Strategic Interpretations*

Kfir Eliaz†, Ran Spiegler‡ and Heidi C. Thysen§

January 14, 2019

Abstract

We study strategic communication when the sender can influence the receiver's understanding of messages' equilibrium meaning. We focus on a "pure persuasion" setting, in which the informed sender wants the uninformed receiver to always choose "accept". The sender's strategy maps each state of Nature to a distribution over pairs consisting of: (i) a multi-dimensional message, and (ii) a "dictionary" that credibly discloses the state-dependent distribution of some of the messsage's components. The receiver does not know the sender's strategy by default; he can only interpret message components that are covered by the dictionary he is provided with. We characterize the sender's optimal persuasion strategy and show that full persuasion is possible when the prior on the acceptance state exceeds a threshold that quickly decreases with message dimensionality. We extend our analysis to situations where interpretation of messages is done by a third party with uncertain preferences, and explore alternative notions of "dictionaries".

*Financial support by ERC Advanced Investigator grant no. 692995 is gratefully acknowledged. We thank Xiaosheng Mu and seminar audiences at Columbia, Penn and the Warwick Economic Theory conference for helpful comments.

†School of Economics, Tel-Aviv University and Economics Dept., Columbia University. kfire@post.tau.ac.il.

‡School of Economics, Tel-Aviv University and Economics Dept., University College London and CFM. E-mail: rani@tauex.tau.ac.il.

§London School of Economics, h.c.thysen@lse.ac.uk.

1 Introduction

In the simplest textbook model of strategic communication, originated by Crawford and Sobel (1982), a "sender" privately observes a state of Nature and chooses a costless message from some given message space. Another agent, referred to as the "receiver", observes the message and takes an action that affects both parties' payoffs. The situation is thus modeled as a two-stage game with one-sided incomplete information.

A hallmark of this conventional approach is that messages have no intrinsic meaning. Their content - namely, the inference the receiver draws from them - is established in Nash equilibrium. Under the standard steady-state interpretation of this solution concept, the receiver has access to a rich dataset that fully reveals the true statistical relation between states and messages. As a result, the receiver knows the meaning of the sender's equilibrium messages and does not need anyone to interpret them for him. (Interpreting out-of-equilibrium messages requires other modes of inference, captured by refinements of Nash equilibrium.)

In this paper we revisit the basic sender-receiver model of strategic communication and depart from the assumption that the receiver is fully capable of interpreting equilibrium messages. We focus on a pure persuasion setting (as in Glazer and Rubinstein (2004, 2006) or Kamenica and Gentzkow (2011)): the receiver has two available actions, "accept" and "reject"; the sender wants the receiver to accept regardless of the state, whereas the receiver wants to accept only in the "acceptance state" (the prior probability of which is $\pi < 1/2$). By default, the receiver lacks access to any data that would shed light on the relation between states and messages. Therefore, he cannot decipher messages by himself; he is like a tourist in a foreign country who does not understand the local language or cultural codes. However, if someone came along and handed the receiver a "*dictionary*" containing some data regarding the statistical steady-state mapping between states and messages, he would have some ability to interpret the message he encounters.

The model we construct thus extends the basic sender-receiver model by making room for the supply of "dictionaries" that provide partial interpretation of equilib-

rium messages. A key feature of our model is that interpretations are *strategic*. The sender himself, or a third party whose preferences may be aligned with the sender's, chooses a dictionary from some feasible set - each *credibly* providing some statistical data regarding the equilibrium state-message mapping. Given a realized message-dictionary pair, the receiver updates his beliefs regarding the state given the message. The receiver's inferences are purely based on the data in the dictionary; he has no other means for extracting the meaning of messages. In addition, the receiver cannot draw inferences from the dictionary itself - i.e., he has no means of updating his beliefs just from the mere fact that he received a particular dataset out of the possible datasets he could have received (unless he also receives data on the statistical steady-state mapping between states and dictionaries, which is an extension we discuss in Section 5.2).

For a concrete example of the kind of situations that motivate our model, consider an employee who is up for promotion and wants to exert effort only when sufficiently confident that he will be promoted. The employee is summoned to the General Manager's office to hear about his prospects at the company. After the meeting is over, the Human Resources manager (who was present at the meeting) explains that when the General Manager says to an employee that "he has a future in the company", this means a 50% chance of getting a promotion. This amounts to an interpretation of the General Manager's verbal message. Yet the HR manager could ignore the General Manager's verbal message altogether and only interpret his body language: "The GM's handshake was feeble; this is definitely bad news". Alternatively, suppose the sender is a political party and the receiver is a representative voter. The party's message is multi-dimensional: each component describes public pronouncements by a different party member. The party's message is interpreted by a media outlet that provides historical data about the match between pronouncements by selected party members and the underlying reality.

In these examples, interpretation is *selective* because it focuses on particular aspects of the sender's multi-dimensional message; and it is *strategic* because the interpreter's interests are not necessarily aligned with the receiver's. One could argue that in both examples, the statistical data the interpreter provides need not be

perfectly credible or unbiased. We abstract from these considerations; our analytical task is to quantify the effect of strategically selective interpretations on the persuasion problem, assuming the statistical data these interpretations involve is accurate. We study the problem under various specifications of the notion of a dictionary and show that strategic interpretations greatly enhance the sender's ability to persuade the receiver.

Preview of the analysis

In Sections 2-4, we present and analyze a model in which a dictionary is defined as a selection of the components of a K-dimensional message. The dictionary interprets these components by disclosing their joint distribution conditional on the state. We show that when the sender himself (or a proxy with fully aligned interests) interprets messages, he can attain full persuasion (i.e. the receiver chooses "accept" with probability one), as long as π is above a cutoff $\pi^*(K)$ that is given by a simple formula. Moreover, this cutoff quickly drops towards zero as K grows larger. Note that full persuasion means that it does not matter whether the sender is able to fully commit to a communication strategy. When the interpreter's preferences are not fully aligned with the sender's, or when $\pi < \pi^*(K)$, full persuasion is not possible. In this case, we characterize the maximal probability of persuasion and describe how key features of the sender's strategy vary with π and the interpreter's preferences.

In Section 5 we examine alternative notions of dictionaries, which allow the sender (or his proxy interpreter) to provide data about other slices of the joint equilibrium distribution over states, messages and dictionaries. For illustration, suppose that every message m consists of two components, m_1 and m_2. The interpreter can provide data about how the individual variables m_1 and m_2 are distributed conditional on the state of Nature, without disclosing data about the *joint* conditional distribution of m_1 and m_2. Alternatively, the interpreter can provide data about the marginal distribution of m_1 (without disclosing how it depends on the state), as well as data about the conditional distribution of m_2 given m_1 and the state. Finally, the sender's dictionary can also provide data about the distribution of *dictionaries* conditional on the state. Such data enables the receiver to draw partial inferences from the type of dictionary he is equipped with.

In all these cases, we need to define how the receiver extrapolates a subjective belief from the marginal and conditional distributions that the sender's dictionary discloses. Following Spiegler (2018), we employ the *maximum entropy* principle - that is, the receiver's (unconditional) subjective belief maximizes (Shannon) entropy subject to the data he receives. This generalization subsumes the basic model of Section 2 as a special case and extends the motivating idea behind it - namely, that the receiver does not infer correlations beyond what his data tells him. We examine, via a series of examples, whether these more elaborate notions of dictionaries enable the sender to outperform the result in Section 3.

Finally, in Section 6 we present and analyze an alternative model of "non-Bayesian" persuasion (suggested to us by Xiaosheng Mu) that is formally related to our basic model. In this model, the sender's strategy induces a non-partitional information structure for the receiver, whose inferences are naive and violate the standard introspection axioms that characterize partitional information structures.

2 A Model

There are two players, a sender and a receive. The sender observes a state of nature $\theta \in \Theta = \{Y, N\}$. The receiver does not observe the state but needs to take an action a, which can be either "yes" or "no". With slight abuse of notation, we denote these actions by $Y$and N, respectively. The two players' payoffs take the values 0 and 1. The sender's payoff is 1 if and only if $a = Y$, while the receiver's payoff is 1 if and only if $a = \theta$. That is, the sender would like the receiver to choose Y in any state, whereas the receiver wants to choose Y if only if $\theta = Y$.

The players' common prior belief over Θ assigns probability $\pi < \frac{1}{2}$ to state Y. Hence, in the absence of any further information, the receiver's optimal action is N. To persuade the receiver to choose Y with some probability, the sender must convey some information about the state. He does so by committing to a strategy that maps each state to a distribution over *reports,* where a report is a pair (m, D) such that (i) $m = (m_1, ..., m_K) \in \{0, 1\}^K$ is a K-dimensional *message* with $K \geq 1$, and (ii) $D \in 2^{\{1,...,K\}}$ is a *dictionary*. Hence, the sender's strategy is a function

$\sigma : \Theta \to \Delta\left(\{0,1\}^K \times 2^{\{1,\ldots,K\}}\right)$. Since the receiver only has two available actions, the assumption that states and message components are binary is without loss of generality and is made for notational simplicity. The probability with which the sender plays the report (m, D) in state θ is denoted $\sigma(m, D \mid \theta)$. With slight abuse of notation, define $\sigma(m \mid \theta) = \sum_D \sigma(m, D \mid \theta)$.

Multi-dimensionality of messages has several interpretations. First, different components of m may represent different communication modes: verbal statements, body language, voice intonation, etc. When the sender represents an organization, different message components can represent utterances made by different organs (party members, corporate executives, spokepersons). Alternatively, we can relax the binary-state assumption (which entails no loss of generality, as mentioned above) and allow the state itself to have multiple dimensions, such that each message component corresponds to a different dimension of the state.

