JOURNAL OF THEORETICAL POLITICS

Article

Bayesian explanations for persuasion

Journal of Theoretical Politics
2023, Vol. 35(3) 147–181

DOI: 10.1177/09516298231185060
journals.sagepub.com/home/jtp
Sage

Andrew T Little
Department of Political Science, UC Berkeley, CA, USA

Abstract

The central puzzle of persuasion is why a receiver would listen to a sender who they know is trying to change their beliefs or behavior. This article summarizes 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 *common interest*, (3) the sender messages are *verifiable information*, (4) the sender cares about their *reputation* for competence/honesty, and (5) the sender can *commit* to a messaging strategy (often called 'Bayesian Persuasion'). After reviewing these approaches with common notation, I discuss which provide insight into prominent empirical findings on campaigns, partisan media, and lobbying. While models focusing on commitment have rapidly become prominent (if not dominant) in the recent theoretical literature on persuasion in political science and economics, the insights they provide are primarily technical, and are not particularly well-suited to explaining most of these phenomena.

Keywords

Cheap talk; costly signaling; persuasion; reputation

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.

Corresponding author:
Andrew T Little, Department of Political Science, UC Berkeley, 210 SSB, Berkeley, CA 94720, USA.
Email: andrew.little@berkeley.edu

This article overviews formal approaches to persuasion, with as much common notation as possible. A more specific definition will come in the context of the formalization, but in general I use persuasion to mean any attempt by a sender to change the beliefs or behavior of a receiver to be 'more favorable'. 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.

Given the constraint of rational updating, under some conditions persuasion is impossible.[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.

Of course, persuasion does sometimes occur both in theories and reality. The bulk of the analysis shows how modifying the assumptions in this benchmark makes persuasion possible. While inevitably nonexhaustive, much applied theoretical work uses at least one of five modifications. First, some kinds of messages are costly, and cheaper to send for those with favorable information (*costly signaling*). Second, the sender and receiver can have partially aligned goals (*common interest*). Third, the sender messages can be checked (*verifiable information*). Fourth, the sender may care about perceptions of their competence/honesty (*reputation* concerns). Finally, senders may be able to *commit* to a strategy where they don't lie so much that their messages still affect the receiver behavior (often called 'Bayesian persuasion', though this isn't ideal). 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 overview how they have been applied to three empirical 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.

1. Insights and trends

The first four explanations contain fundamental insights about communication which can be explained without going into any technical detail.

- Costly signaling models tell us that people may engage in seemingly wasteful, inefficient, or harmful behavior if it shows off that they are a 'type' who is willing to do this.
- Common interest models tell us that communication is easier when the sender and receiver have more closely aligned goals, and so doing what the sender wants can be good for the receiver, or at least better than ignoring him.
- Verifiable information models give a simple explanation for communication: when favorable information can't be faked, messages claiming good news should be believed. More subtly, an important insight from these models is that people may reveal mediocre or even somewhat unfavorable information (from

their own perspective) as well if keeping quiet would make things 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, these incentives may also cause senders to cater to the prior belief of their audience or make stronger claims than are warranted by their information.

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

The key insights from models with commitment are harder to boil down. But this has not hindered their popularity. In fact, 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 (see Kamenica, 2019: for a recent review).

Figure 1 provides 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 to 2021. ‘Cheap talk’ can loosely describe any approach where messages are not costly, though it is often associated with the second explanation (common interests).

The first two increase steadily—perhaps 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 indicates that results for costly signaling and cheap talk include many empirical papers while Bayesian persuasion returns

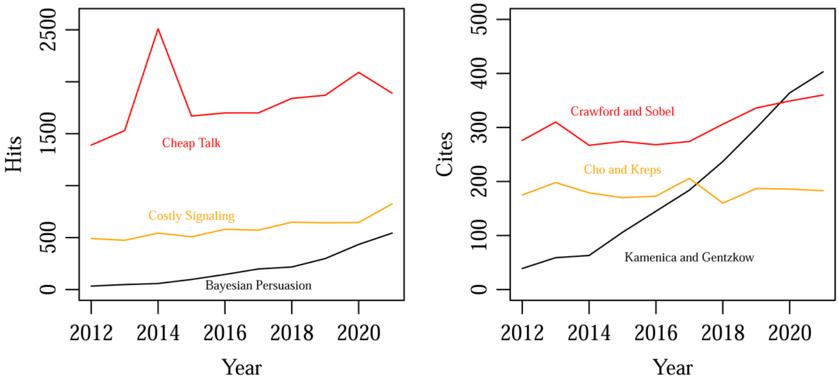

Figure 1. Google scholar search hits for kinds of model (left panel) and citations to influential papers (right panel).

almost entirely theoretical papers. So this figure likely understates the rise of theoretical papers relying on commitment.

The right panel shows the number of citations to three influential theoretical papers: Crawford and Sobel (1982), which is an early example of cheap talk with common interests; Cho and Kreps (1987), which studies belief refinements in costly signaling models;[2] and Kamenica and Gentzkow (2011), which popularized the commitment approach. Over the past decade, Kamenica and Gentzkow (2011) has overtaken these two influential papers in citations per year.[3]

Are models where the sender can commit to a strategy rapidly rising because this approach is uniquely good at explaining particular kinds of political persuasion?

The goal here is not to argue that the answer to this question is always 'no', let alone to argue that persuasion via commitment is never appropriate. Rather I aim to situate this kind of model in a wider framework including other approaches.[4] As always, different assumptions are appropriate for different problems, and this paper aims to give an overview of what kinds of assumptions are useful for different research questions about persuasion.[5]

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 many insights about communication with just two possible things the sender can know and two possible things he can say.

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.) and $\theta = 0$ 'bad news'. It is common to refer to θ as the 'type' of the sender; here there is a good news type and a bad news type. Let $p = Pr(\theta = 1)$ be the prior belief that the information is good news. The receiver does not observe θ, though she might learn something about it from the sender message.

