Bayesian Explanations for Persuasion*

Andrew T. Little†

April 2022

Abstract

The central puzzle of persuasion is why a receiver would ever listen to a sender who they know is trying to change their beliefs or behavior. This paper provides a common formal framework for five approaches to solving this puzzle: (1) some messages are easier to send for those with favorable information (*costly signaling*), (2) the sender and receiver have partially aligned interests (*cheap talk*), (3) the sender messages can be checked (*verifiable information*), (4) the sender cares about perceptions of his competence/honesty (*reputation concerns*), and (5) the sender can "commit" to a messaging strategy (*Bayesian Persuasion*). To explore the relative value of these approaches, I discuss which provide insight into prominent empirical findings on campaigns, partisan media/propaganda, and lobbying. While models focusing on commitment have rapidly become one of the most common (if not the most common) theoretical approach to studying persuasion in political science and economics in the past decade, they are not particularly well-suited to explaining these phenomena.

*Many thanks to Elliot Lipnowski for insightful comments.

†Associate Professor, Department of Political Science, UC Berkeley. andrew.little@berkeley.edu.

Communication and persuasion are central to much if not most of politics. Democratic politicians try to persuade donors to donate and voters to vote for them. For autocrats, relatively free of institutional constraints, persuading others that they are strong leaders who should not be challenged may be even more central. Pundits aim to persuade an audience to adopt their views, or at least persuade an audience to continue paying attention to what they say. Ordinary citizens frequently talk to each other about politics—though certainly far less than political scientists—either to persuade or just for entertainment.

This note overviews formal approaches to persuasion, with as much common notation as possible. I generally focus on political applications, but much of the work cited comes from economics and can be applied more widely. The formal analysis restricts attention to models where the target of persuasion is fully rational, or Bayesian. That is, they understand the speaker's strategy and update their beliefs by Bayes' rule (in addition to standard sequential rationality requirements for decisions). In Section 5 I provide some discussion of when and why it is valuable to loosen this restriction.

One major benefit to the constraint of rational updating is that, combined with some general and reasonable conditions, it implies that persuasion should *not* be possible.[1] Intuitively, if the speaker (or sender) always wants the listener (or receiver) to take certain actions, and faces no constraints on what they say, they would always say whatever makes the listener do what they want. Knowing this, the receiver has no reason to pay attention. That is, starting with the assumption that the receiver is rational implies the pervasiveness of attempts at persuasion is a real puzzle to solve.

The bulk of the analysis shows how modifying the assumptions in this benchmark can provide such a solution. While inevitably non-exhaustive,[2] five modifications which are frequently used in applied models are: (1) some messages are easier to send for those with favorable information (*costly signaling*), (2) the sender and receiver have partially aligned interests (*cheap talk*), (3) the

[1]The title is a nod to Fearon (1995), which is structured in a similar way.

[2]One example I do not cover here is the role of mediators; see Kydd (2003) for an application to conflict and Salamanca (2021) for a recent theoretical discussion.

sender messages can be checked (*verifiable information*), (4) the sender cares about perceptions of his competence/honesty (*reputation concerns*), and (5) the sender can "commit" to a messaging strategy (*Bayesian Persuasion*). I then informally discuss "non-Bayesian" models of persuasion which are driven by receivers being less than fully rational in how they process information.

After describing the differences and commonalities of these models, I discuss how they have been applied to three substantive literatures on campaigns, partisan and state-controlled media, and lobbying. In each case I discuss when the assumptions and predictions of different models seem in line (or not) with empirical results. Finally, I reflect on the general theoretical insights provided by different approaches.

1 Trends

Part of the motivation for writing this is a sense that, following Kamenica and Gentzkow (2011), the number of papers on communication and persuasion in political science (and economics) that focus on persuasion via commitment has been dramatically rising, perhaps becoming the modal approach in applied theory papers. Figure 1 proves some suggestive evidence. The left panel shows the number of Google Scholar search hits for the phrases "Cheap Talk", "Costly Signaling", and "Bayesian persuasion" from 2012-2021.[3] The latter two increase steadily (perhaps entirely explained by more papers being indexed in general), while the latter goes from almost no hits to nearly as many as "Costly Signaling." As discussed further in section 7, informal perusing of the results indicates that results for costly signaling and cheap talk include many empirical papers while Bayesian persuasion returns almost entirely theoretical papers. This could potentially reconcile the lower overall number with the impression that this approach may be the modal approach in applied theory.

The right panel shows the number of citations to an emblematic paper from each tradition,

[3]The other two approaches are a bit harder to pin down with a single phrase, and tend to return fewer search results.

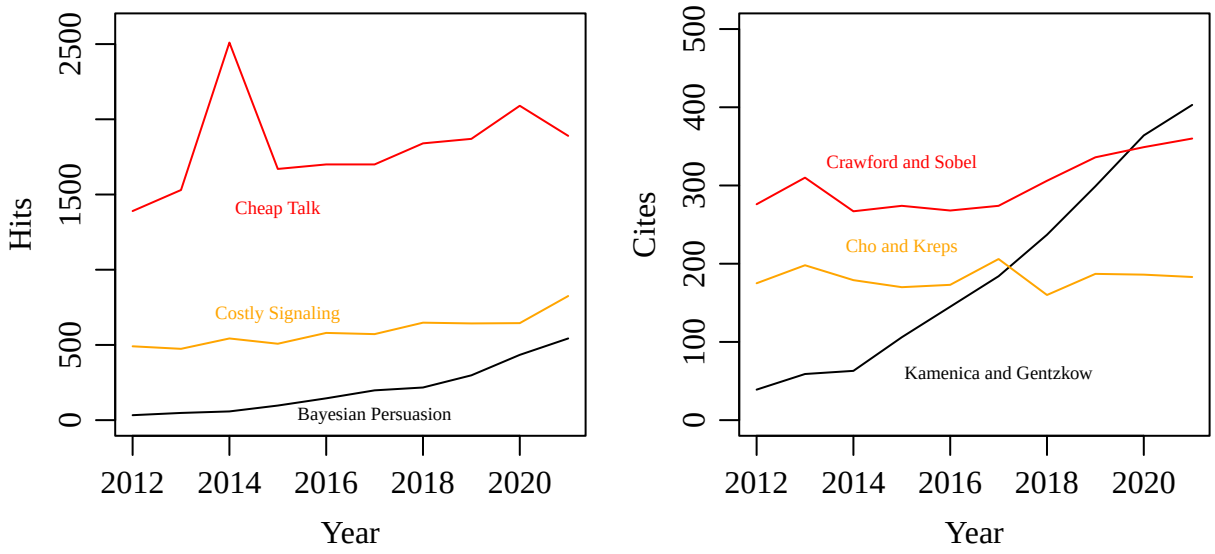

Figure 1: Google scholar search hits for kinds of model and citations (left panel) to influential papers (right panel)

Crawford and Sobel (1982) for cheap talk, Cho and Kreps (1987) for costly signaling,[4] and Kamenica and Gentzkow (2011) for Bayesian persuasion. Over the past decade, Kamenica and Gentzkow (2011) has overtaken these two influential papers in citations per year.[5]

Is the rapid rise of models where the sender can commit to a strategy because this approach is uniquely good at explaining persuasion?

The goal here is not to argue that that the answer to this question is clearly "no," let alone to argue that persuasion via commitment is never appropriate. Rather I aim to situate this kind of model in wider framework including other approaches.[6] Different assumptions are appropriate in different problems, and this note aims to give an overview of what kinds of assumptions are useful for different research questions about persuasion.

[4]It is hard to know the ideal paper here since the natural choice for an original paper using this style of model, Spence (1973), earns a huge number of citations for the point it makes about education, not about costly signaling per se.

[5]It is possible that part of this is driven by the fact that, for example, Crawford and Sobel (1982) is so well known that there is no need to cite it when using a cheap talk model.

[6]A working paper version of Kamenica and Gentzkow (2011) had more of a discussion of these connections, but the emphasis and organization here are different.

A point where I'll be more emphatic is linguistic. Several interlocutors report hearing others claim that any model of persuasion *must* include an assumption of commitment. Earlier, Milgrom and Roberts (1986*b*) propose reserving "persuasion" for models with verifiable information. While there is obviously value in precise in terminology, in this case the linguistic policing causes more confusion than it alleviates. Persuasion is a huge part of human interaction, so it is perfectly appropriate that there are several prominent ways to theorize about it. Any model where one actor tries to change what another thinks and does by some form of communication is one of persuasion. Now, on to the models.

2 The environment

Consider an interaction between a sender (he) and a receiver (she). In all versions the sender knows something which the receiver does not. This is why the receiver might listen to the sender and potentially be persuaded to act differently based on what he says.

Information and actions Formally, the sender first observes some information about a state of the world $\theta \in \Theta$ and picks a message $m \in M$. The sender strategy is the message sent as a function of the information, $m(\theta)$. The receiver observes m and then takes an action $a \in A$. The receiver strategy is the action taken as a function of the message, $a(m)$.

We can get most of our basic insights about communication with a binary state space, and binary messages.

Let $\theta \in \{0, 1\}$ represent the set of things the sender might know, where $\theta = 1$ is the state that the sender "wants the receiver to believe is true." Call $\theta = 1$ "good news" (the economy is booming, the politician is competent, etc.) Let $p = Pr(\theta = 1)$ be the prior belief that the information is good. The receiver does not observe θ, though she might learn something about it from the sender message.

The messages available are also $M = \{0, 1\}$.[7] For the kind of equilibria we study, it is reasonable to interpret a message of $m = 1$ as meaning "the news is good" and $m = 0$ meaning "the news is bad." Though, importantly, how the receiver interprets the message will depend on the sender strategy. For example, a sender strategy of $m(\theta) = \theta$ corresponds to honestly reporting the news, and if the receiver knows the sender uses this strategy she can take the message at face value. On the other hand, if the sender sometimes or always "lies" and sends $m = 1$ when in fact $\theta = 0$—i.e., says the news is good when it is in fact bad—a rational receiver will know to discount claims that things are going well.

Sometimes we will use a binary action set as well: $a \in \{0, 1\}$. In general, interpret $a = 1$ as doing what the sender wants; e.g., voting for the sender's preferred candidate or not joining a protest against him, or setting the policy he prefers. However, it will often make results tidier if we allow the receiver to take a continuous action between 0 and 1 ($a \in [0, 1]$), which can be interpreted as exerting effort on the sender's behalf, or just "supporting" the sender more generally. Another possible interpretation here is that the action is binary, but the receiver may be more or less likely to take it for random and exogenous reasons, and so the sender expected utility depends on the probability that the action is taken.