The role of dictionaries is to grant the receiver "partial access" to the statistical regularities inherent in the sender's strategy. Specifically, when the receiver observes the report (m, D), he learns the collection of conditional distributions $\{\sigma(m_D \mid \theta)\}_{\theta \in \Theta}$, where $m_D = (m_k)_{k \in D}$ and

$$\sigma(m_D \mid \theta) = \sum_{m' \mid m'_D = m_D} \sigma(m' \mid \theta)$$

That is, the receiver learns how the message components in D - *and nothing but them* - are distributed conditional on the state. He cannot make sense of any other aspect of the report - i.e., the message components $m_{\{1,\ldots,K\}-D}$ and the dictionary D itself. Consequently, when $D \neq \emptyset$, the receiver arrives at the following updated belief that the state belongs to Y:

$$\tilde{\Pr}(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m_D \mid \theta = Y)}{\pi \cdot \sigma(m_D \mid \theta = Y) + (1 - \pi) \cdot \sigma(m_D \mid \theta = N)} \tag{1}$$

If $D = \emptyset$, then the receiver cannot interpret the sender's report and therefore assigns his prior belief π to state Y. Compare this definition of the receiver's subjective belief with the correct, rational-expectations posterior probability of Y conditional

on (m, D):

$$\Pr(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m, D \mid \theta = Y)}{\pi \cdot \sigma(m, D \mid \theta = Y) + (1 - \pi) \cdot \sigma(m, D \mid \theta = N)} \tag{2}$$

Note that providing the full dictionary $D^* = \{1, ..., K\}$ does not automatically endow the receiver with rational expectations because such a dictionary does not interpret *itself*; it enables the receiver to draw correct inferences from m, but not from the entire report (m, D). Note also that the distribution over the receiver's posterior $\widetilde{\Pr}(\theta = Y \mid m, D)$ need *not* satisfy Bayes plausibility.

The receiver best-replies to this subjective posterior belief. Equivalently, faced with a report (m, D), he computes its subjective likelihood ratio

$$\rho_\sigma(m, D) = \frac{\sum_{m' | m'_D = m_D} \sigma(m' \mid \theta = Y)}{\sum_{m' | m'_D = m_D} \sigma(m' \mid \theta = N)} \tag{3}$$

and chooses $a = Y$ if and only if $\rho_\sigma(m, D) \geq (1 - \pi)/\pi$.

Discussion: The interpretation of interpretations

The notion of a dictionary in our model formalizes the idea that the receiver has an imperfect understanding of the sender's strategy, which limits his ability to draw inferences from messages; and moreover, that this imperfect understanding is *endogenously determined*.

Under the steady-state view of equilibrium behavior, the sender's strategy σ describes a long-run statistical relation between states and messages. The receiver moves once, against the background of a large dataset consisting of many realizations of $(\theta, m_1, ..., m_K, D)$, resulting from previous interactions between the sender (or different senders having identical objectives) with different receivers. The dataset can be visualized as a large spreadsheet, where each column represents one of the variables and each row represents an observation (an independent draw from the joint distribution over states and reports). Rational expectations correspond to having full access to this dataset. Our model relaxes this assumption and assumes that the receiver is granted access to a subset of columns. The receiver can only rely on the

accessed data for drawing inferences.

Given that we model the situation as a two-player game, a literal interpretation of our model would be that the sender himself interprets his own messages. For example, a General Manager may communicate to an employee that "he has a future in this company" and then add "in the past, when I used this term, that meant a 50% chance of getting a promotion" and provide numerous verifiable examples that substantiate this claim. Similarly, a political candidate campaigning for office may call to "drain the swamp", and then list the bills for government cuts that he initiated or voted for.

Our preferred interpretation is that the two-player model is a reduced form of a larger model in which interpretation is done by a *third party* whose preferences are aligned with the sender's: an accomplice, a spokesperson or a captured media outlet. In reality, such third parties provide selective data that illuminate the meaning of utterances by the agent they serve.[1] The data are quantitative and verifiable, and therefore it is reasonable to assume they are relatively credible - unlike the messages themselves, which are pure "cheap talk".

We could turn the interpreter into an actual third player, producing the following timeline. The sender moves first by choosing a strategy that maps each θ to a distribution over m. The interpreter moves after observing m, and chooses D; unlike the receiver, he has rational expectations. The conditional distribution σ over pairs (m, D) is induced by the sender's and interpreter's strategies. The receiver moves last, having observed the history (m, D), and he best-replies to the belief (1). If the sender and the interpreter have common interests, the situation can be reduced to our two-player formulation, once our notion of equilibrium is appropriately extended to the three-player interaction. The difference is that the three-player model requires σ to satisfy the conditional-independence property $D \perp \theta \mid m$. In Section 3 we will see there is no loss of generality in imposing this property directly on the two-player model, rendering its equivalence to the three-player formulation exact.

[1] The influence of biased media on voters' behavior is discussed in Prat (2018), who argues that media owners with political motives may use their outlets to back their preferred candidates. Strategic interpretation by news commentators can serve this goal.

3 The Basic Result

We begin this section by analyzing our model under rational expectations. In this case, which coincides with Kamenica and Gentzkow's (2011) "prosecutor" example, the probability of persuasion is maximized by the following strategy: In state Y send the message $(1, ..., 1)$ with probability one, and in state N randomize by sending the message $(1, ..., 1)$ with probability $\pi/(1-\pi)$ and the message $(0, ..., 0)$ with the complementary probability. When the receiver gets the message $(0, ..., 0)$, he infers that $\theta = N$ for sure and takes the action N. When he receives the message $(1, ..., 1)$, his posterior belief that $\theta = Y$ is

$$\Pr(\theta = Y \mid m = (1, ..., 1)) = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{\pi}{1-\pi}} = \frac{1}{2}$$

such that he is just willing to play Y. Consequently, the overall probability of persuasion is

$$\pi + (1-\pi) \cdot \frac{\pi}{1-\pi} = 2\pi$$

This result crucially relies on the sender's ability to *commit* to a strategy ex-ante. Without the ability to commit, the probability of persuasion would be *zero* in any Nash equilibrium.

The following example demonstrates that in contrast to the rational-expectations benchmark, our model enables *full* persuasion as an equilibrium outcome.

An example: Full persuasion under $K = 2$

Consider the following sender strategy. In the state Y, he randomizes uniformly between the reports $((1,1),\{1\})$ and $((1,1),\{2\})$ - i.e., he sends the message $(1,1)$ with probability one and interprets a random component. In the state n, the sender randomizes uniformly between the reports $((1,0),\{1\})$ and $((0,1),\{2\})$.

Following every report $(m,\{k\})$ that is played with positive probability under the sender's strategy, the receiver's posterior belief, $\widetilde{\Pr}(\theta = Y \mid m, \{k\})$, is equal to

$$\frac{\pi \cdot \sigma(m_k = 1 \mid \theta = Y)}{\pi \cdot \sigma(m_k = 1 \mid \theta = Y) + (1-\pi) \cdot \sigma(m_k = 1 \mid \theta = N)} = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{1}{2}}$$

Therefore, the receiver weakly prefers playing Y after each of these four reports whenever $\pi \geq \frac{1}{3}$.

The example illustrates the following key points. First, since the sender achieves full persuasion, his strategy would also constitute an equilibrium in the *absence* of commitment. The reason is that the receiver plays Y after any realized report, hence the sender has no incentive to deviate from any realization of his strategy. This is in sharp contrast to the rational-expectations benchmark.

Second, the receiver lacks rational expectations: his beliefs are determined by the strategically chosen dictionaries. Therefore, when he receives the report $((0,1),\{2\})$ he cannot draw and inference from the message's first component. It is as if the sender sends the receiver a muted video clip with no captions or an illustrated instruction manual in a foreign language - in both cases the receiver can only interpret the visuals. The example raises the question of how large should dictionaries be when $K > 2$, or how many components is it optimal to interpret? Theorem 1 below answers this question.

Third, even if the receiver could draw inferences from the realized dictionary itself (i.e., the sender's decision not to interpret a particular message component), there would be no rational basis for that because according to the sender's strategy, the distribution over D is *the same* in both states - i.e., D is independent of θ. One might argue that the receiver should still be "suspicious" of the selective interpretation and therefore discount its informational content. However, this argument is unconventional: By the same token, one could argue that in Crawford and Sobel's (1982) partially informative "interval equilibria", the receiver should be suspicious of the fact that the sender only communicates an interval to which the state of Nature belongs rather than the state itself.

Finally, note that a receiver with rational expectations would not need to draw any inferences from the dictionary itself, because D is independent of θ conditional on m. Thus, such a receiver would be able to restrict attention to m; the realization of D does not reveal any additional information about the state. The following lemma establishes that this property entails no loss of generality.

Lemma 1 *Without loss of generality, we can assume that the sender's strategy satisfies the property* $D \perp \theta \mid m$.

Proof. Suppose that σ violates this property. Let p be the joint distribution over θ, m, D induced by the prior over θ and σ. Consider a deviation to a strategy σ', defined as follows. For every θ, m: $\sigma'(m \mid \theta) = \sigma(m \mid \theta)$ and $\sigma'(D \mid \theta, m) = p(D \mid m)$. The new strategy induces the same joint distributions over (θ, m) and (m, D) as σ, and yet it satisfies $D \perp \theta \mid m$. Because the receiver's subjective likelihood ratio of any report (m, D) is only a function of the distribution over (θ, m), the receiver's strategy remains unchanged after the deviation. And since the distribution over (m, D) remains unchanged, the ex-ante distribution over the receiver's action does not change, which means that the overall probability of persuasion is unchanged. ■

This lemma serves two roles. First, it substantiates the three-player interpretation of our model (described at the end of Section 2), since a distinct interpreter would only be able to condition D on m. Second, the extended model in Section 4 will impose the conditional-independence property as an assumption from the outset, hence it is useful to know that it is without loss of generality in the basic model of Section 2.

We are now ready to state the main result of this section. The result makes use of the following notation, which will also prove useful in later sections:

$$\begin{aligned} S &= \binom{K}{\lfloor K/2 \rfloor} \\ \mathcal{B}^* &= \left\{ (m, D) \mid \sum_k m_k = \left\lfloor \frac{K}{2} \right\rfloor \ ; \ D = \{k \mid m_k = 1\} \right\} \end{aligned}$$

Note that $|\mathcal{B}^*| = S$. We do not provide the proof here because it is a special case of the results we prove in Section 4.

