of Political Economy*, 116(2): 303–350.

Farrell, Joseph, and Robert Gibbons. 1989. “Cheap Talk with Two Audiences.” *American Economic Review*, 79(5): 1214–1223.

Goltsman, Maria, and Gregory Pavlov. 2011. “How to Talk to Multiple Audiences.” *Games and Economic Behavior*, 72(1): 100–122.

Greer, George W. 2000. “In re: the guardianship of Theresa Marie Schiavo, Incapacitated.” *Florida Sixth Judicial Circuit. Retrieved 2006-01-08.*

Hanson, Robin. 2003. “Warning labels as cheap-talk: why regulators ban drugs.” *Journal of Public Economics*, 87(9-10): 2013–2029.

Harris, Nonie M. 2001. “The euthanasia debate.” *J R Army Med Corps.*, 147: 367–370.

Hirsch, Alexander V. 2011. “Experimentation and Persuasion in Political Organizations.”

Lade, Diane C. 2012. “Booster seat bill wants older kids buckled up.” *Sun Sentinel.*

Lee, Samuel, and Petra Persson. 2011. “Circles of Trust.”

Locke, John. 1689. *Two Treatises of Government.*

Lovett, Ian. 2012. “Law on condoms threatens tie between sex films and their home.” *The New York Times.*

Mill, John Stuart. 1859. *On liberty.*

OpenCongress. 2012. “On Passage of the Bill (S. 1813 As Amended).” *http://www.opencongress.org/vote/2012/s/48.*

Spinnewijn, Johannes. 2012. “Heterogeneity, Demand for Insurance and Adverse Selection.”

Suber, Peter. 1999. “Paternalism.” *Philosophy of Law: An Encyclopedia*, 2: 632–635.

Thaler, Richard H., and Cass R. Sunstein. 2003. “Libertarian paternalism.” *American Economic Review*, 93: 175–179.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press.

The Washington Post. 2005. “GOP memo says issue offers political rewards.” *The Seattle Times.*

Van den Steen, Eric. 2006. “The limits of authority: motivation versus coordination.”

Van den Steen, Eric. 2009. “Authority versus Persuasion.” *American Economic Review*, 99: 448–453.

Whiting, Raymond. 2002. *A Natural Right to Die: Twenty-Three Centuries of Debate.* Westport: Greenwood Press.

Yardley, William. 2012. "Big sky, bright sun and melanoma." *The New York Times.*

Zuckerman, Laura. 2012. "Idaho state Senate panel nixes teen tan ban." *Reuters.*

A Appendix: Proofs

Proof of Proposition 1.

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

$$\begin{aligned}\mathbb{E}\left[U_A(a(1),\theta)|s=1\right] &= -p(1)\left[\int_{-\infty}^{+\infty}[(p(1)+b_i-1)^2+\varphi(p(1)-1)^2]f(b_i)db_i\right] \\ &\quad -(1-p(1))\left[\int_{-\infty}^{+\infty}[(p(1)+b_i)^2+\varphi p(1)^2]f(b_i)db_i\right] \\ &= -p(1)\left[(p(1)+\bar{b}-1)^2+\varphi(p(1)-1)^2\right] \\ &\quad -(1-p(1))\left[(p(1)+\bar{b})^2+\varphi p(1)^2\right]+\bar{b}^2-\overline{b^2}.\end{aligned}$$

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 $\mathbb{E}\left[U_A(a(1),\theta)|s=1\right] \geq \mathbb{E}\left[U_A(a(0),\theta)|s=1\right]$. This condition is equivalent to the truth telling condition when the advisor communicates with a single individual with preference bias $\bar{b}$, and can be rewritten as

$$\varphi \geq -1+\frac{2\bar{b}}{p(1)-p(0)}-\frac{c}{(p(1)-p(0))^2}. \qquad (\mathrm{TT1_{pr}})$$

Similarly, the advisor will report $s = 0$ truthfully whenever

$$\varphi \geq -1 - \frac{2\bar{b}}{p(1) - p(0)} - \frac{c}{(p(1) - p(0))^2}. \tag{TT0$_{\text{pr}}$}$$

Thus, truth telling is an equilibrium outcome if and only if the altruism level satisfies $\varphi \geq \overline{\varphi}(c, \bar{b}) = \max\{-1 + \frac{2|\bar{b}|}{p(1)-p(0)} - \frac{c}{(p(1)-p(0))^2},\}$. 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 an advisor who gets the signal $s = 1$, which gives her the posterior belief γ. If the advisor could differentiate messages between the types, she would always report $s = 1$ truthfully to t_1 with prior $\pi_1 \leq 1/2$; and she may 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 1, below we consider the properties of the loss and benefit functions (the analysis of the $s = 0$ case and region $L0$ is analogous).