Utilities To formalize the notion that $a = 1$ is the action the sender wants the receiver take, assume the sender utility is strictly increasing in a. To start, assume that this is the only thing the sender cares about. That is, our first results will assume the sender utility does not *directly* depend on m, meaning messages are "cheap talk." The sender utility also does not depend on his type θ; he only cares about how beliefs about θ influence the sender action. Since the receiver knows the sender wants her to take a high action no matter what, we can say he has "transparent motives" (Lipnowski and Ravid, 2020).

[7] For most of the models there is little if any loss making the message space the same size as the type space.

Formally, say the sender has a *univariate monotone* utility if it can be written:

$$u_S = v(a) \tag{1}$$

where v is a strictly increasing function.

For the receiver, we will use the following utility:

$$u_R = -\theta(1-a)^2 - (1-\theta)a^2, \tag{2}$$

which captures the idea that the receiver wants her action to be close to θ with increasing marginal costs. For example, if the news is whether a policy under consideration would help the economy, then the receiver may want to implement the policy ($a = 1$) if the truth is that the policy is good ($\theta = 1$), and not implement it otherwise. This quadratic formulation is for convenience; the main results generally hold, (sometimes with caveats) as long as the optimal receiver action is increasing in her belief that the sender has good news ($\theta = 1$).

Solution Concept To capture the notion that we want to explain persuasion of a rational receiver (by a rational sender), we will impose standard Perfect Bayesian Equilibrium requirements. Namely, that (1) the sender messaging strategy ($m^*(\theta)$) is optimal for each possible θ, given the receiver strategy ($a^*(m)$), (2) the receiver action is optimal for each message m and the receiver beliefs $Pr(\theta|m)$, and (3) the beliefs $Pr(\theta|m)$ are formed by Bayes' rule, consistent with the messaging strategy (when possible).

Preliminary Analysis The optimal receiver behavior given her beliefs is straightforward. Let $Pr(\theta = 1|m)$ be the probability that the state is 1 given m. When we restrict a to be either 0 or 1, her expected utility for picking $a = 0$ is $-Pr(\theta = 1|m)$ and for picking $a = 1$ is $-(1 - Pr(\theta =$

$1|m)$). So, it is optimal to pick $a = 1$ if and only if:

$$Pr(\theta = 1|m) \geq 1 - Pr(\theta = 1|m)$$

or $Pr(\theta = 1|m) \geq 1/2$. If this is met with equality, either choice gives equal utility, and if it is strict there is a unique best response.

When we allow for a continuous choice on $[0, 1]$, the receiver utility is strictly concave in a, and her optimal action meets the first-order condition:

$$\frac{\partial u_R}{\partial a} = 2aPr(\theta = 1|m) - 2(1 - a)Pr(\theta = 1|m) = 0,$$

which is solved by $a^*(m) = Pr(\theta = 1|m)$. This utility function captures the idea that the sender "wants the receiver to think that $\theta = 1$," as the action taken (and hence the sender utility) is equal the probability she assigns to the sender having good news.

3 Persuasion is impossible

What does it mean for "communication" or "persuasion" to happen? While we will ultimately focus on whether the sender can change the receiver action, first consider whether he can systematically change her beliefs. In one sense he certainly can not. Consider the average posterior belief that the state is good, $\mathbb{E}_m[Pr(\theta = 1|m)]$, where the subscript highlights the fact that we are averaging over the messages. Then:

Theorem 1. *(No persuasion on average) In any PBE,*
(i) The average posterior belief that $\theta = 1$ is equal to the prior belief that $\theta = 1$ ($\mathbb{E}_m[Pr(\theta = 1|m)] = p$), and
(ii) If the posterior belief that $\theta = 1$ is strictly higher for one message that is sent in equilibrium ($Pr(\theta = 1|m = i) > p$ and $Pr(m = i) > 0$ for some $i \in \{0, 1\}$), then the posterior belief must be

strictly lower for the other message ($Pr(\theta = 1|m = i) < p$ for $j = 1 - i$).

In the main text I provide intuition for this and later results; formal proofs are in the appendix. The first part states that the average posterior belief about the whether the sender has good news must equal the prior. This is essentially just a statement of the law of iterated expectations, which in this context is often called the Martingale property of belief updating.[8] The binary restrictions are not needed for this result; it holds for any type space and message set (and any utility function). One way to think of this is from the perspective of the receiver before learning the message: if she expects her belief that $\theta = 1$ after observing m will be higher than her prior on average, then she should just adjust her prior upwards in light of this before receiving the message.

The second part follows immediately from the first; if one message increases the belief that $\theta = 1$, the other message must lower this belief to ensure the average posterior is equal to the prior.

Importantly, Theorem 1 does not mean the receiver will never learn anything from the message: it could be the case that the belief goes up for one message and down for the other. What it does formalize is a sense in which a sender can never systematically increase the beliefs of a receiver (on average).

Does this result put a nail on the coffin of persuasion? Not necessarily. Persuasion is typically not just about when the sender influences *beliefs*, but when he can influence *actions*.[9] A natural way to define whether persuasion-to-action occurs is to compare what the receiver does compared to a benchmark action with "no communication." In words, can the sender use the opportunity to speak to the receiver sender to get her to do something closer to what he wants?

With continuous choices, the benchmark action is just the prior belief: $a_0 = p$. With binary choices, this benchmark action is 1 if $p > 1/2$, 0 if $p < 1/2$, and can be 0 or 1 if $p = 1/2$. To simplify a later result, call the benchmark action 1 for the knife-edged case where $p = 1/2$:

[8]The proof requires just a hair of additional work to deal with the possibility off-path beliefs, which need not be formed by Bayes' rule, but don't affect the average belief precisely because they are off-path.

[9]Another way to put this—coined by Scott Ashworth in some long-ago tweet—is that we aren't just interested in "persuasion that" (beliefs changing as a function of m), but "persuasion to" (actions changing as a function of m).

Definition The *benchmark action* is $a_0 = p$ with the continuous receiver choice, and $a_0 = \mathbf{1}_{p \geq 1/2}$ with a binary receiver choice.

We are now ready to formally define persuasion

Definition A message is *persuasive* if and only if $a^*(m) > a_0$. A *persuasive PBE* is a PBE where a persuasive message is sent with strictly positive probability.

Thinking in causal inference terms, we are making a comparison between an outcome (the action) under one "treatment" (hearing message m) vs. a "control" condition of receiving no new information. A message is persuasive if and only if there is a strictly positive treatment effect.

This definition combined with Theorem 1 leads to a stark result (see Theorem 2 in Lipnowski, Ravid and Shishkin, 2019, for a more general statement):[10]

Theorem 2. *With a univariate monotone sender utility, there is no persuasive PBE.*

The intuition is simple: if a message was persuasive and led to a better action than the benchmark, the sender would always want to send this message. But if they always send the persuasive message, the belief upon observing it must be the same as the benchmark action, meaning it can't be persuasive.

Taking stock We started by observing much of politics involves people talking in order to persuade others to do things. Using a very bare-bones setup with reasonable assumptions, we arrived at a theorem which states that this can never happen. What gives?

More constructively, what might we change about this setting to make persuasion possible? Most formal theories of persuasion can be placed into three categories based on the answer to their question.

[10]Note that in the knife-edged case where $p = 1/2$, the proof relies on the definition of the benchmark action to be 1. If we set it to 0, then any informative equilibrium will be persuasive. Still the above result would hold for any $p \neq 1/2$.

First, we could change something about the utility functions or information structure. As we will see in sections 4.1-4.4, most classic models of communication can be cast in this fashion. To contrast with the last approach, we can call these models of *Bayesian persuasion without commitment*.

Second, we can allow the sender to "commit" to a messaging strategy, as discussed in section 4.5.[11] These are often called models of "Bayesian persuasion," which is deeply misleading as the previous class of models also study persuasion among actors who are Bayesian. Following Gehlbach (2021, Ch. 8) and to contrast the models in section 4.1-4.4, call these models of *Bayesian persuasion with commitment*.

Finally, whether we allow for commitment or not, we can loosen the "Bayesian" part, assuming that either the sender or receiver has non-standard beliefs. As previously promised, we will not formalize this class of explanations but discuss when we may (or may not) want to use this approach in section 5.

4 Bayesian persuasion, with and without commitment

We can think of the different kinds of communication/persuasion models commonly used in political science as different ways of getting around these "impossibility" results.

As above, the description of these models aims to include just enough formalization to convey the main ideas. Appendix B contains more complete descriptions of the equilibria, which are pretty standard (but potentially useful for teaching purposes.)

4.1 Costly Signaling

In costly signaling models, the sender utility is a function of the message they send, and also depends on their type. A simple version is to add a cost for sending $m = 1$ which depends on the

[11] In a sense this could be considered "changing the information structure," but given the prominence of this approach in recent work it is worth placing in a separate category.

sender type,

$$u_S = a - mc_\theta$$

where $c_1 < c_0$, and allow the receiver to take continuous actions. That is, sending message $m = 0$ is free; often this corresponds to "not sending the signal." Sending $m = 1$ incurs a cost, which is higher for the bad news type than the good news type. In the context of costly signaling models, we often call the good news type the "high" (or "strong") type, and the bad news type the "low" (or "weak") type, so I'll use that language for this subsection.

For either binary or continuous receiver actions, this creates the possibility for a persuasive ("separating") PBE where $m^*(\theta) = \theta$ if $c_1 \leq 1 \leq c_0$.[12] In this PBE, the high type gets $1 - c_1 > 0$ for sending the message, so it is worth sending if it reveals to the receiver that $\theta = 1$, and the receiver infers that $\theta = 0$ otherwise. However, since $1 - c_0 < 0$, it is not worth it for the sender with bad information to send $m = 1$ even if this "tricks"the receiver into taking the favorable action ($a = 1$).

While we won't fully characterize the equilibria to this game (or later ones), an important observation which I'll return to is that there is also a "pooling" PBE where the sender always chooses $m = 0$ and the receiver always picks $a = p$.[13]

Even in the separating/persuasive equilibrium, Theorem 1 still applies: the average belief and hence action taken is still p. In fact, if the action is continuous and $v(a)$ is linear in a, the average sender utility is lower in the persuasive PBE than the "pooling on $m = 0$" PBE.[14] So the high (good news) type can benefit from communication relative to the low (bad news) type, but the ability to send costly signals can not make the receiver systematically think that the sender has good news.