Theorem 1 *The maximal probability of persuasion is* $\min\{1, \pi(1+S)\}$. *It can be implemented by the following strategy:*

$$\sigma((1,\ldots,1), D \mid \theta = Y) = \frac{1}{S} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor$$

$$\sigma(m, D \mid \theta = N) = \min\{\frac{1}{S}, \frac{\pi}{1-\pi}\} \text{ for every } (m, D) \in \mathcal{B}^*$$

$$\sigma((0,\ldots,0), D \mid \theta = N) = \max\{0, \frac{1}{S} - \frac{\pi}{1-\pi}\} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor$$

Furthermore, when $\pi \geq 1/(1+S)$, *this strategy is time-consistent.*

The strategy outlined by this result extends the example. In state Y, the sender sends a single message, which we conveniently select to be $(1, ..., 1)$. Each of the components of this message can therefore be regarded as "good news". What happens in state N depends on the relation between the prior π and the number S, which itself depends on K. Suppose K is even, for the sake of the argument here. If $\pi \geq 1/(1+S)$, the sender randomizes uniformly over $\mathcal{B}^*$, which is the set of all messages that consist of an equal number of 1's ("good news") and 0's ("bad news"). Crucially, the dictionary that accompanies each of these messages interprets *only the good news*. If $\pi < 1/(1+S)$, each of these message-dictionary pairs is played with probability $1/S$, and the remaining probability is allocated to the message $(0, ..., 0)$ - i.e. all "bad news".

Unlike the case of the "mixed" messages in $\mathcal{B}^*$, there is considerable freedom in selecting the dictionaries that accompany the "pure" messages $(1, ..., 1)$ and $(0, ..., 0)$. Our construction has the property that the distribution over D conditional on each of these messages is the same as conditional on $\mathcal{B}^*$. Consequently, the strategy has the property that D is unconditionally independent of θ - beyond the *conditional* independence property $D \perp \theta \mid m$, which was established by Lemma 1. This means that even if the receiver could draw inferences from D, he would be unable to learn anything about θ from the realization of D itself. With regards to the question of how large dictionaries should be, the optimal strategy described in Theorem 1 achieves the highest probability of persuasion by interpreting *half* of the components.

Now examine the receiver's reaction to various realizations of the sender's strategy. When he confronts the message $(0, ..., 0)$, each of the realizations of D interprets some "bad news" revealing that $\theta = N$. In contrast, every other realization of (m, D) satisfies $m_k = 1$ for all $k \in D$ and enables him to interpret m_D. He thus learns that the probability of m_D conditional on $\theta = Y$ is one, whereas the probability of m_D conditional on $\theta = N$ is $\min\{1/S, \pi/(1-\pi)\}$. Therefore, his subjective likelihood ratio of (m, D) is

$$\rho_\sigma(m, D) = \frac{1}{\min\{\frac{1}{S}, \frac{\pi}{1-\pi}\}}$$

which is, by definition, weakly above $(1-\pi)/\pi$ and therefore persuades the receiver.

The basic intuition behind this result is basic: strategic interpretation of mixed messages is often *selective*, explaining the meaning of the good news while remaining silent about the bad news. A receiver with rational expectations would realize that the mixed messages in $\mathcal{B}^*$ only occur in state N. However, our receiver can only draw inferences from message components that the sender interprets for him. But since the sender only interprets the components that constitute good news, this selective interpretation manages to convey a false sense that the mixed message is actually good news. Moreover, as K gets large, each $(m, D) \in \mathcal{B}^*$ identifies a *distinct pattern* that is increasingly rare in state N yet occurs with probability one in state Y. Therefore, even if π is quite small and even if $\mathcal{B}^*$ is played with high probability in state N, the receiver will be persuaded by the reports in $\mathcal{B}^*$.

When $\pi \geq 1/(1+S)$, the sender is able to attain full persuasion. This means that the sender's strategy is *time-consistent*: since the receiver plays Y after every report, the sender would not want to deviate from any realized report even if he could. In other words, the assumption that the sender has commitment power is not required in this range of parameters.

The proof of Theorem 1 and Sperner's Theorem

Suppose that the feature that only one message is played in state Y is taken for granted. As before, we can let this message be $(1, ..., 1)$.[2] Then, in order to persuade

[2]The proof of the theorem is considerably more involved than the following sketch may suggest, precisely because we are unable to establish *at the outset* that this assumption entails no loss of

the receiver, any report (m, D) that is played in state N must satisfy $m_A = 1$ for some non-empty collection of components A. If the dictionary that accompanies this message interprets components outside A, the receiver will immediately infer that $\theta = N$. Therefore, the sender will only want to interpret components *inside* A. That is, he will accompany m with a dictionary $D \subseteq A$. Moreover, making this subset larger will only increase the rarity of the pattern m_D in state N without changing the fact that the pattern occurs for sure in state Y; this will serve to increase the receiver's posterior on Y. It follows that the sender will choose $D = A$. In other words, every message m that is played in state N and has at least *some* good news will pin down the dictionary that accompanies it to be the set of 1 components in m.

Now consider two reports $(m, D), (m', D')$ that are played in state N, and suppose $D' \subset D$. Our argument in the previous paragraph then implies $m'_{D'} = m_{D'}$ - i.e. the pattern $m'_{D'}$ also appears in the message m. This means that the receiver considers both reports when calculating his subjective likelihood ratio of (m', D'). If the sender shifted all the weight from (m, D) to (m', D'), this subjective likelihood ratio would remain unchanged. Repeatedly applying this argument, we can conclude that without loss of generality, the sender's optimal strategy satisfies the following property: the collection of dictionaries that are played in state N as part of a persuasive report constitutes an *anti-chain* - i.e., no dictionary contains another. The sender would want this anti-chain to be as large as possible, because this will make the pattern highlighted by each dictionary increasingly rare in state N. A basic result in extremal combinatorics, known as Sperner's Theorem (see Anderson (1987), pp. 2-4), establishes that the largest anti-chain consisting of subsets of $\{1, ..., K\}$ is the set of all subsets of size $\lfloor K/2 \rfloor$. This explains the role of S and $\mathcal{B}^*$ in our result.

Comment: Restricting the set of available dictionaries

Theorem 1 is based on the assumption that the sender can interpret *any* collection of message components of size $\left\lfloor \frac{K}{2} \right\rfloor$. If we think of a dictionary as a physical dataset that the sender can grant access to, our assumption means that when K is large, the

generality; we are able to derive it only later in the course of the proof.

sender has an incredibly large set of datasets at his disposal. Such richness becomes increasingly unrealistic as K grows larger, and we may wish to restrict further the set of available dictionaries. For instance, suppose that the sender can only interpret *individual* message components - i.e. the set of available dictionaries is the set of singletons $\{k\}$, $k = 1, ..., K$. In this case, the sender can attain full persuasion whenever $\pi > 1/(1+K)$; in state N, he would randomize uniformly over all reports (m, D) that satisfy $m_k = 1$ for a unique k and $D = \{k\}$.

4 An Interpreter with Uncertain Motives

Our interpretation that dictionaries are provided by a third player - a strategic "interpreter" - raises a natural question: What if the interpreter's interests are aligned with the receiver? For instance, if we view media outlets as interpreting messages by politicians to voters, then some outlets are partisans while others are objective, and a politician does not necessarily know which outlet the voter follows.[3] While a biased media outlet may be strategically selective in its interpretation (by failing to provide data about messages that address certain issues, or by ignoring messages by particular party members), an objective media outlet would not pick and choose which message components to interpret, but rather interpret them all.

Uncertainty over the interpreter's motives introduces two novel complications to the sender's problem: (i) *multiple interpretations* - a given message may be understood differently by receivers who follow interpreters with opposing interests, and (ii) *externalities* - when the interpreter sides with the receiver, the dictionary that accompanies one message may highlight patterns that appear in other messages (a feature that does not occur under the optimal strategy in the basic model, as our discussion of the proof of Theorem 1 at the end of Section 3 demonstrated). How does the sender cope with these difficulties when designing his optimal strategy?

To address this question, we extend the model of Section 2 by assuming that the receiver may be of *two* types: "rational" with probability λ or "non-rational" with

[3]Kennedy and Prat (2017) provide empricial evidence on the diversity of media outlets that individuals get their news from.

probability $1-\lambda$. A non-rational receiver behaves exactly as in Section 2. That is, given a sender strategy σ, his subjective posterior belief following a report (m, D) is given by (1). In contrast, a rational receiver knows σ and therefore forms the correct, rational-expectations posterior (2).

We view this extended model as a reduced-form approach to accommodating interpreters with opposing interests: with probability λ the interpreter sides with the receiver and with probability $1-\lambda$ he sides with the sender. Recall that to substantiate the three-player interpretation of our model, we showed in Section 3 that when $\lambda = 0$, there is no loss of generality in restricting attention to strategies that satisfy the conditional-independence property $D \perp \theta \mid m$ (since the third player - the interpreter - chooses the dictionary after observing the message but not the state). Whether this conclusion remains true when $\lambda > 0$ is an open question (we conjecture that it does). Hence, to maintain the three-player interpretation of the extended model with $\lambda > 0$, we *assume* that the sender is restricted to strategies that satisfy $D \perp \theta \mid m$. The proof of the results in this section makes subtle use of this property.

When the interpreter's interests are aligned with the sender's, he will use the dictionary that the sender would have wanted him to use. In this case, the receiver will act like the non-rational receiver in the reduced two-player model. When the interpreter's interests are aligned with the receiver's, he wishes to impart his rational expectations to the receiver, using an appropriate dictionary. This would be the full dictionary $D^* = \{1, ..., K\}$. The reason is that since $D \perp \theta \mid m$, the rational-expectation inference of θ from the report (m, D) is based entirely on m. The rational-expectations posterior can thus be written as

$$\Pr(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m \mid \theta = Y)}{\pi \cdot \sigma(m \mid \theta = Y) + (1-\pi) \cdot \sigma(m \mid \theta = N)} = \overset{\sim}{\Pr}(\theta = Y \mid m, D^*)$$

That is, a receiver who obtains the full dictionary D^* from the interpreter would be endowed with rational expectations, and therefore he corresponds to a rational receiver in the reduced-form, two-player model.