The messages available are also $M = \{0, 1\}$.[6] 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 lies and sends $m = 1$ when in fact $\theta = 0$—that is, 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 receiver action set as well: $a \in \{0, 1\}$. In general, interpret $a = 1$ as doing what the sender wants; for example, voting for the sender's preferred candidate, 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, making a choice closer to the sender's ideal point, or just 'supporting' the sender more generally. Equivalently, there may be a lot of actions that the sender wants the receiver to take, and we can interpret a as the proportion of favorable actions taken. A final possible interpretation is that the action is binary, but the receiver may be more or less likely to take it for random and exogenous reasons. We can interpret a in the continuous choice model as the probability of taking the favorable action.

Utilities

Different models will make different assumptions about the sender utility. The key common thread will be that the sender wants the receiver to take a relatively 'high' action. In some models this will be captured by assuming the sender utility is always increasing in a. Other times the sender utility will not be monotone but they at least prefer a higher action than what the receiver would do otherwise. The sender utility will also sometimes directly depend on the message sent (i.e., some messages are costlier than others) or the state θ (i.e., the sender also wants the receiver to take different actions with different news).

For the receiver, we will always use the following utility:

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

which captures the idea that the receiver wants her action to be close to θ with increasing marginal costs.[7] 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 specific functional form 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 ($\theta = 1$).[8]

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 (hereafter, *equilibrium*). 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 the action 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, both choices give 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 cleanly 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 to the probability she assigns to the sender having good news.

3. When Bayesians can't be persuaded

What does it mean for 'communication' or 'persuasion' to happen? One general definition is that a sender persuades a receiver if his message leads to receiver to changes his beliefs in a way that the sender likes. Of course, senders typically don't have direct preferences over what the receiver believes, if such a preference over another person's internal state is even possible. Rather senders typically care about how these beliefs are mapped into an action which the sender cares about.

Still, much empirical work takes the beliefs (or attitudes) of receivers as the main outcome of interest (e.g., Druckman, 2021; Coppock, 2022). Such beliefs and attitudes may be intrinsically important. Further, senders may aim to manipulate beliefs without a specific target action in mind. For example, an interest group may want to move public opinion to be more favorable about policies they support even if there is no particularly relevant election or referendum on the horizon.

In the context of the class of models here, which contain a concrete action taken by the receiver, we will define persuasion with respect to this action. Before getting to that, it will help to ask whether the sender can systematically change the receiver beliefs. In one sense he certainly cannot. 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 In any equilibrium,

(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 ($Pr(\theta = 1|m = j) < p$ 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 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.[9] The binary restrictions are not needed for this result; it holds for any type space and message set (and any utility function).[10] One way to think of this fact is from the perspective of the receiver before learning the message: if she expects her belief that the sender has good news after hearing his message 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 the sender has good news, 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 a rational receiver's belief that he has good news.

Even so if we define persuasion as changing *actions*, it may still be possible.[11] A natural way to define whether persuasion occurs is to compare what the receiver does relative to a benchmark action with 'no communication'. In words, can the sender use the opportunity to speak to the receiver 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 $p = 1/2$:

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

In general, we should define a persuasive message as one that leads to a better outcome for the sender. As we will generally focus on the case where the sender wants a high action, this is, straightforward:

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

To map this to a common empirical setting, we are making a comparison between an outcome (the action) under one 'treatment' (hearing message m) versus a 'control' condition of receiving no new information. A message is persuasive if and only if there is a strictly positive treatment effect.

Whether there is a persuasive equilibrium by this definition depends on the sender utility. Start with the simple assumption that the sender only cares about inducing higher actions. Formally, say the sender has a *univariate monotone* utility if it can be written

$$u_S = v(a) \tag{2}$$

where v is a strictly increasing function.

This definition combined with Theorem 1 leads to a stark result, which shows up in some form in many related papers (see e.g., Theorem 2 in Lipnowski and Ravid, 2020: for a more general statement):[12]

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

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 same message, the belief upon observing it must be the same as prior, meaning it can't be persuasive.

Taking stock

We started by observing that 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 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 this question.

First, we could change something about the utility functions or information structure. As we will see in Sections 4.1 to 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 Sections 4.5. These are often called models of 'Bayesian persuasion', which is unfortunate as the models discussed in Sections 4.1 to 4.4 also study persuasion among actors who are Bayesian. Following Gehlbach (2021: Ch. 8) and to contrast the models in Sections 4.1 to 4.4, call these models of *Bayesian persuasion with commitment.* In Section 4.6, I discuss some models which blend features of different approaches and explore the relationship between them.

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, I will not formalize this class of explanations but discuss when we may (or may not) want to use this approach in Section 5.

Scope

As a last bit of throat-clearing, there are some variants of these approaches and related literatures which I will not cover beyond superficial references.

- The emphasis will remain on situations where an informed sender says something to a less-informed decision-maker. Related work considers when the decision-maker may delegate authority to the informed party (see Gailmard and Patty, 2012; Bendor and Meirowitz, 2004; Dessein, 2002), or how decision-makers with private information may distort their actions (or 'pander') because of the inferences that receivers make (e.g., Canes-Wrone et al., 2001). This also rules out 'career concerns' models where an uninformed sender takes an action which distorts a signal observed by a receiver (e.g., Holmström, 1999).
- The information held by the sender is taken as a given. That is, the models will not include information gathering (e.g., Austen-Smith and Wright, 1992; Patty, 2009).
- We also will not bring other actors into the mix. For example, we will not cover the role of mediators; see Kydd (2003) for an application to conflict and Salamanca (2021) for a recent theoretical discussion. Passing reference will be made to models with multiple senders (e.g., Minozzi, 2011), multiple receivers (Farrell and Gibbons, 1989; Levy and Razin, 2004) or both (Battaglini, 2002), which can modify the conclusions of the different approaches in different ways.

4. When Bayesians can be persuaded, without and with commitment

We can think of the different kinds of communication/persuasion models commonly used in political science as different ways of getting around the 'impossibility' results of Section 3.

As above, the description of these models aims to include just enough formalization to convey the main ideas. Appendix B. contains more complete analyses, which are standard.