[12]This condition may seem restrictive, but if we allow for the message to be a continuous choice $m(\theta) \geq 0$, there is always an continuum of equilibria of this form where $m(0) = 0$ and $c_1 \leq m(1) \leq c_0$.

[13]This is certainly true if the "off-path" belief upon observing $m = 1$ is that $Pr(\theta = 1) \leq p$. Much theoretical work on signaling attempts to identify when such beliefs are reasonable Cho and Kreps (e.g., 1987), a topic beyond the scope here.

[14]This need not be true in the binary action case. If $p < 1/2$, then the sender expected utility can be higher in the persuasive PBE than the pooling PBE since the $\theta = 0$ type gets utility 0 in either case and the $\theta = 1$ type gets $1 - c_1 > 0$ in the persuasive PBE.

Similar arguments work if the cost of sending messages does not depend on type, but the receiver type affects the benefit they receive for taking the action.[15] "Burning money" may serve as a signal that the sender cares a lot about the receiver taking a higher action, which may mean the receiver wants to do so (typically this relies on common interest, which is discussed more in the following section, see Austen-Smith and Banks, 2000).

Combining, costly signaling models are an appropriate way to measure persuasion when (1) there are real costs associated with the action taken, and (2) these relative costs and benefits of the action depend on the information held by the sender.

Often the first part is clear: (political) advertising is costly (Milgrom and Roberts, 1986*a*), as are donations (Gordon and Hafer, 2005; Schnakenberg and Turner, 2021), getting educated (Spence, 1973). Some actions are costly at least in time, like protesting (Lohmann, 1993) or lobbying. Other kinds of costs may be less concrete but still real, like the cost of backing down after making a threat (Fearon, 1997) or signing a treaty (Hollyer and Rosendorff, 2011).

Costs and benefits depending type are often plausible too.[16] Earning a degree requires less effort for smarter and more diligent student. Those who care more about a policy change are more willing to protest/lobby. Etc. However, these models are generally less suited to communication which is just talk.

4.2 Cheap Talk

Models with cheap talk are common in political science as well. To capture the notion that talk is cheap, these models do not include the message itself in the utility function, but typically allow the senders with different types to have different preferences, potentially allowing for honest

[15] We can assume the cost is constant but the benefit differs by type by letting $u_S = b_\theta a - mc$. If $b_0 < 1 < b_1$ there is a separating PBE with $m^*(\theta) = \theta$ by an identical logic.

[16] See Petrova (2008) for a related model where costs are not correlated with type, but are heterogeneous, and so there can be an equilibrium where lower cost types misrepresent their signal while higher cost types do not, making favorable information partially persuasive (as it is still more likely to be sent when the news is good).

communication (Crawford and Sobel, 1982; Green and Stokey, 1980).[17]

The most common way for communication with cheap talk to arise is if there is some degree of common interest among the sender and receiver (though see below for an important exception). In the extreme, if the sender utility is the same as the receiver utility, there can be full communication and persuasion in equilibrium: if both actors want to match the action to the state, the sender has a strong incentive to tell the truth.

To capture this intuition but retain the idea that the sender also tends to want the receiver to take high actions, let the sender utility be:

$$u_S = ba + (1 - b)u_R$$

where $b \in (0, 1)$ scales the magnitude sender bias (relative to the common interest captured by the u_R term). Regardless of whether actions are continuous or binary, there can be a PBE where the sender reveals his information where $m = \theta$ if the bias term is relatively small. Formally, suppose the sender is always "honest" ($m^*(\theta) = \theta$) and so the receiver takes an action equal to his message ($a^*(m) = m$). It is immediate that, given the receiver does what he says, a sender with $\theta = 1$ wants to send $m = 1$. The sender also prefers to send $m = 0$ when $\theta = 0$ if the utility from the sender taking the "right" action $(1 - b)$ is higher than the utility from lying and inducing an action of 1 (b) This is true when $b \leq 1/2$.

In words, one way for cheap talk models to "get around" the impossibility results by allowing for sufficient common interest between the sender and receiver that the benefits of communication to the sender outweigh the costs of revealing less favorable information.

Cheap talk models like this are an effective way to model persuasion in cases where the sender doesn't only care about getting the receiver to take particular actions, but also has some common interest with the receiver (Crawford and Sobel, 1982). This is a natural assumption in many set-

[17] Of course, any speech takes time, which entails some opportunity cost. What really matters for this class of model is that, among the messages which might be sent, there is no difference in direct cost.

tings like policy-making (Gilligan and Krehbiel, 1987) and bureaucratic implementation (Gailmard and Patty, 2012), where all want policies which are objectively "good", but different actors have slightly to widely different views of what is ideal.

Another possibility, which tends to arise in multivariate environments, is that there is some dimension on which the receiver is indifferent, and hence willing to do what the sender wants (Battaglini, 2002; Chakraborty and Harbaugh, 2010; Schnakenberg, 2015; Lipnowski and Ravid, 2020). Intuitively, if the sender wants the receiver to do multiple things, it may be credible for him to say "among the things I want you do to, X is more important to me than Y."

As a quick formal example, return to the benchmark model, except now the receiver takes two actions a_1 and a_2, and let her utility be the same with $a = a_1 + a_2$. Let the sender utility be $a_1 + ra_2$, where $r > 0$. I.e., there are two kinds of actions the receiver can take which are interchangeable from her perspective (and she now wants the sum of the actions to match the state), but if $r < 1$ the sender prefers her to take action 1 and if $r > 1$ he prefers action 2. If the sender has private information about r, he can effectively say "regardless of how much action you choose, please do it on dimension d", where $d \in \{1, 2\}$. I.e., he can't persuade her to take a higher sum action, but can persuade her to do the kind of action he prefers. This general idea can work even if the sender has transparent motives, meaning his preferences do not depend on his private information (Lipnowski and Ravid, 2020).

4.3 Verifiable Information

While costly signaling and cheap talk models allow for persuasion by changing preferences, and other possibility is to change the information structure. On approach in this vein is to assume that messages are *verifiable* or *hard information*. (In fact, this is precisely the feature that Milgrom and Roberts (1986*b*) use to distinguish games of persuasion from cheap talk models, though here I stick with using persuasion to refer to a wider class of approaches.)

A simple way to model this is to change the message space to $M = \{0, 1, \emptyset\}$, and to assume

that upon observing θ the messages the sender can actually choose are $M(\theta) = \{\theta, \emptyset\}$, where we can interpret $\emptyset$ as "saying nothing." I.e., the sender can either reveal the truth or keep quiet.[18]

The assumption that the sender is incapable of sending a lie may seem extreme. One way to interpret this is that the sender is really an "intermediary" who receives a report from a subordinate, and is deciding whether to pass it on to a higher-up. We also need not interpret this literally: similar results arise if the receiver gets a separate signal which indicates whether the message was "correct" (see Dziuda and Salas, 2018, for a nice example with partial lie detection).

With any monotone sender utility and either continuous or binary actions, there is a persuasive PBE where the type with good news reveals this ($m^*(1) = 1$) and the type with bad news either admits it or says nothing ($m^*(0) = 0$ or $m^*(0) = \emptyset$). The receiver is persuaded by seeing $m = 1$ since only the good type sends this message. The key difference between this model and the benchmark is that the sender type with bad information can't pretend there is good news because it is verifiable. He can only admit the news is bad or say nothing, both of which reveal that $\theta = 0$.

This argument becomes more striking when there is a larger number of states and messages. Suppose the state can be any number between 0 and 1 with uniform probability, the sender can either reveal the state or say nothing, and the receiver takes an action equal to her average belief about the state. Consider a potential equilibrium where the sender reveals the truth if and only if it is better than average ($\theta \geq 1/2$). Then upon hearing nothing, the receiver knows the state is between 0 and $1/2$, so the average belief is $1/4$. But then a sender who's information is just slightly unfavorable—formally, between $1/4$ and $1/2$—would rather reveal it than keep quiet. This "unraveling" argument leads to the conclusion that there must be full revelation of information.

This is great for a receiver who wants more information, but what about the sender? The possibility of communication is good for the sender when they have good news to share, but this gain is offset by the loss when there is bad news, since the sender can't prevent the receiver from

[18]The choice set could also be written $M = \{\theta, \{0, 1\}\}$, i.e., the sender can't lie in the sense that if the the state is 0 they can either say "the state is 0" or "the state is 0 or 1."

learning that $\theta = 0$. So it is indeterminate whether being able to persuade the receiver is generally good for the sender.

The assumption of verifiable information is reasonable in some scenarios and not others. Persuasive speech often takes the form of saying "here is *why* you should do what I want," which the receiver can evaluate by seeing if the argument "seems reasonable." Much political communication comes along with data or other forms of evidence to back it up.

Verification need not be perfect, and it could also be costly for one of the actors. For example, the receiver may need to pay a cost to "audit" the signal (e.g., Austen-Smith and Wright, 1992). Sending inherently informative messages may be costly as well. One may need to do some research to acquire persuasive information (Austen-Smith and Wright, 1992; Patty, 2009). We can also think of the decision to hold an election/referendum as a (very costly!) way to provide a noisy signal of the popularity of the incumbent or some policy (Little, 2017*a*).

4.4 Reputation

Often times senders don't care as much about persuasing receivers of some decision-relevant information ("this policy is a good idea"), but of their own competence ("I am the type of expert who knows what policies are a good idea"). Such reputation concerns can increase or decrease the prospect of persuasive communication.

Suppose θ corresponds to the competence of the sender, which then affects the "quality" of information he receives. There is an additional state of the world $\omega \in \{0, 1\}$, with $Pr(\omega = 1) = q$; one common interpretation is that this corresponds to whether a proposed policy change will be a successful. The sender knows his competence and gets a signal which might be informative about the state. When $\theta = 0$ the sender gets an uninformative message $s = \emptyset$, and when $\theta = 1$ the sender observes $s = \omega$.

The receiver now takes two actions $a = (a_\theta, a_\omega)$. Think of a_ω as the "policy choice" and a_θ as the "competence assessment." Suppose the utilities over both have a similar quadratic form as

above, and so the best responses are $a^*_\theta(m) = Pr(\theta = 1|m)$ and $a^*_\omega(m) = Pr(\omega|m)$.

The sender utility is:

$$u_S = ra_\theta + (1 - r)a_\omega$$

Where r scales the "reputation conerns" relative to the "policy concerns".

If $r = 0$ (only policy concerns), there is no persuasive PBE by a similar logic to the main model; if either message led to a better policy everyone would send it, rendering the message uninformative.