Fix the sender's strategy σ. We say that (m, D) *persuades a non-rational receiver*

if $\rho_\sigma(m,D) \geq (1-\pi)/\pi$, where $\rho_\sigma(m,D)$ is the subjective likelihood ratio that a non-rational receiver assigns to the report (m,D), as defined by (3). We say that (m,D) persuades a rational receiver if

$$\rho^*_\sigma(m,D) = \frac{\sigma(m,D \mid \theta = Y)}{\sigma(m,D \mid \theta = N)} = \frac{\sigma(m \mid \theta = Y)}{\sigma(m \mid \theta = N)} = \rho_\sigma(m,D^*) \geq \frac{1-\pi}{\pi}$$

where $\rho^*_\sigma(m,D)$ is the likelihood ratio that a rational receiver assigns to the report (m,D). The second equality is due to the restriction that $D \perp \theta \mid m$ under the sender's strategy σ.

The probability of persuasion induced by a sender strategy σ is the probability that the receiver chooses Y, given our model of how each receiver type responds to reports. Denote

$$\pi^* = \frac{1}{1+S} \qquad\qquad \lambda^* = 1 - \frac{1}{S}$$

We now present a complete characterization of the maximal probability of persuasion (and strategies that implement the optimum), for three different ranges of the values of the parameters λ, π.

Theorem 2 *If $\lambda \geq \lambda^*$, then, the maximal probability of persuasion is 2π, and the following strategy implements this optimum:*

$$\begin{aligned}
\sigma((1,\ldots,1),D^* \quad &| \quad \theta = Y) = 1 \\
\sigma((1,\ldots,1),D^* \quad &| \quad \theta = N) = \frac{\pi}{1-\pi} \\
\sigma((0,\ldots,0),D^* \quad &| \quad \theta = N) = \frac{1-2\pi}{1-\pi}
\end{aligned}$$

When $\lambda \geq \lambda^*$, it is very likely that the receiver is rational. Therefore, the sender focuses on persuading the rational type as often as possible. By doing so, he persuades the non-rational type whenever he persuades the rational type, but he is unable to take advantage of the former's limitation.

Theorem 3 *If $\lambda < \lambda^*$ and $\pi \leq \pi^*$, then the maximal probability of persuasion is $\pi(1 + (1-\lambda)S)$, and the following strategy implements this optimum:*

$$\begin{aligned}
\sigma((1,\ldots,1),D \;\;|\;\; \theta = Y) &= \frac{1}{S} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor \\
\sigma(m,D \;\;|\;\; \theta = N) &= \frac{\pi}{1-\pi} \text{ for every } (m,D) \in \mathcal{B}^* \\
\sigma((0,\ldots,0),D \;\;|\;\; \theta = N) &= \frac{1}{S} - \frac{\pi}{1-\pi} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor
\end{aligned}$$

When $\lambda < \lambda^*$ and $\pi \leq \pi^*$, the sender focuses on persuading the non-rational receiver and exploits this type's inability to make inferences about uninterpreted message components. That is, for this parameter range it is not optimal to try and persuade both the rationals and the non-rationals. In particular, conditional on $\theta = N$, he assigns probability $S\pi/(1-\pi)$ to the set of reports $\mathcal{B}^*$; with the remaining probability, the sender submits a message consisting of "bad news". Recall that what characterizes a report in $\mathcal{B}^*$ is that the message has $\left\lfloor \frac{K}{2} \right\rfloor$ components that constitute "good news" and the dictionary interprets only these components. A rational receiver type recognizes that such a report implies $\theta = N$, whereas a non-rational type finds the report persuasive.

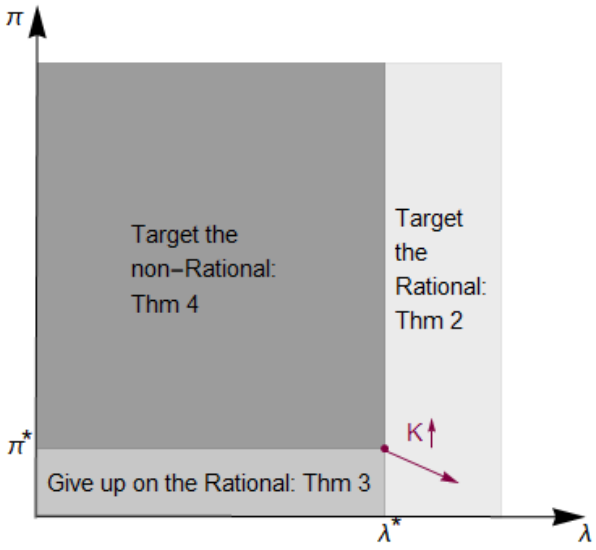
Theorem 4 *If $\lambda < \lambda^*$ and $\pi > \pi^*$, then the maximal probability of persuasion is $1 - \lambda S(1-2\pi)/(S-1)$, and the following strategy implements this optimum:*

$$\begin{aligned}
\sigma((1,\ldots,1),D \;\;|\;\; \theta = Y) &= \frac{1}{S} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor \\
\sigma(m,D \;\;|\;\; \theta = N) &= \frac{1-2\pi}{(1-\pi)(S-1)} \text{ for every } (m,D) \in \mathcal{B}^* \\
\sigma((1,\ldots,1),D \;\;|\;\; \theta = N) &= \frac{1}{S} - \frac{1-2\pi}{(1-\pi)(S-1)} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor
\end{aligned}$$

When $\pi > \pi^*$, the sender can persuade a non-rational receiver with probability one. Moreover, if he played a strategy that exclusively targets the non-rational re-

ceiver, he could ensure that this type will *strictly* prefer to accept - i.e. his subjective likelihood ratio will exceed $(1-\pi)/\pi$ for every report he encounters. The sender can use this slack to play a different strategy that increases the probability of persuading the rational type. Note that when $\lambda = 0$ (as in Section 3), the sender's use of the slack is payoff-irrelevant. For instance, the strategy outlined by Theorem 1 uses the slack to increase the probability of $\mathcal{B}^*$ in state N, compared with the strategy given by Theorem 4. Therefore, the two strategies are slightly different, yet both attain full persuasion when $\lambda = 0$ and $\pi \geq \pi^*$.

In Theorems 3-4 (as in Section 3), the sender's strategy assigns the same distribution over D in each state - it is uniform over $\mathcal{B}^*$. Therefore, the sender's use of dictionaries per se does not convey any information about the underlying state.



(Figure 1)

The strategies outlined by Theorems 2-4 differ in their targeting of receiver types. Figure 1 summarizes these differences. When $\lambda \geq \lambda^*$, the sender ignores the non-rational type and behaves as if the receiver has rational expectations. In contrast,

when $\lambda < \lambda^*$, his strategy targets the non-rational type and maximizes the probability of persuading him. In particular, when $\pi \leq \pi^*$, he ignores the rational type and behaves as if $\lambda = 0$; but when $\pi > \pi^*$, he can fully persuade the non-rational type and has enough slack to increase the probability of persuading the rational type.

Sketch of the proof of Theorems 2-4

The proof of Theorems 2-4 is given in Section 4.1 and proceeds stepwise. First, we simplify the construction of the sender's optimal strategy σ by noting that if the rational receiver is persuaded by a report (m, D), we can set $D = D^*$ without loss of generality. Similarly, if a report (m, D) does not persuade any receiver type, we can set $D = \emptyset$ without loss of generality. Second, we show that under σ, the set of reports that only persuade a non-rational receiver (denoted $\mathcal{B}_\sigma$) does not contain redundancies, in the following sense: for every pair of distinct reports $(m, D), (m', D') \in \mathcal{B}_\sigma$, $m_D \neq m'_D$ - i.e., the non-rational receiver would not be persuaded by the reports (m', D) or (m, D'). In particular, this implies that $D \subsetneq D'$. Our third step establishes that under σ, all reports in $\mathcal{B}_\sigma$ induce the same subjective likelihood ratio $(1 - \pi)/\pi$ for a non-rational receiver. Next, we show that without loss of generality, the optimal strategy satisfies the following restriction: Whenever $\mathcal{B}_\sigma$ is non-empty, it is impossible for both dictionaries D^* and $\emptyset$ to be played with positive probability in state N. Finally, we use all these steps to calculate an upper bound on the probability of persuasion, and show that it is attained by the strategy described in the three theorems.

When $\lambda > 0$, the sender's problem is complicated by the fact that his messages are interpreted differently by different receiver types. As observed at the end of Section 3, when $\lambda = 0$, the reports that are sent in state N (which persuade only a *non-rational* receiver) have the property that the dictionaries in these reports highlight a pattern that does *not* appear in *any* other message that is sent in state N. This is no longer the case when $\lambda > 0$, where the sender may send messages in state N that persuade both receiver types; and these messages may contain patterns that are highlighted by dictionaries that accompany other messages. This makes the proof of the $\lambda > 0$ case more intricate.

The additional complication is especially prominent in one of the key Lemmas

in the proof (Lemma 5), which shows that when $\mathcal{B}_\sigma \neq \emptyset$, optimal persuasion can be achieved by *either* sending a full dictionary in state N with some probability, *or* by sending a null dictionary in N with some probability, but *not both*. In particular, when $\pi > \pi^*$, the sender can send reports in state N that persuade both receiver types. He does so by shifting weight between reports in $\mathcal{B}_\sigma$ and reports with both null and full dictionaries. However, since the full dictionary highlights patterns that appear in messages other than the one it accompanies, this weight-shifting potentially reduces the likelihood ratio of reports that include a full dictionary below the acceptance threshold. The shift must therefore be carefully constructed to avoid this pitfall. This "externality" across reports makes the case of $\lambda > 0$ more intricate.

4.1 The Proof

The proof is presented for an arbitrary value of λ. Certain steps become redundant when $\lambda = 0$ (such that the model of this section is reduced to the basic model of Section 2). In what follows, we point out these steps, so that a reader who is only interested in the $\lambda = 0$ case can skip them.

We begin with a few preliminary definitions and basic observations that simplify notation and the construction of the sender's optimal strategy. Fix a sender's strategy σ. Let $\mathcal{B}_\sigma$ be the set of reports (m, D) that are played with positive probability conditional on $\theta = N$ and persuade a non-rational receiver but not a rational receiver. That is,

$$\mathcal{B}_\sigma = \left\{ (m, D) \mid \sigma(m, D \mid \theta = N) > 0 \text{ and } \rho_\sigma(m, D) \geq \frac{1-\pi}{\pi} > \rho^*_\sigma(m, D) \right\}$$

For any report $(m, D) \in \mathcal{B}_\sigma$ there must be some message m' such that m' is part of a report which is realized with positive probability in state Y and $m'_D = m_D$. Otherwise the non-rational receiver would not be persuaded by the report (m, D).