4.1. Costly signaling

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

$$u_S = a - mc_\theta$$

where $c_1 < c_0$. 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 usually 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 costly signaling and later models, we will primarily focus on when there is *fully informative* equilibrium where the receiver can perfectly infer the sender type. The natural kind of fully informative equilibrium to study is one where the message equals the state:

Definition: An equilibrium is *honest* if $m^*(\theta) = \theta$.

In such an equilibrium $m = 1$ is persuasive, and hence honest equilibria are always persuasive equilibria.

For either binary or continuous receiver actions, there is an honest (or 'seperating') equilibrium if $c_1 \leq 1 \leq c_0$.[13] In this equilibrium, the strong type gets $1 - c_1 > 0$ for sending the costly message (m=1) 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 weak type to send the costly message 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) in the main text, an important observation about this basic costly signaling model is that there is also always a 'pooling' equilibrium where the both sender types say $m = 0$ and the receiver picks the same action regardless of the message.[14] Even if a message could in principle serve as a costly signal of a good news/high type, nothing forces the receiver to interpret it in this fashion.

Even in the separating/honest equilibrium, Theorem 1 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 equilibrium than the 'pooling on $m = 0$' equilibrium.[15] Put another way, the strong type can benefit from communication relative to the weak type, but the ability to send costly signals cannot make the receiver systematically think that the sender is strong.

The costs and benefits of sending messages could depend on the sender information in different ways. If the sender's *benefit* from the receiver taking the high action depends on his private information, this induces some common interest, and so I discuss this more in Section 4.6. Another possibility is that there is a cost to lying, that is, sending a message not equal to the state (e.g., Kartik, 2009).

Combining, costly signaling models are an appropriate way to model 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, 1986a), as are donations (Gordon and Hafer, 2005; Schnakenberg and Turner, 2021), and 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 breaking a treaty (Hollyer and Rosendorff, 2011).

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

4.2. Common interests

The remaining four explanations all rely on ‘cheap talk’ in the sense that messages do not directly enter into the sender utility function. In the political science literature, this phrase usually evokes a model where the sender and receiver have some interest in common, in the style of Crawford and Sobel (1982). As there is some ambiguity here, I won’t take a strong stance on what exactly should count as cheap talk, but will flag that common interest is only one explanation for persuasion where messages aren’t directly costly. Still, it is an important one in many social settings.

In the extreme, if the sender utility is the same as the receiver utility, there can be an honest (and persuasive) 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 idea but retain the premise that the sender also tends to want the receiver to take high actions, let the sender utility be:

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

The $b \in [0, 1]$ term or “bias” captures the relative importance of the sender taking a high action, with $1 - b$ weighting the u_R term which captures the common interest between the sender and the receiver. Regardless of whether actions are continuous or binary, there can be an honest equilibrium if the bias term is relatively small. Formally, suppose the sender reports the news honestly ($m^*(\theta) = \theta$) and so the receiver takes an action equal to his message ($a^*(m) = m$). Given the receiver does what he says, a sender with good news clearly wants to tell the truth ($m = 1$). The sender with bad news faces a tradeoff where telling the truth leads to the correct action for the receiver (good for common interest, bad for inducing high actions) and lying would trick the sender into taking action $a = 1$ (a favorable action, but bad for common interest). Telling the truth gives the sender utility $1 - b$ and lying gives b, and so there is an honest equilibrium if and only if $b \leq 1/2$.

Common interest is a natural assumption in many settings 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 related 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 to do, X is more important to me than Y’.[17] 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).[18]

4.3. Verifiable information

While costly signaling and cheap talk models allow for persuasion by changing assumptions about preferences, another possibility is to change the information structure. One approach in this vein is to assume that messages are *verifiable* or *hard information*.[19]

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'. That is, the sender can either reveal the truth or keep quiet.[20]

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 an example with partial lie detection).

With any monotone sender utility and either continuous or binary actions, there is an honest and persuasive equilibrium where the type with good news reveals this ($m^*(1) = 1$) and the type with bad news admits this ($m^*(0) = 0$) because their only other option is to say nothing which the receiver also interprets as having bad news.[21] 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.

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—in particular, between 1/4 and 1/2—would rather reveal it than keep quiet. In any potential equilibrium where those with news worse than a threshold keep quiet, the type just below the threshold has an incentive to reveal what he knows, effectively saying 'the news may be bad but it's not *that* bad'. This unraveling argument leads to the conclusion that there must be full revelation of information.

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. While the assumption of perfect and free verification may be extreme, the insights of these models extend to cases where verification is imperfect or costly (e.g., Austen-Smith and Wright, 1992).

4.4. Reputation

Often times senders don't care as much about persuading 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') or honesty ('I am the type who will tell you the truth no matter what'). Such reputation concerns can increase or decrease the prospect of persuasive communication.

To see how, 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 successful. The sender knows his competence and gets a signal which might be informative about the state. In particular, incompetent senders ($\theta = 0$) get an uninformative message $s = \emptyset$, and competent senders ($\theta = 1$) observe $s = \omega$.

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 receiver best responses are $a_\theta^*(m) = Pr(\theta = 1|m)$ and $a_\omega^*(m) = Pr(\omega = 1|m)$.

Reputation for good. First consider a case with sender utility:

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

where r scales the reputation concerns relative to the policy concerns.

If $r = 0$ (only policy concerns), there is no persuasive equilibrium by a similar logic to the one in Section 3; if either message led to a better policy everyone would send it, rendering the message uninformative.

If $r = 1$ (only reputation concerns), there is no fully honest equilibrium where the uninformed type sends a distinctive message $m = \emptyset$ because this would lead to a competence assessment of 0 while either $m \in \{0, 1\}$ would lead to a competence assessment of 1. However, there is a 'partially honest' equilibrium where the competent type reveals his signal ($m = s$) and the incompetent type guesses, sending $m = 1$ with probability q and $m = 0$ with probability $1 - q$. In this equilibrium, the receiver learns nothing about the sender competence; if not, the incompetent type would send whatever message makes him appear more competent. However, since the sender is more likely to say the policy is good when this is true, the receiver does get some information 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 an equilibrium with some learning/persuasion about both the expert competence and the ideal policy,[22] but a fair amount of lying.