If $r = 1$ (only reputation concerns), there is a PBE where the competent type sends $m = s$ and the incompetent type sends $m = 1$ with probability p and $m = 0$ with probability $1 - p$. In this equilibrium, the receiver learns nothing about the sender competence—this is precisely what makes the incompetent type indifferent between the two message. However, since the sender is more likely to say the policy is good when this is true, the receiver does get some information (or, is persuaded) on this dimension.

The appendix also contains an analysis of the intermediate case where $r \in (0, 1)$, which unsurprisingly blends features of these extremes. There is always a PBE with some learning/persuasion about both the expert competence and the ideal policy, but a fair amount of lying.

Models with reputation for competence are natural when studying political actors whose reputations hinge on views of their ability (Backus and Little, 2020).[19] As the example above shows, this may lead to some persuasion about the state of the world the receiver fundamentally cares about, but ensuring full communication typically requires adding some features discussed above like partial alignment of interest or partially verifiable messages (Ottaviani and Sorensen, 2006; Backus and Little, 2020). Reputation also plays an important role in repeated interactions, where being seen as an honest or competent type can incude a sender to listen to future messages (Sobel, 1985; Kuvalekar, Lipnowski and Ramos, 2022).

[19]Within political science, models with reputation concerns more often focus on beliefs about the competence of decision-makers themselves, and how incentives to "pander" or "posture" can distort policies away from what the decision-maker knows to be ideal (e.g., Canes-Wrone, Herron and Shotts, 2001).

Reputation concerns can also reduce honest communication if the receiver has a strong prior belief and hence doubts sources of contrary information (e.g., Prendergast, 1993; Morris, 2001; Gentzkow and Shapiro, 2006), and are a poor incentive for honesty when senders can have more moderate information, which can give incentives to exaggerate to appear more informed (Ottaviani and Sorensen, 2006; Backus and Little, 2020).

4.5 Commitment

Finally, we explore how persuasion might be possible (or expanded) by allowing the sender to *commit* to a messaging strategy. Communication models with this feature have been around for a while, e.g., Bénabou and Tirole (2002) can be interpreted as studying a "rational" self committing to a messaging strategy to a "deciding self" who has self-control problems.[20] However, use of this approach exploded in popularity following Kamenica and Gentzkow (2011), who provided a general treatment and several techniques to make analyzing models with this assumption tractable (see also Rayo and Segal, 2010).

One way to think about the commitment assumption is that prior to observing θ, the sender picks a messaging strategy $Pr(m = 1|\theta)$, which is then "implemented" upon the realization of θ. Write this $\mu = (\mu_0, \mu_1)$, where $\mu_i = Pr(m = 1|\theta = i)$. The receiver observes this strategy as well as the result (m). Another way to think of this is that the sender is setting up an "experiment" which maps the true state to a probability distribution over outcomes, and this outcome will be revealed to the receiver (see Luo, 2018, for an example which emphasizes this interpretation).[21]

A PBE to the model with commitment is then a μ, $a^*(m)$, and $Pr(\theta|m)$ such that μ maximizes the sender expected utility given $a^*(m)$, $a^*(m)$ is optimal given μ and $Pr(\theta|m)$, and $Pr(\theta|m)$ is formed by Bayes' rule when possible.

[20] While not published until after Kamenica and Gentzkow (2011), an early version of Gehlbach and Sonin (2014) was circulated in 2008 if not earlier.

[21] Yet another interpretation is that the sender picks m after observing θ, but does not need to pick a sequentially rational choice for both realizations of θ, but rather can pick a messaging strategy which maximizes his *ex ante* utility. I.e., the game form is not changed, but the solution concept is.

Continuous action, weakly concave utility Before getting to the "standard" case of Bayesian persuasion with commitment, which focuses on binary receiver actions, to contrast this kind of model to others it will be instructive to consider when commitment makes persuasion possible in the continuous action setting.

First suppose the sender utility is linear and strictly increasing in a, i.e., can be written $v(a) = \alpha + \beta a$ for some $\alpha \in \mathbb{R}$ and $\beta > 0$. For any messaging strategy μ, Theorem 1 tells us that $\mathbb{E}_m[Pr(\theta = 1)] = p$. By the linearity of the expectation operator, the expected utility for any messaging strategy is $\alpha + \beta p$.[22]

So, with linear utilities, the sender is indifferent between any messaging strategy, and any μ can be part of a PBE. Technically speaking, there can be a persuasive PBE in this setting: e.g., there is nothing to stop the sender from picking $m = \theta$ (or $\mu_0 = 0$, $\mu_1 = 1$), in which case both messages are fully informative about the state (and $m = 1$ is persuasive). However, such equilibrium selection under indifference is an unsatisfying way to explain the pervasiveness of attempts to persuasion. Further, since any strategy is an equilibrium, this will not be a useful benchmark to bring to applied models.[23]

There are many other kinds of v functions to consider beyond the linear case, but one natural family is set set of strictly concave v functions. That is, there are "diminishing marginal returns to persuasion." In this case, there is a sharp negative result that there can be no persuasive PBE. This follows from the fact that any informative messaging strategy induces a mean-preserving spread of the posterior belief, which is bad for a sender with concave preferences. See Appendix B.5 for

[22] Formally,

$$\begin{aligned}\mathbb{E}_m[v(a^*(m;\mu))] &= Pr(m=1;\mu)(\alpha + \beta Pr(\theta = 1|m;\mu)) + Pr(m=0;\mu)(\alpha + \beta Pr(\theta = 1|m;\mu)) \\ &= \alpha + \beta(Pr(m=1;\mu)Pr(\theta = 1|m=1;\mu) + Pr(m=0;\mu)Pr(\theta = 0|m=0;\mu)) \\ &= \alpha + \beta p\end{aligned}$$

[23] A quick observation which I've never seen formalized: take the utility from the cheap talk model, with any $b < 1$, meaning the sender puts some weight on the receiver making a good decision. With commitment the sender would choose full information revelation, since the average action is the same no matter what and so he might as well help the receiver make a good choice.

a formal statement or Remark 1 in (Kamenica and Gentzkow, 2011) for a more general result.

Binary Action What if the action taken by the receiver is binary?

Recall that with binary actions, the receiver can take action 1 if and only if $Pr(\theta = 1|m) \geq 1/2$.[24]

If $p \geq 1/2$, then in an uninformative equilibrium, $a^*(m) = 1$ for both m, giving the highest possible utility to the sender. In other words, there is no need for persuasion here, since the receiver is already going to do what the sender wants. So, there is no value to committing to an informative information structure; why risk screwing up a good thing?[25]

The interesting case is $p < 1/2$. Here, in an uninformative equilibrium the receiver would always take action $a = 0$. If the sender always reported the news honestly ($m = \theta$), then the receiver would take action $a = 1$ upon observing $m = 1$ and $a = 0$ upon observing $m = 0$. This gives the sender expected utility $p > 0$. So, a fully informative equilibrium is better than an uninformative one for the sender. (And the receiver! More on this below.)

However, the sender can do even better. To see why, suppose the sender also occasionally sends sends $m = 1$ when $\theta = 0$ ($\mu_0 > 0$). If such lies are sufficiently rare, then the receiver will still be pretty confident that $\theta = 1$ upon observing $m = 1$, and takes action 1. So the sender can increase how often he gets the sender to pick $a = 1$ and hence improves his payoff by sometimes lying. The optimal strategy for the sender is to lie as much as possible, subject to the constraint that the receiver still takes the favorable action when observing $m = 1$.

When are models of this form appropriate to study persuasion? Much attention gets focused on the commitment assumption; see Gehlbach (2021, ch. 8) for a valuable discussion. A common (and often appropriate) justification is that the choice being studied is not an individual act of persuasion, but setting up a more durable institution (biased media, and advertising campaign) which will

[24]The general argument easily extends to the case of a more general threshold.

[25]Recall we assumed that if $p = 1/2$ the receiver takes action $a = 1$. If not there are a few cases to consider here but no real insight.

produce a bias to information.[26] It can also be appropriate in individual settings where the sender is choosing to "commission a study" of some form which where the receiver will inevitably observe the result. If the choice of how much bias to introduce is made before the sender even knows the revelation of information, this is equivalent to picking the messaging strategy which maximizes *ex ante* utility.

Equally important for commitment to lead to informative communication and persuasion is that the sender has convex preferences over the receiver beliefs. Loosely speaking, if the sender just wants the receiver to generally have a high belief that $\theta = 1$, then being able to commit to a messaging strategy is not very helpful, if at all. However, there are many political situations where the sender has a more specific goal, in particular when the receiver is taking a binary (or discrete) action if her belief is above a threshold (e.g., "is the politician good enough for me to vote for her?" or "are citizens mad enough for a politician to pick a reform.") In this case, it also important the the sender have a pretty good sense of what the receiver prior belief is and the threshold for action, otherwise the situation is better approximated by just wanting to think $\theta = 1$.

There are scenarios where these assumptions are reasonable. Further, both assumptions can be loosened. On partial commitment, see Luo and Rozenas (2018) for an interesting application to electoral fraud and monitoring, and Lipnowski, Ravid and Shishkin (2019) for a more general analysis. Titova (2020) shows that under some assumptions outcomes that can be achieved with commitment can also arise in an analogous model with a continuous type space and verifiable information. Other paper study persuasion of an audience wit ha heterogenous prior or preferences (Gehlbach and Sonin, 2014; Alonso and Câmara, 2016), or a single listener with private information (Kolotilin et al., 2017).

[26] A drawback of this interpretation is that if a choice like media bias intended to influence a wide range of people, in a wide range of situations, over a non-trivial period of time, it is likely that their "thresholds" for doing the action that the sender wants are highly heterogeneous. It possible for uncertainty about the threshold to remove the possibility of successful persuasion with commitment; e.g., if the distribution of thresholds is uniform on $[0, 1]$, the model becomes equivalent to the continuous action case with linear utility over receiver beliefs, where persuasion is not possible. However, with many realistic distributions a sender can still benefit from commitment with heterogeneous thresholds or prior beliefs (e.g., Alonso and Câmara, 2016; Kolotilin et al., 2017), though the scope for doing so is often diminished.

5 Non-Bayesian Persuasion

We have restricted the formal analysis to persuasion with Bayesian or rational receivers. Two common justifications for focusing on such an approach are that:

1. All models must make some assumptions about beliefs, and without placing any restrictions on beliefs we often (if not usually) can't make concrete predictions about behavior. Particularly in situations with high stakes, political actors can have strong incentives to form correct beliefs. Even if this doesn't lead to perfectly correct beliefs, it is a natural place to start.

2. If the goal is to explain why persuasion is attempted/succeeds, it is "too easy" to get this result by assuming that receivers believe whatever senders say.