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 $a_0 = p_{\pi_1}(0) = \frac{\pi_1(1-\gamma)}{\pi_1(1-\gamma)+(1-\pi_1)\gamma}$.[29] 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 derivatives with respect to π_1):

$$\begin{aligned} l''(\pi_1) &= 2\left[-a_1'(a_1 - \gamma) + a_0'(a_0 - \gamma)\right]' \\ &= 2\left[-(a_1')^2 - a_1''(a_1 - \gamma) + (a_0')^2 + a_0''(a_0 - \gamma)\right]. \end{aligned}$$

It can be shown that the following two equalities hold:

$$(a_1')^2 + a_1''(a_1 - \gamma) = \frac{\gamma^2(1-\gamma)^2}{[\pi_1\gamma + (1 - \pi_1)(1 - \gamma)]^4}\left[1 + 2(2\gamma - 1)(1 - 2\pi_1)\right] \geq 0,$$

[29] Similarly to the previously introduced notation, $p_\pi(s)$ represents the posterior belief of an individual with prior π about θ: $\Pr_\pi(\theta = 1|s)$.

$$(a_0')^2+a_0''(a_0-\gamma) = \frac{\gamma(1-\gamma)}{[\pi_1(1-\gamma)+(1-\pi_1)\gamma]^4}\left[\gamma(1-\gamma)+2(2\gamma-1)(\pi_1(1-\gamma)^2-(1-\pi_1)\gamma^2)\right].$$

Condition $\pi_1 \leq 0.5$ further ensures that the following two inequalities hold:

$$\pi_1\gamma+(1-\pi_1)(1-\gamma) \leq \pi_1(1-\gamma)+(1-\pi_1)\gamma,$$

$$\gamma(1-\gamma)\left[1+2(2\gamma-1)(1-2\pi_1)\right] > \gamma(1-\gamma)+2(2\gamma-1)(\pi_1(1-\gamma)^2-(1-\pi_1)\gamma^2).$$

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 $\pi_1 = 0$. Thus $l(\pi_1)$ 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 prior π_2. As before, let $\tilde{a}_1 = p_{\pi_2}(1)$ and $\tilde{a}_0 = p_{\pi_2}(0)$ denote the action choices of t_2 after messages 1 and 0, respectively. The potential benefit from misreporting is

$$b(\pi_2) = (\tilde{a}_1-\gamma)^2-(\tilde{a}_0-\gamma)^2.$$

Similar calculations yield that $b(\pi_2)$ is strictly increasing for $\pi_2 \leq \frac{\gamma^2}{\gamma^2+(1-\gamma)^2}$, $b(\pi_2)$ is strictly concave for priors $\pi_2 \geq \frac{\gamma^2}{\gamma^2+(1-\gamma)^2}$, $b'(1) < 0$, and $b(1) = 0$.

These properties 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 $b(\pi_2)$ since parabola $-(\pi-\gamma)^2$ has its peak at $\gamma > 0.5$. Figure 5 illustrates typical loss and benefit functions.

The difference in the advisor's expected payoffs from lying and truthful reporting is $\frac{1}{2}b(\pi_2)-\frac{1}{2}l(\pi_1)$.

These properties of $l(\pi_1)$ and $b(\pi_2)$ imply that the region $L1$, where $b(\pi_2) > l(\pi_1)$, has the form shown in Figure 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. 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 $b(\pi_2)$).

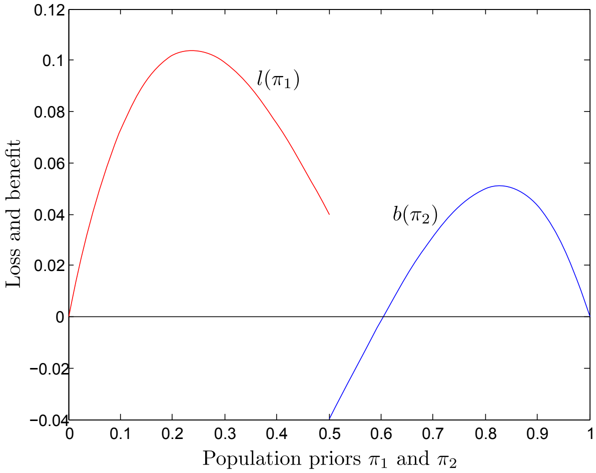

Figure 5: Loss and benefit functions from misreporting $s = 1$, $\gamma = 0.6$.

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 is 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:

$$\begin{aligned} b(\pi_2) &= (\tilde{a}_1 - \gamma)^2 - (\tilde{a}_0 - \gamma)^2 \\ &= \frac{(2\gamma - 1)\pi_2(1-\pi_2)}{(\pi_2\gamma + (1-\pi_2)(1-\gamma))(\pi_2(1-\gamma) + (1-\pi_2)\gamma)}(\tilde{a}_1 + \tilde{a}_0 - 2\gamma). \end{aligned}$$

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.[30]

[30] While we do not show analytically that $l(0.5) < \max_{\pi_2} b(\pi_2)$ *if only if* γ is below a

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 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 $g_1(\pi_1) = 0$ 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 $\pi_2 < 0.5$.[31] 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 the 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 1 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 Propositions 2 and 3 See online Appendix B.

specific threshold, computational exercises suggest that this is the case.

[31] 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].