Reputation for bad. Now suppose the sender's 'non-reputation' motive is aligned with the receiver utility, and so:

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

If $r = 0$, there is an honest and persuasive equilibrium by the logic of the common interest model. The only additional thing to check is that is incentive compatible to report getting an uninformative message ($m = \emptyset$ when $s = \emptyset$) because being honest about not knowing the state renders an intermediate action of $a = q$ optimal and the sender does not care that this gives him away as incompetent since there are no reputation concerns.

However, if r is sufficiently large, this honest reporting is no longer possible since the incompetent sender would rather deviate to pretending to have an informative message (sending $m = 0$ or $m = 1$) if this makes him look competent, even if it leads to a worse policy choice. When $r = 1$ this utility function is the same as the previous one, and hence the equilibrium again involves the competent types revealing the truth and the incompetent types guessing with a mixed strategy that renders the message partially informative about the state but not informative about competence.

Summary. Reputation models are natural when studying political actors whose care about views of their ability (Backus and Little, 2020).[23] As modeled here, reputation concerns can lead to more or less information transmission depending on whether the sender would be able to communicate well without this incentive. More generally, reputation concerns can 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, let's consider 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, for example, Bénabou and Tirole (2002) can be interpreted as studying a 'rational' self committing to a messaging strategy to a 'deciding self' with present bias.[24] 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). Kamenica (2019) and Bergemann and Morris (2019) provide recent reviews of this literature and the broader study of 'information design', respectively.

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).[25]

An equilibrium 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.

Continuous action, weakly concave utility. Before getting to the 'standard' case of Bayesian persuasion with commitment, which focuses on binary receiver actions, 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, that is, 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 $\mathbb{E}_m[Pr(\theta = 1)] = p$. The expected utility for any messaging strategy is $\alpha + \beta p$.

So, with linear utility, the sender is indifferent between any choice of μ, and hence there is an equilibrium with every possible messaging strategy. The benefit to higher actions when the news is good gets perfectly offset by the lower actions when the news is bad.

Technically speaking, there can be a persuasive equilibrium in this setting: for example, 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 persuade. Further, since any messaging strategy is possible in equilibrium, this will not be a useful benchmark to bring to applied models.

There are many other kinds of sender utility functions to consider beyond the linear case, but one commonly used family is the 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 equilibrium. This follows from the fact that any informative messaging strategy makes the sender action more volatile, and the concave utility effectively makes the sender risk averse. See Appendix B.5. for a formal statement or Remark 1 by 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 $Pr(\theta = 1|m) \geq 1/2$. 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?[26]

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, an honest equilibrium is better than an uninformative one for the sender (and the receiver).

However, the sender can do even better. To see why, suppose the sender also occasionally lies and says the news is good when it is bad (sends $m = 1$ when $\theta = 0$, or $\mu_0 > 0$). If such lies are sufficiently rare, then the receiver will still be confident that the news is likely good when the sender claims it is, 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 $m = 1$.

Commitment and welfare. Another useful way to think about when commitment leads to persuasion is that it will do so when informative equilibria are better for the sender. With continuous actions and weakly concave preferences, the sender is not *ex ante* better off in a persuasive equilibrium. The possibility of persuasion 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 learning that $\theta = 0$.

However, with binary preferences and $p < 1/2$, the benchmark action is the worst possible for the sender, and hence he tends to prefer persuasive messaging strategies.

This perspective also helps us quickly see when commitment leads to more communication in some of our other models. In a costly signaling model where $m = 0$ is free and $m = 1$ is costly, the sender would prefer to commit to a model with no communication (pooling on $m = 0$) since this reduces the amount of inefficient messaging (see Little, 2017a: for a discussion of this point in the context of models that treat authoritarian elections as costly signaling).

On the other hand, take the common interest 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.

In the verifiable information and reputation models, whether the sender prefers persuasion depends on whether his preference over the receiver action(s) is concave or convex.

In sum, the idea that commitment leads to more communication and persuasion is sensitive to other assumptions made about the environment.

Is the commitment assumption too strong?. 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 of when it does and does not apply. There are certainly situations where a full commitment assumption is a stretch, such as most examples of interpersonal communication. Still, the are a few reasons this line of attack doesn't always hit the mark.

First, the conclusions of models with full commitment are partially if not completely robust to allowing for some chance of deviating from the committed strategy. Intuitively, if there is a small chance that the sender can deviate from the probability of falsifying good news that they commit to, they can simply 'commit' to lying a bit less to offset this and their behavior is effectively the same from the receiver perspective.[27] Even if pure commitment is rare, like any assumption it is worth exploring how things play out in the special case where this is true.[28] See Lipnowski et al. (2019) for a general analysis of partial commitment, and Luo and Rozenas (2018) and Prato and Turner (2022) for applications to electoral fraud and legislative oversight.

Second, commitment can often be partially microfounded with other mechanisms; see Section 4.6 for examples. This is convenient from a theoretical perspective, as we can take a model with commitment as a reduced form way to capture any force which makes people sometimes willing to tell the truth even when lying helps them in the short term. However, if our aim is to explain why persuasion happens in real situations, we should keep in mind that models relying on commitment for convenience are not saying that this is the real mechanism.

Finally, a common (and often appropriate) justification for commitment is that the choice being studied is not an individual act of persuasion, but setting up a more durable institution (e.g., choosing general guidance to media outlets about what degree of government criticism will be allowed) which will produce a bias in the information produced.[29] It can also be appropriate in individual settings where the sender is choosing to 'commission a study' of some form 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.

Even when we accept the commitment assumption, equally important for this approach 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 the news is good, then being able to commit to a messaging strategy is not very helpful, if at all.