I think point 1 is very important, but does not imply we should never consider models with incorrect beliefs. However, it does mean that deviations from correct beliefs should have some combination of theoretical and empirical justification. We generally want to focus on particular classes of deviations from correct beliefs, which can be motivated by empirical results (ideally ones that isolate the particular mistake being assumed). These goals usually work well together, since good empirical work isolating incorrect beliefs usually entails a particular kind of mistake which can be included in our theories (see Benjamin, 2019, for a recent overview).

My views on point 2 are more mixed. It is true that if we are going to go through the work of setting up and solving formal models, we generally want to go beyond the obvious; one does not need fancy math or solution concepts to say "persuasion happens because people believe what they are told." And we are often too quick to resort to blaming puzzling behavior and beliefs—particularly among those we disagree with—on irrationality. Still, as raised in response to point 1: if we restrict ourselves to deviations from rationality with strong empirical grounding, this will constrain us from overly reductive behavioral explanations. Further, even if some of our explanation for persuasion is that some people are credulous, there are non-obvious implications for when this leads to more or less information transmission (Kartik, Ottaviani and Squintani,

2007), spillover effects to non-credulous individuals Little (2017*b*), and whether the ability to lie freely to a partially credulous audience actually helps senders (Little and Nasser, 2018).[27] Other recent papers explore how non-Bayesian updating can expand the amount of persuasions possible with commitment, potentially leading to the receiver always doing what the sender wants (Levy, Moreno de Barreda and Razin, 2022).

6 Empirical Insights

In this section I discuss some prominent classes of empirical findings in light of the analysis above.[28] These are all enormous literatures and the aim here is not to be anywhere near comprehensive, but to give a general sense of some key findings and how the different theoretical approaches shed light on when we should and not expect to find persuasive effects.

6.1 Campaigns

A useful place to start is the study of political campaigns: advertising, canvassing, mailers, etc. Political campaigns typically have transparent motives: they aim to get citizens to vote in in a particular way. Theorems 1 and 2 can be seen as a formal representation of an old conventional wisdom that campaigns should have minimal effects; see Kalla and Broockman (2018) for a recent overview and meta-analysis in support of this conclusion.

However, other forms of campaigning do appear to have small to moderate effects on voters. Research with credible designs finds evidence of some persuasion in many contexts, such as television advertising in the United States (e.g., Huber and Arceneaux, 2007; Spenkuch and Toniatti,

[27] See also Horz (2021). A more subtle version of this bias is that receivers update properly on the information which they do see, but struggle to make inferences about information which is hidden, they can be persuaded on average (Enke, 2020; Eyster and Rabin, 2005; Jin, Luca and Martin, 2021). Broockman and Kalla (2022) provide evidence about the persuasive power of biased media which is easiest to interpret in this light. Another example of persuasion occuring due to incorrect Bayesian updating is Mullainathan, Schwartzstein and Shleifer (2008).

[28] One large strand of work I avoid here is lab and survey experiments on persuasion (see Druckman, 2021, for a recent overview with more of an experimental focus).

2018), mail and phone contact in Italy (Kendall, Nannicini and Trebbi, 2015), banners on streets in Spain (Esteban-Casanelles, 2020), and clientalist appeals in Benin (Wantchekon, 2003).[29]

The models in section 4 can provide insight into when and why we might expect campaigns to succeed at persuasion. While "senders" in this setting generally just want voters to behave in a certain way, trying to identify and emphasize common interest is often used as a persuasive strategy (Broockman and Kalla, 2016; Druckman, 2021).

Campaigning is also costly, either in money or in the time of volunteers. However, a straightforward application of costly signaling arguments may be somewhat tenuous, as candidates with more favorable information (they are aligned with the voters, the other candidate is corrupt, etc.) may not face much lower costs. Still, this could be plausible if it is easier to get volunteers or raise money for better candidates.

While exaggeration and stretching of the truth are common, much campaign information is also at least partially verifiable. Kendall, Nannicini and Trebbi (2015) are explicit about this, emphasizing that their intervention gives voters "hard and verifiable information" about the positions and valence of candidates. Outside of experimental settings, campaigns that have better information to share (their candidate does have a good record; the other candidate truly has a dodgy past) may invest more in advertising to share this.

These effects can be magnified if some voters are credulous and take messages at face value, though given the common wisdom that politicians and campaigns will do whatever they can to get votes this seems like a domain where such effects should be limited. A more plausible non-Bayesian mechanism (to me, at least) is that campaigning may simply raise certain "considerations" to the top of voters minds (Zaller et al., 1992). This could explain why campaign effort tends to increase steadily up to elections, as this is when effects will be most salient/decay the least (Acharya et al., 2019).

[29]These studies focus on persuading voters to select certain candidates for parties. Persuasive effects are often stronger when trying to persuade voters to turn out, particularly with canvassing (Gerber and Green, 2000) or social pressure (Gerber, Green and Larimer, 2008).

Finally, reputation and commitment seem to have limited explanatory power in this context. Starting with commitment, given the high pressure of campaigns it seems unlikely that politicians can commit to refrain from exaggerating (if not outright lying) if doing so would increase their chance of winning. It is possible that a reputation for honesty restrains more extreme lies, but I don't know of any work that explores the implications of this class of models for campaigns.

6.2 Partisan and State-Controlled Media

A key difference between campaigns and partisan media/state-controlled media is that outlets themselves typically care not just about winning a current election, but about their reputation in the longer term, as well as gaining viewership. While this difference will affect interpretation, the overall empirical results from this literature are similar, with well-identified studies typically finding small to medium sized persuasive effects.

A major strand of this literature in the US studies the effect of Fox News on political attitudes and voting, by studying the rollout of the channel (DellaVigna and Kaplan, 2007), where Fox lies in channel listings (Martin and Yurukoglu, 2017), or experimentally inducing a change in media diet (Broockman and Kalla, 2022). These studies consistently find that watching Fox makes views more conservative in their attitudes and voting. Partisan news may exert a particularly strong influence if the slant of the provider is less known, e.g., in local news (Martin and McCrain, 2019). Research on state-controlled media, particularly in more autocratic contexts, also consistently finds persuasive effects (e.g., Adena et al., 2015; Enikolopov, Petrova and Zhuravskaya, 2011).

As alluded above, many of the dynamics discussed with respect to campaigns apply here, though media outlets care more about their reputation, either for intrinsic reasons or to increase viewership/advertising revenue (Petrova, 2011). This provides some restraint even for sources with a strong desire to influence beliefs, as they can't persuade people who find the news too biased to be useful (Gehlbach and Sonin, 2014). However, short-term incentives to appear competent may lead media to cater to viewer's prior beliefs (Gentzkow and Shapiro, 2006).

The commitment assumption may also be more reasonable when looking at the long-term strategies of partisan media outlets (whether privately owned or government-controlled). High-level decision-makers are typically not dictating how every individual story should be covered, but give broader guidance, which can loosely be thought of as "how often to lie when the state is bad." The tradeoffs captured in models with commitment seem reasonable and important here: the more one lies, the less persuasive it is when they say their favored side is doing well. Further, more manipulation of information renders news outlets less informative, which may decrease viewership (Gehlbach and Sonin, 2014). However, this general insight could also obtain from a model focused on reputation concerns: the more the outlet skews their coverage, the less they may be trusted in the future. Distinguishing between commitment and reputation could be a fruitful avenue for future theoretical and empirical work.

Another important puzzle to explain is that partisan media seems to not only have persuasive power with individual messages, but can shift beliefs on average and over long periods of time. Recall that in all of the standard models the average posterior belief that $\theta = 1$ must be equal to the prior (Theorem 1). For example, in the reputation or verifiable information models, the sender can persuade the receiver *when the information is favorable*, but not on average.

While it is probably possible to come up with ways to explain persistent persuasion on average, a simple explanation for this that that views don't fully adjust for the bias in news sources (see Broockman and Kalla, 2022, for direct evidence of this in the case of heavy Fox News consumers). Brundage, Little and You (2022) discuss evidence that this kind of *selection neglect* could explain the persuasiveness of partisan media.

6.3 Lobbying

An immediate difference between the media examples and lobbying is that the latter may serve a non-informational purpose. In fact, as reviewed by Bombardini and Trebbi (2020), most empirical work on lobbying focuses more on quid-pro-quo explanations. Even with a research design

that pins down that more time lobbying or more donations causes politicians to vote in a way the lobbyist wants, we can't directly infer that the politician beliefs about ideal policy were affected.

However, this is some evidence consistent with an lobbying as persuasion drawing on ideas from the cheap talk, costly signaling, and verifiable information approaches. The fact that lobbyists spend most of the time with those who are ideologically aligned has a natural interpretation from the cheap talk approach, where common interest facilitates communication (Grossman and Helpman, 2001), though other theories could explain this pattern as well (e.g., Snyder Jr, 1991; Hall and Deardorff, 2006). Hirsch et al. (2021) develop a model where a key role of lobbyists is to screen clients to put them in touch with politicians with common interest, which helps the politician learn about the merit of client requests. Gordon and Hafer (2005) find that lobbying may reduce enforcement activity, consistent with a costly signaling model. As with the previous examples, it seems plausible that lobbyists care about their reputation or can persuade more by committing to a messaging strategy, but I am unaware of any empirical work that explores particular predictions of these styles of model.

7 Theoretical Insights

Now to editorialize a bit: each of the four models without commitment contain a fundamental insight about communication which is easy to apply to many social settings:

- Costly signaling models tell us that people may engage in seemingly wasteful/inefficient/harmful behavior if it shows off that they are a "type" who is willing to do this.
- Cheap talk models tell us that communication is easier when the sender and receiver have more closely aligned incentives. When they are unaligned, "talk is cheap" and should be ignored.
- Verifiable information models tell us that when information is easy to check, people may

reveal unfavorable information (from their own perspective) if keeping quiet would make them look even worse

- Reputation models teach us that when speakers care about a reputation for competence, they may tell the truth in order to seem smart. However, senders may have incentives to cater to the prior belief of their audience or make stronger claims than are warranted by their information.

A valuable thing about these classes of models is that these forces are straightforward to apply to specific examples of political communication and persuasion. This is useful for applied theorists; e.g., when we see people taking seemingly inefficient actions to try and induce others to do what they want, it is natural to build on a costly signaling model to explain this. It is also useful for empirical scholars who want to motivate their analysis or interpret their results without having to write an original formal model.

If I may issue a challenge to proponents of using models of Bayesian persuasion models that rely on a commitment assumption, it isn't that they need to do more to justify the commitment assumption. It's that I'm not sure what clean and widely applicable insight comes from these models.