Definition 1 *We say that a message m' **justifies** the pair (m, D) if: (i) the pair (m, D) satisfies $\sigma(m, D \mid \theta = N) > 0$ and $\rho_\sigma(m, D) \geq \frac{1-\pi}{\pi}$; (ii) $\sigma(m', D' \mid \theta = Y) > 0$ for some dictionary D', and (iii) $m'_D = m_D$.*

In the following observation, Part (ii) is redundant when $\lambda = 0$.

Observation 1 *There is no loss of generality in assuming that σ has the following properties for every (m, D) that is played with positive probability under σ: (i) If (m, D) persuades no receiver type, then $D = \emptyset$; (ii) If (m, D) persuades a rational receiver, then $D = D^*$. In particular, every report $(m, D) \in \mathcal{B}_\sigma$ satisfies $D \neq D^*, \emptyset$; and if (m, D) persuades a rational receiver, it also persuades a non-rational receiver.*

Proof. Fix a sender strategy σ and consider a report (m, D) that is played with positive probability under σ.

First, assume that (m, D) persuades no receiver type. Suppose the sender deviates to a strategy that always replaces the report (m, D) with $(m, \emptyset)$, but otherwise coincides with σ. By definition, the probability of persuasion conditional on this report is weakly higher than under the original report. However, we need to check that the deviation does not lower the probability of persuasion conditional on other realizations (m', D'). Because the deviation does not change the distribution of messages conditional on any state, it does not affect the response of any receiver type to such (m', D').

Second, assume that (m, D) persuades a rational receiver. Suppose the sender deviates to a strategy that replaces (m, D) with (m, D^*), but otherwise coincides with σ. Since the deviation does not affect the distribution of messages conditional on any state, it does not change the response of a rational receiver to any realized report, and it does not change the response of a non-rational receiver to any $(m', D') \neq (m, D)$. However, the deviation ensures that the non-rational receiver reacts to (m, D^*) exactly like the rational receiver, because $\rho_\sigma(m, D^*) = \rho^*_\sigma(m, D)$ for every D. ■

Observation 2 *There is no loss of generality in restricting attention to strategies with the following property: If the reports (m, D) with $D \neq \emptyset$ and $(m', \emptyset)$ are both realized with positive probability under σ, then $m'_D \neq m_D$.*

Proof. Assume the contrary - i.e. $m'_D = m_D$. Suppose the sender deviates to a strategy that replaces $(m', \emptyset)$ with (m', D) but otherwise coincides with σ. By Observation 1, $(m', \emptyset)$ does not persuade any type prior to the deviation. And since the deviation does not affect the distribution of messages conditional on any state, it does not change the response of any receiver type to any report $(m'', D'') \neq (m', \emptyset)$. Therefore, the deviation weakly raises the probability of persuasion. ■

Henceforth, we will restrict attention to strategies that satisfy the properties of Observations 1 and 2. In addition, whenever we refer to a report with some generic dictionary, we mean that the dictionary is non-empty.

An important point regarding our proof strategy

We will now proceed with a sequence of lemmas that involve modifications of the sender's strategy. We will allow some of the modified strategies to violate the constraint that D and θ are independent conditional on m, but we will maintain the constraint that both receiver types do not draw inferences from D - i.e., they only use the joint distribution over (θ, m) to form beliefs for any given D. At the end, we will arrive at a strategy that does satisfy both restrictions, and this justifies our method of proof. Therefore, from now on, we can rewrite $\rho^*_\sigma(m, D)$ as

$$\rho^*_\sigma(m) = \frac{\sigma(m \mid \theta = Y)}{\sigma(m \mid \theta = N)}$$

because the rational receiver's likelihood ratio of any report (m, D) will only be a function of m.

The following lemma establishes that without loss of generality, the set $\mathcal{B}_\sigma$ of reports that only persuade the non-rational receiver has a simple structure: every dictionary that features in $\mathcal{B}_\sigma$ effectively interprets only the particular message it is coupled with.

Lemma 2 *Without loss of generality, an optimal sender strategy σ satisfies the following property: $m'_D \neq m_D$ for every pair of distinct reports $(m, D), (m', D') \in \mathcal{B}_\sigma$.*

Proof. Let σ be an optimal sender strategy. We will modify it in two phases into a new strategy that satisfies the property in the statement of the lemma and induces the same probability of persuasion.

In the first phase, we construct a partition $\{T_1, ..., T_L\}$ of $\mathcal{B}_\sigma$ as follows. For every $l = 1, 2, ...$, select an arbitrary report $(m^l, D^l) \in \mathcal{B}_\sigma - \cup_{h<l} T_h$, and define

$$T_l = \{(m, D) \in \mathcal{B}_\sigma - \cup_{h<l} T_h \mid m_{D^l} = m^l_{D^l}\}$$

Modify σ as follows. For each $l = 1, ..., L$ and any $(m, D) \in T_l$, $D \neq D^l$, shift the probability of (m, D) conditional on $\theta = N$ to the report (m, D^l). By definition, both (m, D) and (m^l, D^l) persuade only a non-rational receiver. Perform the following additional modification. By the definition of $\mathcal{B}_\sigma$, there must be a message m that justifies (m^l, D^l). That is, $m_{D^l} = m^l_{D^l}$, and there is a dictionary D such that (m, D) is played with positive probability in Y. If $D \neq D^*$, then shift the probability of every such (m, D) conditional on Y to (m, D^l). By construction, $m_{D^l} = m^l_{D^l}$. Therefore, (m, D^l) persuades a non-rational receiver. And since the deviation does not affect the distribution over messages conditional on any state, it does not change the response of any receiver type to any other realized report.

Let us now turn to the second phase. Start this phase by shifting the probability of any $(m, D^L) \in T_L$ conditional on $\theta = N$ to some report in T_L, denoted $(\tilde{m}^L, D^L)$. This effectively transforms T_L into a singleton $\{(\tilde{m}^L, D^L)\}$. By the construction of the first phase, every $(m, D^L) \in T_L$ satisfies $m_{D^L} = \tilde{m}^L_{D^L}$. Therefore, the deviation does not change the non-rational receiver's subjective likelihood ratio of $(\tilde{m}^L, D^L)$, such that he continues to be persuaded by this report. Moreover, by the construction of the first phase, for every $l < L$ and every $(m, D^l) \in T_l$, $m_{D^l} \neq \tilde{m}^L_{D^l}$. Therefore, the deviation does not affect a non-rational receiver's subjective likelihood ratio of $(m, D^l) \in T_l$ for all $l < L$.

Now suppose that for some $l < L$, we have transformed the cells $T_{l+1}, ..., T_L$ into singletons $\{(\tilde{m}^{l+1}, D^{l+1})\}, ..., \{(\tilde{m}^L, D^L)\}$ in such a manner. Suppose that there is some $(m, D^l) \in T_l$ such that $m_{D^h} \neq \tilde{m}^h_{D^h}$ for every $h > l$. Rename this report as $(\tilde{m}^l, D^l)$, and shift the probability of any (m, D^l) conditional on N to $(\tilde{m}^l, D^l)$.

Alternatively, suppose that for every $(m, D^l) \in T_l$ there is some $h > l$ such that $m_{D^h} = \tilde{m}^h_{D^h}$. For any such (m, D^l), shift its probability conditional on N to one of the reports $(\tilde{m}^h, D^h)$ satisfying $\tilde{m}^h_{D^h} = m_{D^h}$. By the same logic as in the previous paragraph, the deviation in these two alternative cases does not affect a non-rational receiver's subjective likelihood ratio of any report.

Finally, we need to check that the changes in the second phase do not affect the response of a rational receiver to a report (m, D) that persuaded him prior to these changes. By Observation 1, $D = D^*$, and there is no $D' \subset D^*$ such that (m, D') is played with positive probability. Since all the changes made in the second phase do not shift weight to messages that are not in $\mathcal{B}_\sigma$, they cannot affect the rational receiver's response to a report outside of $\mathcal{B}_\sigma$.

At the end of the second phase, $\mathcal{B}_\sigma$ has been effectively transformed into the set $\{(\tilde{m}^1, D^1)\}, ..., \{(\tilde{m}^L, D^L)\}$, which by construction satisfies the property in the lemma's statement. ■

From now on, we restrict attention to sender strategies σ that satisfy Lemma 2.