The key point here is not that convexity is never a reasonable assumption, just that it should be justified. There are certainly many natural political situations where the sender does indeed have convex preferences over the receiver's beliefs/action. For example, many models include the common situation where the sender has a more specific goal of getting the receiver to take 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 'does our adversary value this territory enough to fight back if we invade?') In this case, it is also important that the sender has a pretty good sense of what the receiver prior belief is and the threshold for action, otherwise the situation is better approximated by the continuous case. Still several papers do study persuasion of an audience with a heterogenous prior or preferences (Gehlbach and Sonin, 2014; Alonso and Câmara, 2016), or a single listener with private information (Kolotilin et al., 2017) or even an unknown prior (Kosterina, 2018).

4.6. Hybrids

The goal of the preceding subsections was to highlight 'pure' cases of each explanation for persuasion. Many other papers explore models which combine features of the different approaches, or otherwise study the relationship between them. There are many potential combinations so the aim here is to give a few examples rather than explore all possibilities.

One natural combination of costly signaling and common interest arises if the sender private information is about his own benefit for the receiver taking the favorable action. For example, a lobbyist may try to persuade a legislator that a proposed change to a regulation would be very beneficial for their industry. This is related to common interest because when the news is good ('the new policy would be very helpful'), both benefit more from the receiver taking higher actions.[30] In this setting, even '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 (Austen-Smith and Banks, 2000).

Several settings naturally combine components of costly signaling and verifiable information.[31] 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, 2017a).

Models focused on reputation typically do not lead to full communication without adding some other features 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 induce a sender to listen to future messages (Sobel, 1985; Kuvalekar et al., 2022; Best and Quigley, 2020).

Following Kamenica and Gentzkow (2011), several theoretical papers ask whether the sender can end up approximating the best outcome they could achieve with commitment via some mechanism other than commitment. One commonly studied mechanism is repetition. If a sender plays a repeated version of the game studied here, combined with some feedback about past play (Best and Quigley, 2020), private lying costs (Pei, 2020), or some types keeping promises (Fudenberg et al., 2020), persuasion is possible, and in special cases the sender behavior approaches that in the one-shot game with commitment. 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. It bears repeating that this line of work, while potentially justifying a commitment assumption in other models, does not mean that real-world persuasion is actually driven by the ability for senders to commit to a strategy.

In sum, depending on the context it is often natural to combine different mechanisms for persuasion. Often (if not usually) it does not make sense to think of them as competing explanations, but as highlighting different features which may or may not be present in different real-world settings.

5. Non-Bayesian persuasion

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

1. All models must make some assumptions about beliefs. We can always pick beliefs that make persuasion occur practically by assumption, such as when the receiver takes any message at face value. At the other extreme, 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, and so this is a natural place to start.
2. Loosening the correct belief assumption for even one actor can create tricky practical issues (what do others know about these incorrect beliefs?) and philosophical issues (do other aspects of solution concepts often justified by common knowledge of rationality still make sense?). See Fudenberg (2006: section 3.6) for further discussion of this point.

3. Small deviations from the rational benchmark may not dramatically change the conclusions that would be reached with a more standard model

These points are collectively important, but do 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 can 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).

As some examples, receivers may be partially ‘credulous’ (Kartik et al., 2007; Little, 2017b; Horz, 2021), which can have non-obvious implications beyond making persuasion easier. A more subtle bias is that individuals or struggle to make inferences when information which is hidden or what they observe is nonrepresentative (Enke, 2020; Eyster and Rabin, 2005; Jin et al., 2021). Other recent papers explore how non-Bayesian updating can expand the amount of persuasion possible with commitment, potentially leading to the receiver always doing what the sender wants (e.g., Levy et al., 2022).

In sum, while Bayesian explanations for persuasion have generated substantial insights and should arguably be our default approach, in some settings allowing for deviations from this benchmark may lead to simpler explanations with more explanatory power.

6. Empirical applications

In this section, I discuss some prominent classes of empirical findings in light of the analysis above.[32] These are all enormous literatures and the aim here is not to be anywhere near comprehensive, nor claim that any class of results must be explained by a particular kind of model. Rather the idea is to give a general sense of some key findings and how the different theoretical approaches shed light on when we should and should 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 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 diverse contexts, such as television advertising in the United States (e.g., Huber and Arceneaux, 2007; Spenkuch and Toniatti, 2018), mail and phone contact in Italy (Kendall et al., 2015), banners on streets in Spain (Esteban-Casanelles, 2020), and clientalist appeals in Benin (Wantchekon, 2003).[33]

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—using a blend of factual and emotional appeal—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 is less obvious, as it isn’t clear that candidates with more favorable information (they are aligned with the voters, the other candidate is corrupt, etc.) face lower costs. Still, this could be plausible if it is easier to get volunteers or raise money for better candidates. 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.

While exaggeration and stretching of the truth are common, much campaign information is also at least partially verifiable. Kendall et al. (2015) are explicit about this, emphasizing that their intervention gives voters ‘hard and verifiable information’ about the positions and valence of candidates.

What about non-Bayesian explanations? All the previous mechanisms can be magnified if some voters are credulous and take messages at face value. However, 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. One plausible non-Bayesian mechanism is that campaigning may simply raise certain ‘considerations’ to the top of voters minds (Zaller et al., 1992; Aragonès et al., 2015; Dragu and Fan, 2016). 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).

Finally, reputation and commitment seem less suited to this context. For 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 are that outlets themselves typically care not just about winning a current election, but about (1) their reputation in the longer term, and (2) gaining viewership. While this difference will affect interpretation, the big picture painted by empirical results from this literature is similar, with well-identified studies typically finding modest but often politically consequential 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 viewers more conservative in their attitudes and voting. Partisan news may exert a particularly strong influence if the slant of the provider is less known, for example, 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 et al., 2011).

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 is plausible 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 news 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 (Petrova, 2011; Gehlbach and Sonin, 2014). 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. As discussed in Section 4.6, recent theoretical work illustrates that reputation concerns in a dynamic model can lead to behavior approximating the optimal static strategy with commitment.

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. One simple explanation for this is that viewers 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 et al. (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, there is some evidence consistent with 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, 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.