Some possibilities are:

1. That senders who can commit to a messaging strategy can do better (and make receivers better off too)? Ok, but the idea that actors who could commit to behavior could make themselves individually or collectively better off is an insight that comes from several canonical models (the prisoners' dilemma, trust games, bargaining models, etc.). Further, as discussed in Appendix B.5, this result is not true in all persuasion settings; often the ability to commit to a strategy leads to less communication and persuasion.

2. That commitment is move valuable when senders have convex preferences over receiver beliefs? This is an interesting technical observation, but I have a hard time seeing how to

apply it to real political settings.

3. That senders face a tradeoff between sending favorable messages more often and the persuasive value of favorable messages? This is an important insight, but one that also shows up in models without commitment assumptions.

Perhaps because of this shortcoming, it seems like models of Bayesian persuasion with commitment have had much more impact on the theoretical literature on persuasion than the empirical literature.[30] There are a few ways to interpret this; it could just be that that theoretical innovations naturally influence theorists before spreading into empirical work. However, it has been over a decade since the early influential papers on persuasion with commitment, which strikes me as a long gestation period. A more cynical view is that the draw of this class of models is a combination of the fact they are nice to work with technically, and provide something "new" to write papers about, regardless of the usefulness in explaining real world phenomena. I think these factors explain a lot of why so much contemporary formal theory of communication focuses on the role of commitment assumptions, but am open to being persuaded.

References

Acharya, Avidit, Edoardo Grillo, Takuo Sugaya, Eray Turkel et al. 2019. "Dynamic Campaign Spending." *URL: http://stanford. edu/avidit/campaigns. pdf* .

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *The Quarterly Journal of Economics* 130(4):1885–1939.

Alonso, Ricardo and Odilon Câmara. 2016. "Persuading voters." *American Economic Review* 106(11):3590–3605.

[30]Let alone outside of academia; a recent Op-Ed on the topic put it mildly when saying "Bayesian persuasion hasn't been widely embraced by policymakers." https://www.nytimes.com/2022/05/25/opinion/bayesian-persuasion.html.

Austen-Smith, David and Jeffrey S Banks. 2000. "Cheap talk and burned money." *Journal of Economic Theory* 91(1):1–16.

Austen-Smith, David and John R Wright. 1992. "Competitive lobbying for a legislator's vote." *Social choice and Welfare* 9(3):229–257.

Backus, Matthew and Andrew T Little. 2020. "I don't know." *American Political Science Review* 114(3):724–743.

Battaglini, Marco. 2002. "Multiple referrals and multidimensional cheap talk." *Econometrica* 70(4):1379–1401.

Bénabou, Roland and Jean Tirole. 2002. "Self-confidence and personal motivation." *The quarterly journal of economics* 117(3):871–915.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics: Applications and Foundations 1* 2:69–186.

Bombardini, Matilde and Francesco Trebbi. 2020. "Empirical models of lobbying." *Annual Review of Economics* 12:391–413.

Broockman, David and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Broockman, David and Joshua Kalla. 2022. "The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with Fox News viewers.".

Brundage, Matthew, Andrew Little and Soo You. 2022. "Selection Neglect and Political Beliefs.". manuscript.

Canes-Wrone, Brandice, Michael C Herron and Kenneth W Shotts. 2001. "Leadership and pandering: A theory of executive policymaking." *American Journal of Political Science* pp. 532–550.

Chakraborty, Archishman and Rick Harbaugh. 2010. "Persuasion by cheap talk." *American Economic Review* 100(5):2361–82.

Cho, In-Koo and David M Kreps. 1987. "Signaling games and stable equilibria." *The Quarterly Journal of Economics* 102(2):179–221.

Crawford, Vincent P and Joel Sobel. 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society* pp. 1431–1451.

DellaVigna, Stefano and Ethan Kaplan. 2007. "The Fox News effect: Media bias and voting." *The Quarterly Journal of Economics* 122(3):1187–1234.

Druckman, James N. 2021. "A Framework for the Study of Persuasion." *Annual Review of Political Science* 25.

Dziuda, Wioletta and Christian Salas. 2018. "Communication with detectable deceit." *Available at SSRN 3234695* .

Enikolopov, Ruben, Maria Petrova and Ekaterina Zhuravskaya. 2011. "Media and political persuasion: Evidence from Russia." *American Economic Review* 101(7):3253–85.

Enke, Benjamin. 2020. "What you see is all there is." *The Quarterly Journal of Economics* 135(3):1363–1398.

Esteban-Casanelles, Teresa. 2020. "The Effects of Exposure to Electoral Advertising: Evidence from Spain.".

Eyster, Erik and Matthew Rabin. 2005. "Cursed equilibrium." *Econometrica* 73(5):1623–1672.

Fearon, James D. 1995. "Rationalist explanations for war." *International organization* 49(3):379–414.

Fearon, James D. 1997. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of Conflict Resolution* 41(1):68–90.

Gailmard, Sean and John W Patty. 2012. "Formal models of bureaucracy." *Annual Review of Political Science* 15:353–377.

Gehlbach, Scott. 2021. *Formal models of domestic politics*. Cambridge University Press.

Gehlbach, Scott and Konstantin Sonin. 2014. "Government control of the media." *Journal of public Economics* 118:163–171.

Gentzkow, Matthew and Jesse M. Shapiro. 2006. "Media Bias and Reputation." *Journal of Political Economy* 114(2):280–316.

Gerber, Alan S and Donald P Green. 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American political science review* 94(3):653–663.

Gerber, Alan S, Donald P Green and Christopher W Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American political Science review* 102(1):33–48.

Gilligan, Thomas W and Keith Krehbiel. 1987. "Collective decisionmaking and standing committees: An informational rationale for restrictive amendment procedures." *Journal of Law, Economics, & Organization* 3(2):287–335.

Gordon, Sanford C and Catherine Hafer. 2005. "Flexing muscle: Corporate political expenditures as signals to the bureaucracy." *American Political Science Review* 99(2):245–261.

Green, Jerry R and Nancy L Stokey. 1980. A two-person game of information transmission. Technical report Discussion paper.

Grossman, Gene M and Elhanan Helpman. 2001. *Special interest politics*. MIT press.

Hall, Richard L and Alan V Deardorff. 2006. "Lobbying as legislative subsidy." *American Political Science Review* 100(1):69–84.

Hirsch, Alexander V, Karam Kang, B Pablo Montagnes and Hye Young You. 2021. "Lobbyists as gatekeepers: Theory and evidence.".

Hollyer, James R and B Peter Rosendorff. 2011. "Why do authoritarian regimes sign the convention against torture? Signaling, domestic politics and non-compliance." *Signaling, Domestic Politics and Non-Compliance (June 1, 2011)* .

Horz, Carlo M. 2021. "Propaganda and skepticism." *American Journal of Political Science* 65(3):717–732.

Huber, Gregory A and Kevin Arceneaux. 2007. "Identifying the persuasive effects of presidential advertising." *American Journal of Political Science* 51(4):957–977.

Jin, Ginger Zhe, Michael Luca and Daniel Martin. 2021. "Is no news (perceived as) bad news? An experimental investigation of information disclosure." *American Economic Journal: Microeconomics* 13(2):141–73.

Kalla, Joshua L. and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112(1):148–166.

Kamenica, Emir and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101(6):2590–2615.

Kartik, Navin, Marco Ottaviani and Francesco Squintani. 2007. "Credulity, lies, and costly talk." *Journal of Economic theory* 134(1):93–116.

Kendall, Chad, Tommaso Nannicini and Francesco Trebbi. 2015. "How Do Voters Respond to Information? Evidence from a Randomized Campaign." *American Economic Review* 105(1):322–

53.

URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20131063*

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk and Ming Li. 2017. "Persuasion of a privately informed receiver." *Econometrica* 85(6):1949–1964.

Kuvalekar, Aditya, Elliot Lipnowski and Joao Ramos. 2022. "Goodwill in communication." *Journal of Economic Theory* p. 105467.

Kydd, Andrew. 2003. "Which side are you on? Bias, credibility, and mediation." *American Journal of Political Science* 47(4):597–611.

Levy, Gilat, Inés Moreno de Barreda and Ronny Razin. 2022. "Persuasion with correlation neglect: a full manipulation result." *American Economic Review: Insights* 4(1):123–38.

Lipnowski, Elliot and Doron Ravid. 2020. "Cheap talk with transparent motives." *Econometrica* 88(4):1631–1660.

Lipnowski, Elliot, Doron Ravid and Denis Shishkin. 2019. "Persuasion via weak institutions." *Available at SSRN 3168103* .

Little, Andrew T. 2017*a*. "Are non-competitive elections good for citizens?" *Journal of Theoretical Politics* 29(2):214–242.

Little, Andrew T. 2017*b*. "Propaganda and credulity." *Games and Economic Behavior* 102:224–232.

Little, Andrew T and Sherif Nasser. 2018. "Unbelievable lies." *Unpublished manuscript, University of California, Berkeley* .

Lohmann, Susanne. 1993. "A signaling model of informative and manipulative political action." *American Political Science Review* 87(2):319–333.

Luo, Zhaotian. 2018. "Discriminatory Persuasion." *Available at SSRN 3075042* .

Luo, Zhaotian and Arturas Rozenas. 2018. "Strategies of election rigging: trade-offs, determinants, and consequences." *Quarterly Journal of Political Science* 13(1):1–28.

Martin, Gregory J and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review* 107(9):2565–99.

Martin, Gregory J and Joshua McCrain. 2019. "Local news and national politics." *American Political Science Review* 113(2):372–384.

Milgrom, Paul and John Roberts. 1986*a*. "Price and advertising signals of product quality." *Journal of political economy* 94(4):796–821.

Milgrom, Paul and John Roberts. 1986*b*. "Relying on the information of interested parties." *The RAND Journal of Economics* pp. 18–32.

Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109(2):231–265.

Mullainathan, Sendhil, Joshua Schwartzstein and Andrei Shleifer. 2008. "Coarse thinking and persuasion." *The Quarterly journal of economics* 123(2):577–619.

Ottaviani, Marco and Peter Normal Sorensen. 2006. "Reputational Cheap Talk." *RAND Journal of Economics* 37(1):155–175.

Patty, John W. 2009. "The politics of biased information." *The Journal of Politics* 71(2):385–397.

Petrova, Maria. 2008. "Inequality and media capture." *Journal of public Economics* 92(1-2):183–212.