Corollary 1 *Let $(m, D), (m', D') \in \mathcal{B}_\sigma$. If there is a message m^* that justifies both (m, D) and (m', D') , then $D \not\subseteq D'$ and $D' \not\subseteq D$.*

Proof. Assume, by contradiction, that there exist $(m, D), (m', D') \in \mathcal{B}_\sigma$ that are justified by a message m^* and $D \subseteq D'$. This means that $m^*_D = m_D$ and $m^*_{D'} = m'_{D'}$. Therefore, $m_{D \cap D'} = m^*_{D \cap D'} = m'_{D \cap D'}$. But $D \cap D' = D$, which implies that $m_D = m'_D$, in contradiction to Lemma 2. ■

Corollary 2 *Let m^* be a message that is played with positive probability in state Y under σ. Then, the number of reports that m^* justifies is at most S.*

Proof. By Corollary 1, if m^* justifies two reports (m, D) and (m', D'), then D and D' do not contain one another. It follows that the set of all dictionaries that are part of reports justified by m^* constitutes an anti-chain - i.e. a collection of subsets of $\{1, ..., K\}$ that do not contain one another. By Sperner's Theorem, the maximal size of such a collection is S. ■

Lemma 3 *For every report $(m, D) \in \mathcal{B}_\sigma$ there exists a message m' that justifies (m, D), for which $\rho^*_\sigma(m') > \rho_\sigma(m, D)$. Furthermore, we can assume without loss of generality that if $\sigma(m', D' \mid \theta) > 0$, then $D' = D^*$.*

Proof. Let $(m, D) \in \mathcal{B}_\sigma$. There must be a message m' that justifies (m, D) - otherwise, (m, D) would not persuade a non-rational receiver. Assume, by contradiction, that $\rho_\sigma(m, D) \geq \rho^*_\sigma(m')$ for *any* such message m'. In addition, observe that if $m'_D = m_D$ and yet m' does not justify (m, D), then Definition 1 implies that $\sigma(m', D' \mid \theta = Y) = 0$ for any D'. It follows that for every m' for which $m'_D = m_D$,

$$\rho_\sigma(m, D) = \frac{\sum_{m''|m''_D=m_D} \sigma(m'' \mid \theta = Y)}{\sum_{m''|m''_D=m_D} \sigma(m'' \mid \theta = N)} \geq \frac{\sigma(m' \mid \theta = Y)}{\sigma(m' \mid \theta = N)} = \rho^*_\sigma(m')$$

Furthermore, the inequality is strict for $m' = m$ - otherwise, we would have $\rho^*_\sigma(m) \geq (1-\pi)/\pi$, contradicting the fact that $(m, D) \in \mathcal{B}_\sigma$ (and therefore does not persuade a rational receiver). Hence, cross multiplying the denominators and summing over all m' with $m'_D = m_D$ yields,

$$[\textstyle\sum_{m''|m''_D=m_D} \sigma(m''| \theta = Y)] \cdot [\sum_{m'_D|m'_D=m_D} \sigma(m'| \theta = N)] >$$
$$[\textstyle\sum_{m''|m''_D=m_D} \sigma(m''| \theta = Y)] \cdot [\sum_{m'_D|m'_D=m_D} \sigma(m'| \theta = N)]$$

which cannot be true since both sides of the strict inequality are identical.

Suppose that $\sigma(m', D' \mid \theta) > 0$ and yet $D' \neq D^*$. Since $\rho^*_\sigma(m') > \rho_\sigma(m, D) \geq (1-\pi)/\pi$ the rational receiver type is persuaded by (m', D'). So by Observation 1 there is no loss by deviating to a strategy in which (m', D') is replaced by (m', D^*).

■

Lemma 4 *Without loss of generality, $\rho_\sigma(m, D) = (1-\pi)/\pi$ for all $(m, D) \in \mathcal{B}_\sigma$.*

Proof. Let $(\underline{m}, \underline{D})$ and $(\bar{m}, \bar{D})$ be two reports in $\mathcal{B}_\sigma$ such that $\rho_\sigma(\underline{m}, \underline{D}) \leq \rho_\sigma(m, D) \leq \rho_\sigma(\bar{m}, \bar{D})$ for each $(m, D) \in \mathcal{B}_\sigma$. Assume that $\rho_\sigma(\underline{m}, \underline{D}) < \rho_\sigma(\bar{m}, \bar{D})$. Suppose that the sender deviates from σ to a strategy $\hat{\sigma}$ that shifts a weight of $\varepsilon > 0$ from $(\underline{m}, \underline{D})$

to $(\bar{m}, \bar{D})$ in state N. By Lemma 2, $\bar{m}_{\underline{D}} \neq \underline{m}_{\underline{D}}$ and $\underline{m}_{\bar{D}} \neq \bar{m}_{\bar{D}}$. Therefore,

$$\begin{aligned} \rho_{\hat{\sigma}}(\underline{m}, \underline{D}) &= \frac{\sum_{m|m_{\underline{D}}=\underline{m}_{\underline{D}}} \sigma(m \mid \theta = Y)}{\sum_{m|m_{\underline{D}}=\underline{m}_{\underline{D}}} \sigma(m \mid \theta = N) - \varepsilon} > \rho_{\sigma}(\underline{m}, \underline{D}) \geq \frac{1-\pi}{\pi} \qquad (4) \\ \rho_{\hat{\sigma}}(\bar{m}, \bar{D}) &= \frac{\sum_{m|m_{\bar{D}}=\bar{m}_{\bar{D}}} \sigma(m \mid \theta = Y)}{\sum_{m|m_{\bar{D}}=\bar{m}_{\bar{D}}} \sigma(m \mid \theta = N) + \varepsilon} < \rho_{\sigma}(\bar{m}, \bar{D}) \end{aligned}$$

By our initial assumption, $\rho_{\hat{\sigma}}(\underline{m}, \underline{D}) < \rho_{\hat{\sigma}}(\bar{m}, \bar{D})$ for sufficiently small ε. By (4), this implies that $\rho_{\hat{\sigma}}(\bar{m}, \bar{D}) > \frac{1-\pi}{\pi}$. By Lemma 2 $\rho_{\hat{\sigma}}(m, D) = \rho_{\sigma}(m, D)$ for every $(m, D) \in \mathcal{B}_{\sigma} - \{(\underline{m}, \underline{D}), (\bar{m}, \bar{D})\}$. Since the deviation does not involve reports outside $\mathcal{B}_{\sigma}$, it cannot lower the probability of persuading a rational receiver. It follows that the deviation weakly raises the probability of persuasion.

Therefore, we can assume without loss of generality that $\rho_{\sigma}(m, D)$ is the same for all $(m, D) \in \mathcal{B}_{\sigma}$. Suppose this likelihood ratio exceeds $(1-\pi)/\pi$. Pick an arbitrary $(m, D) \in \mathcal{B}_{\sigma}$. By Lemma 3, there exists a report (m', D^*) such that m' justifies (m, D) and $\rho^*_{\sigma}(m') > \rho_{\sigma}(m, D)$, such that a receiver of any type is persuaded by (m', D^*). Suppose the sender deviates to a strategy $\hat{\sigma}$ that shifts a weight $\varepsilon > 0$ from (m, D) to (m', D^*) in state N. Note that $(m', D^*) \notin \mathcal{B}_{\sigma}$, by Observation 1. We can choose ε to be small enough such that a non-rational receiver's subjective likelihood ratio of any report in $\mathcal{B}_{\sigma}$ remains above $(1-\pi)/\pi$. Furthermore, since $\rho^*_{\sigma}(m') > (1-\pi)/\pi$, the deviation will satisfy $\rho^*_{\hat{\sigma}}(m') > (1-\pi)/\pi$ if ε is sufficiently small, such that a receiver of any type will still be persuaded by (m', D^*). Therefore, the deviation weakly raises the probability of persuasion. ■

The next Lemma, which is a key step in the proof of the $\lambda > 0$ case (it is irrelevant for the $\lambda = 0$ case), illustrates the complications that arise in this case when the sender wants to shift weight from one report to another. Due to the externalities across reports (i.e., a dictionary in one report may highlight message components that appear in other reports), this shifting of weight must be done in a particular manner to not adversely affect the likelihood ratios associated with the reports.

Lemma 5 *Without loss of generality, we can restrict attention to sender strategies*

σ that satisfy the following property: If $\mathcal{B}_\sigma \neq \emptyset$, then either D^* or $\emptyset$ is not part of any *report in state N.*

Proof. Assume, by contradiction, that $\mathcal{B}_\sigma \neq \emptyset$, and that $\sigma(m, D^* \mid \theta = N) > 0$ and $\sigma(m^0, \emptyset \mid \theta = N) > 0$ for some m, m^0. Let $(\hat{m}, \hat{D}) \in \mathcal{B}_\sigma$. By Lemma 3, there exists a message m^* that justifies $(\hat{m}, \hat{D})$ such that $\rho_\sigma(m^*, D^*) > (1-\pi)/\pi$. Let G denote the number of reports in $\mathcal{B}_\sigma$ that m^* justifies. Now perform a two-step procedure.

Step 1. Suppose there exists m^* as described above such that $\sigma(m^*, D^* \mid \theta = N) > 0$. If not, proceed to Step 2. Let ε be arbitrarily close to zero (positive or negative). Suppose the sender deviates to a strategy $\hat{\sigma}$ that satisfies

$$\begin{aligned}\hat{\sigma}(m^*, D^* \mid \theta = N) &= \sigma(m^*, D^* \mid \theta = N) + \varepsilon \\ \hat{\sigma}(m^0, \emptyset \mid \theta = N) &= \sigma(m^0, \emptyset \mid \theta = N) + (G-1)\varepsilon\end{aligned}$$

and

$$\hat{\sigma}(m, D \mid \theta = N) = \sigma(m, D \mid \theta = N) - \varepsilon$$

for all $(m, D) \in \mathcal{B}_\sigma$ for which $m^*_D = m_D$; otherwise, $\hat{\sigma}$ coincides with σ. By the definition of m^* (see the previous paragraph), $\rho_\sigma(m^*, D^*) > (1-\pi)/\pi$. Hence,

$$\rho_{\hat{\sigma}}(m^*, D^*) = \frac{\sigma(m^*, D^* \mid \theta = Y)}{\sigma(m^*, D^* \mid \theta = N) + \varepsilon} > \frac{1-\pi}{\pi} \tag{5}$$

for sufficiently small ε (positive or negative). Note also that we constructed $\hat{\sigma}$ such that if $\rho_\sigma(m, D^*) > 0$ for some $m \neq m^*$, then $\rho_{\hat{\sigma}}(m, D^*) = \rho_\sigma(m, D^*)$.

Let $(m, D) \in \mathcal{B}_\sigma$. By Lemma 2, $m'_D \neq m_D$ for every message m' that is part of some report in $\mathcal{B}_\sigma$. This, together with the specification of $\hat{\sigma}$, implies that $\rho_{\hat{\sigma}}(m, D) = \rho_\sigma(m, D)$. This is immediate when $m^*_D \neq m_D$. When $m^*_D = m_D$,

$$\rho_{\hat{\sigma}}(m, D) = \frac{\sum_{m' \mid m'_D = m_D} \sigma(m' \mid \theta = Y)}{\sum_{m' \mid m'_D = m_D} \sigma(m' \mid \theta = N) + \varepsilon - \varepsilon} = \rho_\sigma(m, D)$$

It follows that when the sender deviates from σ to $\hat{\sigma}$, the probability of persuasion changes by $(1-\pi)[1-(1-\lambda)G]\varepsilon$. For any $G \neq 1/(1-\lambda)$, we can find an

arbitrarily small ε (which may be positive or negative) such that the deviation is strictly profitable, a contradiction. If $G = 1/(1-\lambda)$, then we can pick $\varepsilon < 0$ such that (5) is satisfied, and raise the absolute value of ε until either $\hat{\sigma}(m^*, D^* \mid \theta = N)$ or $\hat{\sigma}(m^0, \emptyset \mid \theta = N)$ hits zero.[4] Repeat this type of deviation for every $(\hat{m}, \hat{D}) \in \mathcal{B}_\sigma$ and m^* as defined above.