7. Theoretical takeaways

As discussed in Section 1, models based on costly signaling, common interest, verifiable information, and reputation all contain core insights that have been useful both for formal theorists and more widely. If I may issue a challenge to proponents of using models of Bayesian persuasion with commitment, it isn't that they need to do more to justify any particular 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)? True, but the idea that actors who could commit to behavior could make themselves individually or collectively better off is one of the most widely-known insights from game theory (the prisoners' dilemma, trust games, bargaining models, etc.). Further, as discussed above, this result is not true in all settings; often the ability to commit to a strategy leads to less communication and persuasion.
2. That commitment is more valuable when senders have convex preferences over receiver beliefs? This is an interesting technical observation, but I usually 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.[34] This could be driven by the fact 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. This class of models is often nice to work with technically, and has generated many elegant results. These are valuable features. But an approach drawing so much time and attention from social scientists should aspire to more. I am skeptical that models relying on commitment will have a much wider impact in the future, but am open to being persuaded.

Acknowledgements

Many thanks to Nageeb Ali, Elliot Lipnowski, John Patty, Harry Pei, Doron Ravid, Konstantin Sonin, and two anonymous referees for helpful comments and discussion.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Andrew T Little https://orcid.org/0000-0001-9911-4291

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

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

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 $\mathbb{E}_m[Pr(\theta = 1|m)] = p$ holds regardless of the off-path belief.

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 equilibrium, 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 equilibrium 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 superscript and write $\mu_\theta = Pr(m = 1|\theta = 1)$, that is, unless otherwise noted μ will refer to the probability of sending $m = 1$.

Bayes' rule pins down beliefs for a message m such that $\mu_0^m + \mu_1^m > 0$. If $\mu_0^m + \mu_1^m = 0$, 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 equilibrium 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 types 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 an equilibrium with this messaging strategy.[35]

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 equilibrium with such a messaging strategy, it must be the case that $a^*(m) = Pr(\theta = 1|m)$. It 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) and (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 an equilibrium, it must be the case that $c_1 \leq 1$ and $c_0 \geq 1$, precisely the conditions for a fully separating equilibrium identified above.

If $\mu_1 < 1$, then the good news type must be indifferent between both messages, 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 equilibrium 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:

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

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. Common interests

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 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 equilibrium requires $b \leq 1/2$.

Unlike the costly signaling case, there can also be an equilibrium where the sender picks the opposite message as her signal. This is because in this equilibrium, 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' equilibrium: 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 equilibrium 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 equilibria 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 $\{\theta, \varnothing\}$. 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 equilibrium $a^*(1) = 1$, and by a similar argument $a^*(0) = 0$.

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

Given this, messages of $m = 0$ or $m = \varnothing$ 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 $\varnothing$ must be $Pr(\theta = 1|m \in \{0, \varnothing\}) = 0$. 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 where $u_S = ra_\theta + (1 - r)a_\omega$ for a general $r \in [0, 1]$. The version with $u_S = ra_\theta + (1 - r)u_R$ is left as an exercise for the reader.

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 equilibrium where all 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 an equilibrium where the good types report honestly. There is no equilibrium 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 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:

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

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

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 equilibrium 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 $\mu_b \rightarrow 0$, the payoff to sending $m = 1$ is always strictly higher than sending $m = 0$. As $\mu_b \rightarrow 1$, 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, and hence a unique equilibrium where the good types are honest.

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 equilibrium the messages are uninformative ($Pr(\theta = 1|m) = p$ for $m \in \{0, 1\}$), and there is no persuasive equilibrium.

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 cannot be an equilibrium, and hence there is no persuasive equilibrium.

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:

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

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

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.

Notes

1. The title is a nod to Fearon (1995), which gives a typology for causes of war structured in a similar way.
2. 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.
3. 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 this style of cheap talk model.
4. 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.
5. 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 (1986b) propose reserving 'persuasion' for models with verifiable information. This is unfortunate. Persuasion is a huge part of human interaction, so it is perfectly appropriate that there are several prominent ways to theorize about it.
6. For most of the models there is little if any loss of generality in making the message space the same size as the type space.
7. Another commonly used utility function which leads to the same optimal action for any belief about θ and hence the same equilibria in any version of the model is $u_R = -(\theta - a)^2$.
8. Another natural utility function would use a linear loss: $u_R = -\theta(1 - a) - (1 - \theta)a$. With this utility function the best response is to pick $a = 1$ if $Pr(\theta = 1) \geq 1/2$ and $a = 0$ if $Pr(\theta = 0) \leq 1/2$. That is, this makes the receiver behave as she would if the action space was restricted to $\{0, 1\}$.
9. The proof requires just a hair of additional work to deal with the possibility of off-path beliefs, which need not be formed by Bayes' rule, but don't affect the average belief precisely because they are off-path.
10. When considering more general information structures, Kamenica and Gentzkow (2011) use this fact as the key constraint on what kinds of posterior belief distributions are possible ('Bayes plausibility').
11. Another way to put this—coined by Scott Ashworth in a now-deleted 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).
12. 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 $p \neq 1/2$.
13. This condition may seem restrictive, but if we allow for the message to be a continuous choice $m(\theta) \geq 0$, there is always a continuum of equilibria of this form where $m(0) = 0$ and $c_1 \leq m(1) \leq c_0$.
14. This can always be true if the 'off-path' belief upon observing $m = 1$ is that $Pr(\theta = 1|m = 1) = p$. Much theoretical work on signaling attempts to identify when such beliefs are reasonable (e.g., Cho and Kreps, 1987), a topic beyond the scope here.
15. 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 equilibrium than the pooling equilibrium since the $\theta = 0$ type gets utility 0 in either case and the $\theta = 1$ type gets $1 - c_1 > 0$ in the persuasive equilibrium.
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).