Petrova, Maria. 2011. "Newspapers and parties: How advertising revenues created an independent press." *American Political Science Review* 105(4):790–808.

Prendergast, Canice. 1993. "A Theory of Yes Men." *American Economic Review* 83(4):757–770.

Rayo, Luis and Ilya Segal. 2010. "Optimal information disclosure." *Journal of political Economy* 118(5):949–987.

Salamanca, Andrés. 2021. "The value of mediated communication." *Journal of Economic Theory* 192:105191.

Schnakenberg, Keith E. 2015. "Expert advice to a voting body." *Journal of Economic Theory* 160:102–113.

Schnakenberg, Keith E and Ian R Turner. 2021. "Helping friends or influencing foes: Electoral and policy effects of campaign finance contributions." *American Journal of Political Science* 65(1):88–100.

Snyder Jr, James M. 1991. "On buying legislatures." *Economics & Politics* 3(2):93–109.

Sobel, Joel. 1985. "A theory of credibility." *The Review of Economic Studies* 52(4):557–573.

Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87(3):355–374.

Spenkuch, Jörg L and David Toniatti. 2018. "Political advertising and election results." *The Quarterly Journal of Economics* 133(4):1981–2036.

Titova, Maria. 2020. Persuasion with verifiable information. Technical report UCSD Working Paper.

Wantchekon, Leonard. 2003. "Clientelism and voting behavior: Evidence from a field experiment in Benin." *World politics* 55(3):399–422.

Zaller, John R et al. 1992. *The nature and origins of mass opinion*. Cambridge university press.

A Proofs

Proof of Theorem 1

Proof Write the average posterior belief as:

$$\mathbb{E}_m[Pr(\theta = 1|m)] = Pr(m = 0)Pr(\theta = 1|m = 0) + Pr(m = 1)Pr(\theta = 1|m = 1)$$

If $Pr(m = i) = 1$ for some $i \in \{0, 1\}$, then by Bayes' rule $Pr(\theta = 1|m = i) = p$. The belief upon observing the other message ($m = j \neq i$), $Pr(\theta = 1|m = j)$ is unconstrained, but $Pr(m = j) = 0$, so the result holds regardless of the off-path belief.

If $Pr(m = i) > 0$ for both $i \in \{0, 1\}$ then $Pr(\theta = 1|m = i)$ must both be formed by Bayes rule, and hence:

$$\begin{aligned}\mathbb{E}_m[Pr(\theta = 1|m)] &= Pr(m = 0)Pr(\theta = 1|m = 0) + Pr(m = 1)Pr(\theta = 1|m = 1) \\ &= Pr(\theta = 1, m = 0) + Pr(\theta = 1, m = 1) \\ &= Pr(\theta) = p\end{aligned}$$

For part ii, if $Pr(m = i) > 0$ and $Pr(\theta = 1|m = i) > p$, then if $Pr(\theta = 1|m = j) \geq p$ the average posterior belief would be strictly higher than p, a contradiction. ∎

Proof of Theorem 2

Proof With continuous actions $a_o = p$. In a persuasive PBE, there must exist an $i \in \{0, 1\}$ such that $Pr(\theta = 1|m = i) > p$. By theorem 1, this implies $Pr(\theta = 1|m = j) < p$ for the other message $j = 1 - i$. Since $a^*(m) = Pr(\theta = 1|m)$, for the sender to be playing a best response this implies the sender always picks $m^*(\theta) = i$. But if the sender always sends i, then consistency requires that $Pr(\theta = 1|m = i) = p$, a contradiction.

For the binary action case, if $p \geq 1/2$, then $a_0 = 1$, and so it is immediate that there can be no persuasive message.

If $p < 1/2$, then $a_0 = 0$. If there is a persuasive message $m = i$, then it must be the case that $Pr(\theta = 1|m = j) < 1/2$, and hence $a^*(j) = 0$. So for the sender to play a best response, it must be the case that $m^*(\theta) = i$, and hence $Pr(\theta = 1|m = i) = p < 1/2$ and $a^*(i) = 0$, hence i is not persuasive. ∎

B Formal analysis of models in section 4

In this section we provide a complete description of the PBE of the models in section 4.

For this general analysis we will need notation for potential mixed strategies. In general, let $\mu_\theta^m = Pr(m|\theta)$ be the probability that a sender with news θ sends message m. For most of the models where the type/message space is $\Theta = M = \{0, 1\}$, a more compact way to describe the strategy is to drop the super script and write $\mu_\theta = Pr(m = 1|\theta = 1)$, i.e, unless otherwise noted μ will refer to the probability of sending $m = 1$.

To be more formal, Bayes' rule pins down beliefs for a message m such that $\mu_0^m + \mu_1^m > 0$. If $\mu_0^m + \mu_1^m$ we say message m is *off-path*, and place no constraint on the posterior belief upon observing this message.

B.1 Costly Signaling

Let's start by considering pure strategy equilibria. There are four possibilities here.

First, as discussed in the main text, the good news type can send $m = 1$ and the bad news type $m = 0$. If so the belief upon observing $m = 1$ is $Pr(\theta = 1|m = 1) = 1$ and upon observing $m = 0$ is $Pr(\theta = 1|m = 0) = 0$. Regardless of whether the action is binary or continuous, $a^*(m) = m$. The last thing we need to check is that both types of sender want to send this message given $a^*(m) = m$.

For the good news type, this requires $1 - c_1 \geq 0$ or $c_1 \leq 1$. For the bad news type, we need $0 \geq 1 - c_0$ or $c_0 \geq 1$. Combining, this requires $c_1 \leq 1 \leq c_0$.

Second, there could be separating PBE where both types send the opposite message: $m(\theta) = 1 - \theta$. This quickly falls apart: if so the receiver knows the news is bad when $m = 1$ and hence $a^*(1) = 0$. So the bad news type utility for sending this message is $0 - c_0$, and can always benefit from deviating to $m = 0$ (which gives a minimum payoff of 0).

Third, there could be a pooling equilibrium where both type send $m = 0$. If so, the belief upon observing $m = 0$ is $Pr(\theta = 1|m = 0) = p$. The belief upon observing $m = 1$ is unconstrained since this is off path. If we set this belief to 0 it is immediate that (whether using binary or continuous actions) neither type could deviate to $m = 1$ since it incurs a cost $c_\theta > 0$ and leads to a (weakly) lower action. So there is a always a PBE with this messaging strategy.[31]

Finally, there could be a pooling equilibrium where both types send $m = 1$. If the off-path belief upon observing $m = 0$ is $\hat{p}_0$, then this requires:

$$p - c_\theta \geq \hat{p}$$

The binding constraint is for the low type, or $c_0 \leq p - \hat{p}$.

Mixed Strategies Now consider any equilibrium where both messages are sent with positive probability, in which case we the posterior belief upon observing message m must be:

$$Pr(\theta = 1|m = 1) = \frac{p\mu_1}{p\mu_1 + (1-p)\mu_0} \tag{3}$$

$$Pr(\theta = 1|m = 0) = \frac{p(1-\mu_1)}{p(1-\mu_1) + (1-p)(1-\mu_0)} \tag{4}$$

In any PBE with such a messaging strategy, it must be the case that $a^*(m) = Pr(\theta = 1|m)$. It

[31] As mentioned in the main text, we won't grapple with the question of whether this off-path belief is reasonable, as this kind of analysis can be found elsewhere.

is useful to define:

$$d_1 = Pr(\theta = 1|m = 1) - Pr(\theta = 1|m = 0)$$

For type θ to send $m = 1$ it must be the case that $a^*(1) - c_\theta \geq a^*(0)$, or

$$d_1 \geq c_\theta$$

This inequality is easier to meet for $\theta = 1$ than $\theta = 0$ since $c_1 < c_0$. In words, the type with good news (often called the "strong type") is more apt to send the costly message. Further, if the $\theta = 0$ type picks an interior strategy, he must be indifferent, in which case the $\theta = 1$ type must strictly prefer to send $m = 1$. And if the $\theta = 1$ type plays a mixed strategy, the $\theta = 0$ type must strictly prefer sending $m = 0$.

Combining there are two possible equilibria kinds of equilibria where both messages are sent: the bad news type never sends $m = 1$ and the good news type sometimes (or always) sends $m = 1$, and the bad news type sometimes (but *not* always) sends $m = 1$, and the good news type always sends $m = 1$.

Case 1: $\mu_0 = 0$, $\mu_1 > 0$. Here the bad news type never sends the costly message, and the good news type sometimes does. Plugging these into equations 3-4 gives $Pr(\theta = 1|m = 1) = 1$ and

$$Pr(\theta = 1|m = 0) = \frac{p(1 - \mu_1)}{p(1 - \mu_1) + (1 - p)}$$

If $\mu_1 = 1$, then $Pr(\theta = 1|m = 0) = 0$, and $d_1 = 1$. For this to be a PBE, it must be the case that $c_1 \leq 1$ and $c_0 \geq 1$, precisely the conditions for a fully separating PBE identified above.

If $\mu_1 < 1$, then the good news type must be indifferent between both message, or $d_1 = c_1$. Solving μ_1, this holds when:

$$\mu_1 = \frac{c_1 - (1 - p)}{c_1 p}$$

which is between 0 and 1 when $c_1 \in (1 - p, 1)$. It is possible that this PBE holds but the fully

separating one does not if $1 - p < c_1 < c_0 < 1$.

Case 2: $\mu_0 \in (0, 1), \mu_1 = 1$

This case requires the $\theta = 0$ type to be indifferent between both messages, or $d_1 = c_0$. Solving gives:

$$\mu_0 = \frac{p}{1-p} 1 - c_0 c_0$$

which is between 0 and 1 if $c_0 in(p, 1)$, So, this equilibrium can hold when the fully separating one does not if $c_1 < p < c_0 < 1$.

B.2 Cheap Talk

As with the costly signaling model, let's first consider the case of pure strategy equilibria. If the sender reveals the news honestly ($m(\theta) = \theta$), the receiver effectively learns θ and takes action $a^*(m) = m$. The good news type gets a utility of $1 + b$ for sending $m = 1$ and a utility of of 0 for sending $m = 0$, so will always send $m = 1$. The bad news type gets utility $1 - b$ for sending $m = 0$ and b for sending $m = 1$, so this PBE requires $b \leq 1/2$.

Unlike the costly signaling case, there can also be a PBE where the sender picks the opposite message as her signal. This is because in this PBE, sending the opposite message induces the same action as sending the "correct" message above, so the incentive compatibility constraints are the same. Think of this as the "opposite day" PBE: as long as both actors are aware that the sender says the opposite of the truth, the same amount of information can be conveyed.