Step 2. By Step 1, we can restrict attention to strategies in which either the message m^* or the null dictionary are not part of any report in N. If the former is true, we are done. If only the latter is true, then consider some $(\hat{m}, \hat{D}) \in \mathcal{B}_\sigma$ and m^* as defined above. Let $m^{**} \neq m^*$ be some message for which $\sigma(m^{**}, D^* \mid \theta = N) > 0$ (if none exists, we are done). By definition, m^{**} does not justify *any* report in $\mathcal{B}_\sigma$. Consider an alternative strategy $\hat{\sigma}$ that satisfies

$$\begin{aligned}
\hat{\sigma}(m^*, D^* \mid \theta = Y) &= \sigma(m^*, D^* \mid \theta = Y)(1-\varepsilon) \\
\hat{\sigma}(m^{**}, D^* \mid \theta = Y) &= \sigma(m^{**}, D^* \mid \theta = Y) + \sigma(m^*, D^* \mid \theta = Y)\varepsilon \\
\hat{\sigma}(m^{**}, D^* \mid \theta = N) &= \sigma(m^{**}, D^* \mid \theta = N) + \frac{\pi}{1-\pi}\sigma(m^*, D^* \mid \theta = Y)\varepsilon \\
\hat{\sigma}(m^0, \emptyset \mid \theta = N) &= \sigma(m^0, \emptyset \mid \theta = N) + (G-1)\frac{\pi}{1-\pi}\sigma(m^*, D^*)\varepsilon
\end{aligned}$$

and

$$\hat{\sigma}(m, D \mid \theta = N) = \sigma(m, D \mid \theta = N) - \frac{\pi}{1-\pi}\sigma(m^*, D^* \mid \theta = Y)\varepsilon$$

for every $(m, D) \in \mathcal{B}_\sigma$ for which $m_D = m_D^*$. Otherwise, $\hat{\sigma}$ coincides with σ.

First, we show that all the reports that contain D^* persuades both receivers under the new strategy. Because m^* justifies $(\hat{m}, \hat{D}) \in \mathcal{B}_\sigma$, we have that $\sigma(m^*, D^* \mid \theta = Y) > 0$. Since we are in the case where $\sigma(m^*, D^* \mid \theta = N) = 0$, we have that $\rho_{\hat{\sigma}}(m^*, D^*) > (1-\pi)/\pi$. By Observation 1, $\rho_\sigma(m^{**}, D^*) \geq (1-\pi)/\pi$. Therefore,

$$\rho_{\hat{\sigma}}(m^{**}, D^*) = \frac{\sigma(m^{**}, D^* \mid \theta = Y) + \sigma(m^*, D^* \mid \theta = Y)\varepsilon}{\sigma(m^{**}, D^* \mid \theta = N) + \frac{\pi}{1-\pi}\sigma(m^*, D^* \mid \theta = Y)\varepsilon} \geq \frac{1-\pi}{\pi}$$

In addition, by the construction of $\hat{\sigma}$, $\rho_{\hat{\sigma}}(m, D^*) = \rho_\sigma(m, D^*)$ for every (m, D^*) with

[4] This case of equality is the reason we need to qualify the statement of the Lemma by saying that it is without loss of generality.

$m \neq m^*, m^{**}$ that is realized with positive probability.

Next, we show that all the reports in $\mathcal{B}_\sigma$ persuade the non-rational receiver under the alternative strategy. Consider a report $(m', D') \in \mathcal{B}_\sigma$. Let m'' be a message that justifies (m', D'). The message m'' is *not* part of any report in state N. To see why, note that there are two options: Either m'' is part of a report that contains D^*, or m'' is part of a report in $\mathcal{B}_\sigma$. By Step 1, (m'', D^*) is not realized in state N if m'' justify (m', D') and the latter option contradicts Lemma 2. Therefore, it follows that

$$\sum_{m|m_{D'}=m'_{D'}} \sigma(m \mid \theta = N) = \sigma(m', D' \mid \theta = N)$$

and the non-rational receiver's likelihood ratio of (m', D') under $\hat{\sigma}$ is therefore $\rho_{\hat{\sigma}}(m', D') = \rho_\sigma(m', D')$ if $m^*_{D'} \neq m'_{D'}$ and

$$\rho_{\hat{\sigma}}(m', D') = \frac{\sum_{m|m_{D'}=m'_{D'}} \sigma(m \mid \theta = Y) - \sigma(m^*, D^* \mid \theta = Y)\varepsilon}{\sigma(m', D' \mid \theta = N) - \frac{\pi}{1-\pi}\sigma(m^*, D^* \mid \theta = Y)\varepsilon} \geq \frac{1-\pi}{\pi}$$

if $m^*_{D'} = m'_{D'}$, where the inequality follows from the fact that $\rho_\sigma(m', D') \geq \frac{1-\pi}{\pi}$ since $(m', D') \in \mathcal{B}_\sigma$.

Therefore, when the sender deviates from σ to $\hat{\sigma}$, the probability of persuasion changes by

$$[1 - (1-\lambda)G]\,\pi\sigma(m^*, D^* \mid \theta = Y)\varepsilon$$

As in Step 1, for any $G \neq 1/(1-\lambda)$, we can find an arbitrarily small ε (which may be positive or negative) such that the deviation is strictly profitable, a contradiction. If $G = 1/(1-\lambda)$, then we can select $\varepsilon < 0$ such that either $\hat{\sigma}(m^{**}, D^* \mid \theta = N)$ or $\hat{\sigma}(m^0, \emptyset \mid \theta = N)$ hits zero. ■

The remainder of proof proceeds in two steps. First, we derive an upper bound on the probability of persuasion (as a function of π and λ). Then, we show that the strategies outlined in Theorems 2-4 implement this bound.

Let σ be a sender strategy. To obtain an upper bound on the probability of persuasion under σ, note that the probability of persuasion in state Y cannot exceed

one.

Let M^* denote the set of messages that are part of some report in state Y. Denote $I = |M^*|$. Let $\mathcal{C} = \{C_1, \cdots, C_L\}$ be a partition of $\mathcal{B}_\sigma$, where each cell C_l is defined by the (distinct) subset of messages $J(l) \subseteq M^*$ that justify every report in the cell. Therefore, $L \leq 2^I - 1$. For the final piece of notation we let $g(l) =\mid C_l \mid$ and $\beta(l) = \sum_{(m,D)\in C_l} \sigma(m, D \mid \theta = N)$.

Consider some $(m, D) \in C_l \subseteq \mathcal{B}_\sigma$ and a message $m' \in J(l)$. Since m' justifies (m, D), $m'_D = m_D$. By Lemma 2, there cannot be a dictionary $D' \subset D^*$ such that $(m', D') \in \mathcal{B}_\sigma$. Hence, by Observation 2, if the message m' is sent in the state N, then it must be sent with D^*. It follows that for any $l = 1, ..., L$, a non-rational receiver's likelihood ratio of a report $(m, D) \in C_l \subseteq \mathcal{B}_\sigma$ is

$$\frac{\sum_{m'\in J(l)} \sigma(m' \mid \theta = Y)}{\sigma(m, D \mid \theta = N) + \sum_{m'\in J(l)} \sigma(m', D^* \mid \theta = N)} = \frac{1-\pi}{\pi},$$

where the equality follows from Lemma 4. This equation can be rewritten as

$$\sigma(m, D \mid \theta = N) = \sum_{m'\in J(l)} [\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y) - \sigma(m', D^* \mid \theta = N)] \tag{6}$$

Note that the R.H.S. remains the same for any $(m, D) \in C_l$. Hence, if we write the above equation for each $(m, D) \in C_l$ and sum over all the reports in C_l we obtain the following equation:

$$\beta(l) = g(l) \cdot \sum_{m'\in J(l)} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y) - \sigma(m', D^* \mid \theta = N)\right] \tag{7}$$

Combining (6) and (7) implies that for every l and every $(m, D) \in C_l$, $\sigma(m, D \mid \theta = N) = \beta(l)/g(l)$, such that the the non-rational receiver's likelihood ratio of all reports in C_l is

$$\frac{\sum_{m'\in J(l)} \sigma(m' \mid \theta = Y)}{\frac{\beta(l)}{g(l)} + \sum_{m\in J(l)} \sigma(m', D^* \mid \theta = N)} = \frac{1-\pi}{\pi}, \tag{8}$$

Solving for $\beta(l)$ in (8) and summing over l give us

$$\begin{aligned}\sum_{l=1}^{L}\beta(l) &= \sum_{l=1}^{L} g(l) \sum_{m'\in J(l)} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y) - \sigma(m', D^* \mid \theta = N)\right] \\ &= \sum_{m'\in M^*} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y) - \sigma(m', D^* \mid \theta = N)\right] \sum_{l\in 1,\ldots,L | m\in J(l)} g(l)\end{aligned}$$

where the second equality follows from changing the order of summation. Now, observe that if $\sigma(m', D^* \mid \theta = N) > 0$, then it must be the case that $\rho^*_\sigma(m') \geq (1-\pi)/\pi$. Therefore,

$$\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y) - \sigma(m', D^* \mid \theta = N) \geq 0$$

By definition, $\sum_{l\in J^{-1}(m')} g(l)$ is the number of reports that are justified by m'. By Corollary 2, this number is at most S. Therefore,

$$\begin{aligned}\sum_{l=1}^{L}\beta(l) &\leq \sum_{m'\in M^*} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y) - \sigma(m', D^* \mid \theta = N)\right] S \qquad (9) \\ &= \left[\frac{\pi}{1-\pi} - \sum_{m'\in M^*} \sigma(m, D^* \mid \theta = N)\right] S\end{aligned}$$

where the final equality follows since $\sum_{m'\in M^*} \sigma(m' \mid \theta = Y) = 1$.