17. 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 but replacing a with $a_1 + a_2$. Let the sender utility be $a_1 + ra_2$, where $r > 0$. That is, 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 high of an action you choose, please do it on dimension d', where $d \in \{1, 2\}$. That is, he can't persuade her to take a higher sum action, but can persuade her to do the kind of action he prefers.
18. Lipnowski and Ravid (2020) also provide a simple example where the state is binary but still captures the key idea here well. There are three possible receiver actions: a bad status quo choice, and two possible reforms which the receiver will only enact if confident that the chosen one is best. A nice way to think about why persuasion works here is that the receiver utility is non-monotone in the belief that one policy is best, as intermediate beliefs lead to the bad status quo. So the sender can benefit from informing the receiver, though if he prefers one reform over the other this reduces the amount of possible persuasion.
19. In fact, this is precisely the feature that Milgrom and Roberts (1986b) use to distinguish games of persuasion from cheap talk models, though I prefer using persuasion to refer to a wider class of approaches.
20. The choice set could also be written $M = \{\theta, \{0, 1\}\}$, that is, the sender can't lie in the sense that if the state is 0 they can either say 'the state is 0' or 'the state is 0 or 1'.
21. This does rely on an off-path belief that $Pr(\theta = 1|m = \varnothing) = 0$. There is another fully informative equilibrium, where $m^*(0) = \varnothing$, where the only off-path message ($m = 0$) can only come from the type with bad news.
22. Our definition of a persuasive message equilibrium relied on there being only one action that the sender cares about. However we can naturally extend the definition of being persuasive about ω or persuasive about θ as increasing the action on the respective dimension.
23. 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 et al., 2001).
24. While not published until after Kamenica and Gentzkow (2011), an early political economy application, Gehlbach and Sonin (2014), was circulated in 2008 if not earlier.
25. 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. That is, the game form is not changed, but the solution concept is.
26. 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.
27. See Ederer and Min (2022) for an example of similar offsetting when there is a chance that lies are detected.
28. See also Lin and Liu (2022), who discuss when allowing the receiver to observe the distribution of sender messages does and does not allow the sender to play the same strategy they would without commitment.
29. A drawback of this interpretation is that if a choice like media bias is 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 is possible for uncertainty about the threshold to undermine the possibility of successful persuasion with commitment; for example, 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.

30. Formally, 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 equilibrium with $m^*(\theta) = \theta$ by an identical logic.
31. For a combination of commitment and verification, see Glazer and Rubinstein (2004), who study a model where the receiver has commitment power, and chooses what 'aspect' of a sender message to verify.
32. One large strand of work I do not include is lab and survey experiments on persuasion (see Druckman, 2021: for a recent overview with more of an experimental focus).
33. These studies focus on persuading voters to select certain candidates for parties. Persuasive effects are often stronger when trying to get voters to turn out, particularly with canvassing (Gerber and Green, 2000) or social pressure (Gerber et al., 2008).
34. 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.
35. 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.

References

Acharya A, Grillo E, Sugaya T, et al. (2019) Dynamic Campaign Spending. http://stanford.edu/avidit/campaigns.pdf.

Adena M, Enikolopov R, Petrova M, et al. (2015) Radio and the Rise of the Nazis in Prewar Germany. *The Quarterly Journal of Economics* 130(4): 1885–1939.

Alonso R and Câmara O (2016) Persuading voters. *American Economic Review* 106(11): 3590–3605.

Aragonès E, Castanheira M and Giani M (2015) Electoral competition through issue selection. *American journal of political science* 59(1): 71–90.

Austen-Smith D and Banks JS (2000) Cheap talk and burned money. *Journal of Economic Theory* 91(1): 1–16.

Austen-Smith D and Wright JR (1992) Competitive lobbying for a legislator's vote. *Social choice and Welfare* 9(3): 229–257.

Backus M and Little AT (2020) I don't know. *American Political Science Review* 114(3): 724–743.

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

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

Bendor J and Meirowitz A (2004) Spatial models of delegation. *American Political Science Review* 98(2): 293–310.

Benjamin DJ (2019) Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1* 2: 69–186.

Bergemann D and Morris S (2019) Information design: A unified perspective. *Journal of Economic Literature* 57(1): 44–95.

Best J and Quigley D (2020) Persuasion for the long run. *Available at SSRN 2908115.*

Bombardini M and Trebbi F (2020) Empirical models of lobbying. *Annual Review of Economics* 12: 391–413.

Broockman D and Kalla J (2016) Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science (New York, N.Y.)* 352(6282): 220–224.

Broockman D and Kalla J (2022) The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with Fox News viewers.

Brundage M, Little A and You S (2022) Selection Neglect and Political Beliefs. manuscript.

Canes-Wrone B, Herron MC and Shotts KW (2001) Leadership and pandering: A theory of executive policymaking. *American Journal of Political Science* 45(3): 532–550.

Chakraborty A and Harbaugh R (2010) Persuasion by cheap talk. *American Economic Review* 100(5): 2361–82.

Cho I-K and Kreps DM (1987) Signaling games and stable equilibria. *The Quarterly Journal of Economics* 102(2): 179–221.

Coppock A (2022) Persuasion in parallel. In *Persuasion in Parallel*. University of Chicago Press.

Crawford VP and Sobel J (1982) Strategic information transmission. *Econometrica: Journal of the Econometric Society* 1431–1451.

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

Dessein W (2002) Authority and communication in organizations. *The Review of Economic Studies* 69(4): 811–838.

Dragu T and Fan X (2016) An agenda-setting theory of electoral competition. *The Journal of Politics* 78(4): 1170–1183.

Druckman JN (2021) A framework for the study of persuasion. *Annual Review of Political Science* 25.

Dziuda W and Salas C (2018) Communication with detectable deceit. *Available at SSRN 3234695*.

Ederer F and Min W (2022) Bayesian Persuasion With Lie Detection. Technical Report National Bureau of Economic Research.

Enikolopov R, Petrova M and Zhuravskaya E (2011) Media and political persuasion: Evidence from Russia. *American Economic Review* 101(7): 3253–85.

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

Esteban-Casanelles T (2020) The Effects of Exposure to Electoral Advertising: Evidence from Spain.