There is also always a pooling PBE where both types send either message, call this m^*. A simple way to make this work is to set the off-path belief when observing the other message $m' \neq m^*$ to p, and hence the response to this message is the same. As a result, both types of sender are indifferent between sending m' and m^*. Such a "babbling equilibrium" always exists in cheap talk games.

There can also be mixed strategy PBE but they add little insight.

B.3 Verifiable Information

To reduce cases, again focus on continuous actions.

Recall that in our verifiable information model, the type with news θ chooses from message set $\{\theta, \emptyset\}$. A nice feature of this setup is that when receiving message $m = 1$, it must be the case that $\theta = 1$ because this information set can never be reached if $\theta = 0$. As a result, in any PBE $a^*(1) = 1$, and by a similar analysis $a^*(0) = 0$.

It is possible that the $\theta = 1$ type could still send $m = \emptyset$ if this also induced an action of 1. However, if $a^*(\emptyset) = 1$ then the $\theta = 0$ types have a strict preference to send $m = \emptyset$, and so it can't be the case that $Pr(\theta = 1|\emptyset) = 1$. So, in any PBE, the $\theta = 1$ types always send $m = 1$.

Given this, messages of $m = 0$ or $m = \emptyset$ are either only sent by the $\theta = 0$ types or are off path. If the off-path belief for one of these is greater than zero, then the action taken in response to the other must be zero, and hence the $\theta = 0$ type would deviate. So, any on- or off-path belief upon observing 0 or $\emptyset$ must be $Pr(\theta = 1|m \in \{0, \emptyset\}) = 0$. Given this, the $\theta = 0$ type can send either of these messages or mix between the two, completing the description of the equilibrium.

B.4 Reputation

Consider the model for a general $r \in [0, 1]$.

There are effectively three types here: the good type who knows $\theta = 0$, the good type who knows $\theta = 1$, and the bad type. As in the other models, the probability of being a good type is p, and let the probability that the other state of the world is 1 is q.

As discussed in section B.2, there is always a babbling PBE where both types send the same message (or the same mixed strategy over both messages) and the receiver picks beliefs that leads to low actions when observing any off-path message.

The interesting equilibria to focus on are ones where both messages are sent with positive probability.

First consider a PBE where the good types report honestly. There is no PBE where the bad types always send $m = 0$; if so sending message $m = 1$ would induce actions $a_\omega = 1$ and $a_\theta = 1$, giving the highest possible payoff.

If the bad types always send $m = 1$, then then upon observing $m = 0$ the receiver knows the state is 0 and knows the sender is the good type, giving payoff r.

If sending $m = 1$, the receiver is uncertain about both. The posterior probability that the sender is the good type and the state is 1 become:

$$Pr(\theta = 1|m = 1) = \frac{pq}{pq + 1 - p} \tag{5}$$

$$Pr(\omega = 1|m = 1) = \frac{pq + (1 - p)q}{pq + 1 - p} \tag{6}$$

It is sequentially rational to send $m = 1$ if:

$$r \leq r\frac{pq}{pq + 1 - p} + (1 - r)\frac{q}{pq + 1 - p}$$

rearranging gives this holds if:

$$r \leq \frac{q}{q + 1 - p}$$

That is, if the sender cares relatively little about his reputation for competence (and more about trying to induce a higher action), he will always send $m = 1$ when uninformed. In equilibrium, upon observing $m = 1$ the sender makes a higher policy choice, but also evaluates the sender more poorly.

If r is below this threshold, then in any PBE where the good types are honest the bad types must play a mixed strategy.

If so, the posterior beliefs about competence and the state are:

$$Pr(\theta = 1|m = 1) = \frac{pq}{pq + (1-p)\mu_b}$$
$$Pr(\omega = 1|m = 1) = \frac{pq + (1-p)\mu_b q}{pq + (1-p)\mu_b}$$
$$Pr(\theta = 1|m = 0) = \frac{p(1-q)}{p(1-q) + (1-p)(1-\mu_b)}$$
$$Pr(\omega = 1|m = 0) = \frac{(1-p)(1-q)(1-\mu_b)}{p(1-q) + (1-p)(1-\mu_b)}$$

As $\mu_b \to 0$, the payoff to sending $m = 1$ is always strictly higher than sending $m = 0$. As $\mu_b \to 1$, this approaches the case where the bad type always sends $m = 1$. Further, as, the bad type sends $m = 1$ more often (μ_b increases), this lowers the relative reputational benefit of sending $m = 1$, and also lowers the action taken in response to $m = 1$. So, if $r > \frac{q}{q+1-p}$, then the sender prefers to send $m = 1$ if μ_b is sufficiently low, but prefers to send $m = 0$ if μ_b is sufficiently high. As a result there is a unique μ_b which makes this bad type indifferent.

B.5 Commitment

First, here is a formal claim about commitment not leading to persuasion when the sender has a strictly concave utility in the receiver action:

Theorem 3. *With continuous actions and commitment, if the sender utility is strictly concave in a, then in any PBE the messages are uninformative ($Pr(\theta = 1|m) = p$ for $m \in \{0, 1\}$), and there is no persuasive PBE.*

Proof For any messaging strategy, the sender expected utility is:

$$\begin{aligned} E_m[v(a^*(m)] &= Pr(m=0)v(a^*(0)) + Pr(m=1)v(a^*(1)) \\ &= Pr(m=0)v(Pr(\theta=1|m=0)) + Pr(m=1)v(Pr(\theta=1|m=1)) \\ &\leq v(Pr(\theta=1|m=0)Pr(m=0) + Pr(\theta=1|m=1)Pr(m=1)) = v(p) \end{aligned}$$

where the inequality in the third line follows from Jensen's inequality (treating $Pr(\theta = 1|m)$ as a random variable). If $Pr(\theta = 1|m = 0) \neq Pr(\theta = 1|m = 1)$. This proves that the maximal possible utility is $v(p)$, which is attained for any uninformative message strategy (i.e., if and only if $Pr(\theta = 1|m = 0) = Pr(\theta = 1|m = 1)$). If $Pr(\theta = 1|m = 0) \neq Pr(\theta = 1|m = 1)$ and both messages are sent with positive probability, then the strict concavity implies this inequality is strict. So, any informative strategy can not be a PBE, and hence there is no persuasive PBE. ∎

Complete proof of optimal commitment strategy Suppose the sender always sends $m = 1$ when $\theta = 1$ ($\mu_1 = 1$) and sends $m = 1$ with probability $\mu_0 \in (0, 1)$ when $\theta = 0$ (and hence sends $m = 0$ with probability $1 - \mu_0$. The posterior beliefs upon observing the messages is:

$$\begin{aligned} &Pr(\theta = 1|m = 0) = 0 \\ &Pr(\theta = 1|m = 1) = \frac{Pr(\theta = 1, m = 1)}{Pr(m = 1)} = \frac{p}{p + (1-p)\mu_0} \end{aligned}$$

It is sequentially rational for the receiver to pick $a = 1$ upon observing $m = 1$ if $Pr(\theta = 1|m = 1) \geq 1/2$, or:

$$\frac{p}{p + (1-p)\mu_0} \geq 1/2 \Longrightarrow \mu_0 \leq \frac{p}{1-p}$$

In the range of p we are interested in ($p < 1/2$), $\frac{p}{1-p} \in (0, 1)$. So, there is a unique "maximum" amount of lying where the receiver still follows the signal. The optimal sender strategy is to pick

this maximum amount of lying.

Recall we have restricted the messaging strategy to have no lying when $\theta = 1$ ($\mu_1 = 1$). Can the sender get a higher *ex ante* utility by using a strategy where the sender does not always pick $m = 1$ when $\theta = 1$? Keeping $\mu_0 = p/(1-p)$, the answer is clearly no: if sending $\mu_1 < 1$, the belief upon observing $m = 1$ would no longer be greater than equal to $1/2$, and so the receiver would never pick $a = 1$. In general, if $\mu_1 < 1$, the sender would need to pick a lower μ_0 in order to keep the receiver willing to pick $a^*(1) = 1$.

More abstractly, we can think about the sender problem here as picking a *distribution of posterior beliefs* for the receiver, subject to the constraint that these beliefs are formed by Bayes rule. The goal is to pick a distribution of posterior beliefs that maximizes the probability that this posterior is at least $1/2$. If the belief upon observing both messages is the same, it must be $p < 1/2$, meaning $a = 1$ with probability zero. So, WLOG, let $m = 1$ be the message that induces a higher posterior belief. This is true if and only if $\mu_1 > \mu_0$, which implies $Pr(\theta = 1|m = 0) < 1/2$. Since the receiver only (potentially) takes action $a = 1$ upon observing $m = 1$, we can write the probability of $a = 1$ as a function of μ:

$$Pr(a = 1|\mu) = \begin{cases} 0 & Pr(\theta = 1|m = 1) < 1/2 \\ p + (1-p)\mu_0 & o/w \end{cases}$$

Maximizing this is equivalent to maximizing $p + (1-p)\mu_0$ subject to the constraint that:

$$Pr(\theta = 1|m = 1) = \frac{p\mu_1}{p\mu_1 + (1-p)\mu_0} \geq 1/2$$
$$\mu_0 \leq \mu_1 \frac{p}{1-p}$$

Since we want μ_0 to be as large as possible, the optimal strategy involves setting $\mu_1 = 1$, as assumed above.

Does commitment always lead to more persuasion? The fact that there more persuasion than in the benchmark improves the payoff makes both the sender and receiver better off. That the sender can improve his payoff with commitment power is true by definition, but the fact that this renders the receiver more informed and better off is more surprising. However, does commitment power always lead to more persuasion and a more informed sender?

Here are two examples where it does not.

First, take the costly signaling model from section 4.1 with continuous actions. Recall that the sender is better off in the pooling equilibrium since the average action is always p, but the high type sender has to incur a cost in the separating equilibrium. It is immediate that, with commitment power, the sender would choose to always send $m = 0$. Thus commitment leads to less persuasion and leaves the receiver worse off.

Second, take the verifiable information model from section 4.3, and assume that the sender has strictly concave preferences over the receiver action, and so $pv(1) + (1-p)v(0) < v(p)$. Again, it immediately follows that if the sender could commit to a strategy he would choose to always send $m = \emptyset$, and the receiver would not be persuaded. Intuitively, if the gain from revealing good news is outweighed by the loss to revealing bad information, the sender would like to commit to keeping quiet.

While the observation that commitment can make both actors better off by increasing persuasion, is a valuable insight, it is worth nothing that the opposite can hold as well.