Eyster E and Rabin M (2005) Cursed equilibrium. *Econometrica* 73(5): 1623–1672.

Farrell J and Gibbons R (1989) Cheap talk with two audiences. *The American Economic Review* 79(5): 1214–1223.

Fearon JD (1995) Rationalist explanations for war. *International organization* 49(3): 379–414.

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

Fudenberg D (2006) Advancing beyond advances in behavioral economics. *Journal of Economic Literature* 44(3): 694–711.

Fudenberg D, Gao Y and Pei H (2020) A reputation for honesty. *arXiv preprint arXiv:2011.07159*.

Gailmard S and Patty JW (2012) Formal models of bureaucracy. *Annual Review of Political Science* 15: 353–377.

Gehlbach S (2021) *Formal Models of Domestic Politics*. New York, NY: Cambridge University Press.

Gehlbach S and Sonin K (2014) Government control of the media. *Journal of Public Economics* 118: 163–171.

Gentzkow M and Shapiro JM (2006) Media bias and reputation. *Journal of Political Economy* 114(2): 280–316.

Gerber AS and Green DP (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 AS, Green DP and Larimer CW (2008) Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review* 102(1): 33–48.

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

Glazer J and Rubinstein A (2004) On optimal rules of persuasion. *Econometrica* 72(6): 1715–1736.

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

Grossman GM and Helpman E (2001) *Special Interest Politics*. Cambridge, MA: MIT Press.

Hall RL and Deardorff AV (2006) Lobbying as legislative subsidy. *American Political Science Review* 100(1): 69–84.

Hirsch AV, Kang K, Montagnes BP, et al. (2021) Lobbyists as gatekeepers: Theory and evidence.

Hollyer JR and Rosendorff BP (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).*

Holmström B (1999) Managerial incentive problems: A dynamic perspective. *The review of Economic studies* 66(1): 169–182.

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

Huber GA and Arceneaux K (2007) Identifying the persuasive effects of presidential advertising. *American Journal of Political Science* 51(4): 957–977.

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

Kalla JL and Broockman DE (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 E (2019) Bayesian persuasion and information design. *Annual Review of Economics* 11: 249–272.

Kamenica E and Gentzkow M (2011) Bayesian persuasion. *American Economic Review* 101(6): 2590–2615.

Kartik N (2009) Strategic communication with lying costs. *The Review of Economic Studies* 76(4): 1359–1395.

Kartik N, Ottaviani M and Squintani F (2007) Credulity, lies, and costly talk. *Journal of Economic theory* 134(1): 93–116.

Kendall C, Nannicini T and Trebbi F (2015) How do voters respond to information? evidence from a randomized campaign. *American Economic Review* 105(1): 322–53. https://www.aeaweb.org/articles?id=10.1257/aer.20131063.

Kolotilin A, Mylovanov T, Zapechelnyuk A, et al. (2017) Persuasion of a privately informed receiver. *Econometrica* 85(6): 1949–1964.

Kosterina S (2018) Persuasion with unknown beliefs. *Work. Pap., Princeton Univ., Princeton, NJ.*

Kuvalekar A, Lipnowski E and Ramos J (2022) Goodwill in communication. *Journal of Economic Theory* 105467.

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

Levy G, Moreno de Barreda I and Razin R (2022) Persuasion with correlation neglect: A full manipulation result. *American Economic Review: Insights* 4(1): 123–38.

Levy G and Razin R (2004) It takes two: An explanation for the democratic peace. *Journal of the European economic Association* 2(1): 1–29.

Lin X and Liu C (2022) Credible Persuasion. *arXiv preprint arXiv:2205.03495.*

Lipnowski E and Ravid D (2020) Cheap talk with transparent motives. *Econometrica* 88(4): 1631–1660.

Lipnowski E, Ravid D and Shishkin D (2019) Persuasion via weak institutions. *Available at SSRN 3168103*.

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

Little AT (2017b) Propaganda and credulity. *Games and Economic Behavior* 102: 224–232.

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

Luo Z (2018) Discriminatory Persuasion. *Available at SSRN 3075042*.

Luo Z and Rozenas A (2018) Strategies of election rigging: Trade-offs, determinants, and consequences. *Quarterly Journal of Political Science* 13(1): 1–28.

Martin GJ and Yurukoglu A (2017) Bias in cable news: Persuasion and polarization. *American Economic Review* 107(9): 2565–99.

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

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

Milgrom P and Roberts J (1986b) Relying on the information of interested parties. *The RAND Journal of Economics* 17(1): 18–32.

Minozzi W (2011) A jamming theory of politics. *The Journal of Politics* 73(2): 301–315.

Morris S (2001) Political correctness. *Journal of Political Economy* 109(2): 231–265.

Ottaviani M and Sorensen PN (2006) Reputational cheap talk. *RAND Journal of Economics* 37(1): 155–175.

Patty JW (2009) The politics of biased information. *The Journal of Politics* 71(2): 385–397.

Pei H (2020) Repeated communication with private lying cost. *arXiv preprint arXiv:2006.08069*.

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

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

Prato C and Turner I (2022) Institutional Foundations of the Power to Persuade.

Prendergast C (1993) A theory of yes men. *American Economic Review* 83(4): 757–770.

Rayo L and Segal I (2010) Optimal information disclosure. *Journal of Political Economy* 118(5): 949–987.

Salamanca A (2021) The value of mediated communication. *Journal of Economic Theory* 192: 105191.

Schnakenberg KE (2015) Expert advice to a voting body. *Journal of Economic Theory* 160: 102–113.

Schnakenberg KE and Turner IR (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 JM (1991) On buying legislatures. *Economics & Politics* 3(2): 93–109.

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

Spence M (1973) Job market signaling. *The Quarterly Journal of Economics* 87(3): 355–374.

Spenkuch JL and Toniatti D (2018) Political advertising and election results. *The Quarterly Journal of Economics* 133(4): 1981–2036.

Titova M (2020) Persuasion with verifiable information. Technical report UCSD Working Paper.

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

Zaller JR et al (1992) *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.