We now employ inequality (9) to derive an upper bound on the probability of persuasion in state N. By Lemma 5, we only need to consider three cases. (When $\lambda = 0$, Lemma 5 is not required, and cases 2 and 3 below coincide.)

Case 1: $\mathcal{B}_\sigma = 0$. In this case, we know from Section 2 that the maximal probability of persuasion is 2π.

Case 2: $\mathcal{B}_\sigma \neq \emptyset$, and no report that is played in state N includes the dictionary D^*.

Therefore, $\sigma(m', D^* \mid \theta = N) = 0$ for every $m' \in M^*$. Plugging that into (9) gives us

$$\sum_{l=1}^{L} \beta(l) \leq \frac{\pi}{1-\pi} S$$

which implies that the overall probability of persuasion is bounded from above by

$$\pi + (1-\pi)(1-\lambda)\frac{\pi}{1-\pi} S = \pi[1 + S(1-\lambda)] \tag{10}$$

Case 3: $\mathcal{B}_\sigma \neq \emptyset$, and no report that is played in state N includes the dictionary $\emptyset$. This means that

$$\sum_{l=1}^{L} \beta(l) = 1 - \sum_{m' \in M^*} \sigma(m', D^* \mid \theta = N)$$

Plugging this into (9) yields

$$1 - \sum_{m' \in M^*} \sigma(m', D^* \mid \theta = N) \leq \left[\frac{\pi}{1-\pi} - \sum_{m' \in M^*} \sigma(m', D^* \mid \theta = N) \right] S$$

or, equivalently,

$$\sum_{m' \in M^*} \sigma(m', D^* \mid \theta = N) \leq \frac{\pi S - (1-\pi)}{(1-\pi)(S-1)}$$

The L.H.S of this inequality is the probability of persuading a rational receiver in state N. The R.H.S is an upper bound on this probability. Note that if $\pi < \pi^*$, we obtain a contradiction because the R.H.S is negative. It follows that Case 3 is only possible when $\pi \geq \pi^*$. As to the non-rational receiver, the overall probability of persuading him is at most one. The resulting upper bound on the overall probability of persuasion in this case is

$$\pi + (1-\pi)(1-\lambda) + (1-\pi)\lambda \frac{\pi S - (1-\pi)}{(1-\pi)(S-1)} = 1 - \frac{\lambda S(1-2\pi)}{S-1} \tag{11}$$

Combining Cases 1-3, it is easy to check that the tight upper bound on the overall probability of persuasion is 2π when $\lambda \geq \lambda^*$; (10) when $\lambda < \lambda^*$ and $\pi \leq \pi^*$; and (11) when $\lambda < \lambda^*$ and $\pi > \pi^*$. Our final step is to verify that the strategy outlined in Theorems 2-4 implements the upper bound. Without loss of generality, denote $m^* = (1, 1, \ldots, 1)$ and $m^0 = (0, 0, \ldots, 0)$. Note that the strategy outlined in Theorems 2-4 satisfies the feature that D is independent of θ conditional on m, and therefore a rational receiver indeed relies purely on m to draw inferences regarding the underlying state. There are three case to consider.

Case 1: Let the sender's strategy σ be the one outlined in Theorem 2. This is the standard case of Kamenica and Gentzkow (2011) described in Section 2, which induces a probability of persuasion 2π.

Case 2: Let the sender's strategy σ be as in Theorem 3. Note that the strategy is only feasible when $S\pi/(1-\pi) \leq 1$. By construction, every $(m, D) \in \mathcal{B}_\sigma$ and every (m^*, D') that is played in state Y satisfies $\rho_\sigma(m, D) = \rho_\sigma(m^*, D') = (1-\pi)/\pi$, such that the report persuades a non-rational receiver. As to a rational receiver, note that the message fully reveals whether $\theta = Y$. Therefore, the rational receiver is only persuaded by the reports (m^*, D'). The probability of persuasion is therefore

$$\pi + (1-\pi)S\frac{\pi}{1-\pi}(1-\lambda) = \pi[1 + S(1-\lambda)]$$

Case 3: Let the sender's strategy σ be as in Theorem 4. Note that this strategy is only feasible when $S\pi/(1-\pi) \geq 1$. By construction, every (m, D) that is played with positive probability satisfies $\rho_\sigma(m, D) = S$, which is by assumption greater than $(1-\pi)/\pi$. Therefore, a non-rational receiver is persuaded with probability one. As to the rational receiver, note that the only message that is played in state Y is m^*. Moreover,

$$\rho^*_\sigma(m^*) = \frac{1}{1 - \frac{(1-2\pi)S}{(1-\pi)(S-1)}}$$

which exceeds $(1-\pi)/\pi$. Therefore, the probability of persuasion is

$$\lambda\left[\pi+(1-\pi)\left(1-\frac{(1-2\pi)S}{(1-\pi)(S-1)}\right)\right]+1-\lambda=1-\frac{\lambda S(1-2\pi)}{S-1}$$

These calculations establish that the strategy outlined in Theorems 2-4 implements the above-derived upper bounds on the probability of persuasion for each of the three parameter ranges.

5 Alternative Notions of Dictionaries

The notion of a dictionary employed in Sections 2-4 postulates that data regarding the sender's strategy takes the form of a conditional distribution of a single subset of message components (where the conditioning is on θ). In this section we discuss alternative forms of data that the sender could choose to provide and explore their implications for the maximal probability of persuasion. We focus on the case of $\lambda = 0$ as in Sections 2-3.

5.1 Splitting Dictionaries

Suppose that the sender wishes to accompany a particular message m with an interpretation of its first two components, m_1 and m_2. So far, we assumed that he can only use a dictionary that discloses the joint distribution over (m_1, m_2) conditional on θ. Alternatively, the sender could interpret the two components *separately*: instead of providing a single dictionary $D = \{1, 2\}$, he would simultaneously provide *two* smaller dictionaries $\{1\}$ and $\{2\}$. When the receiver obtains these two small dictionaries, he learns $(\sigma(m_1 \mid \theta))$ and $(\sigma(m_2 \mid \theta))$, but he has no data about the *joint* distribution of (m_1, m_2) conditional on θ. In line with our approach that the receiver does not draw inferences beyond the data he receives, we assume that he regards the two components as *independent* conditional on θ. In other words, even when m_1 and m_2 act as correlated signals of θ, the receiver neglects this correlation.

More generally, define a *composite dictionary* $\mathcal{D}$ to be a collection of mutually disjoint subsets of $\{1, ..., K\}$. Given a sender strategy σ, the receiver's posterior belief on state Y after being confronted with the report $(m, \mathcal{D})$ is

$$\frac{\pi \prod_{D \in \mathcal{D}} \sigma(m_D \mid \theta = Y)}{\pi \prod_{D \in \mathcal{D}} \sigma(m_D \mid \theta = Y) + (1 - \pi) \prod_{D \in \mathcal{D}} \sigma(m_D \mid \theta = N)}$$

The notion of a composite dictionary and our definition of the receiver's posterior belief are very similar to De Barreda et al. (2018), who study persuasion where the sender has multiple communication channels and the receiver neglects their correlation. The key difference is that in our model, the sender can vary the composite dictionary with the message he submits.

This assumption, as well as the assumption in Sections 2-4, are both special cases of the *maximum entropy principle* (which dates back to Jaynes (1957), and is applied in Spiegler (2018) in a similar context of games with players who extrapolate a belief from partial data). The principle states that given partial data regarding an underlying joint probability distribution, the extrapolated subjective belief maximizes (Shannon) entropy subject to being consistent with the available data. The maximum-entropy principle can be regarded as a single organizing modeling assumption for generalizations of our model - as we will see again in Section 5.3.

Do composite dictionaries empower the sender? Suppose K is even, and recall the strategy outlined in Theorem 1, which attains full persuasion when $\pi \geq 1/(1 + S)$. Suppose the sender deviates from this strategy by splitting each of these dictionaries into its constituent singletons (without changing the conditional distribution over messages). Can this new strategy achieve full persuasion for $\pi < 1/(1 + S)$? There are two conflicting forces. On the one hand, splitting dictionaries into their constituent singletons highlights patterns that are more frequent in state N. For example, $\sigma(m_1 = 1 \mid \theta = N) \geq \sigma(m_1 = m_2 = 1 \mid \theta = N)$. This force lowers the subjective likelihood ratio associated with reports in state N. On the other hand, splitting dictionaries creates a correlation-neglect effect that makes the combination

of these individual patterns appear more informative of state Y than they really are. Recall that in state Y, $m_1 = m_2 = 1$ with probability one, whereas in state N, m_1 and m_2 sometimes take different values. By regarding these correlated message components as independent, the receiver exaggerates the extent to which the realization $m_1 = m_2 = 1$ is indicative of state Y.

Which of the two forces wins? Under the original strategy, each report in state N induces a subjective likelihood ratio of S. Let us calculate the subjective likelihood ratio of each report in state N under the new strategy. Since we kept the conditional distribution over m unchanged, the probability of $m_k = 1$ conditional on $\theta = N$ is $1/2$ for every $k = 1, ..., K$. Moreover, every m that is played in state N has exactly $K/2$ 1's. Thus, the subjective likelihood ratio for each report in state N is

$$\frac{1}{(\frac{1}{2})^{K/2}} = 2^{K/2} < S$$

Therefore, the new strategy does not lead to better subjective likelihood ratios. We conclude that the dictionary-splitting deviation does *not* help the sender, in the sense that it *cannot* achieve full persuasion for a wider range of priors.

This calculation can be extended to any deviation that involves any pattern of composite dictionaries, while retaining the same state-contingent distribution over m as in the optimal strategy of Section 3. However, it remains an open question whether there exists some strategy for the seller that utilizes a *different* conditional message distribution and involves composite dictionaries, which outperforms the strategy of Section 3. We conjecture that the answer is *negative*. If this is the case, the conclusion would be that selective interpretation is a more powerful form of persuasion than leveraging correlation neglect.

5.2 Self-Referential Dictionaries

Although the sender's strategy maps each state to a distribution over *pairs* of elements (message and dictionary), so far we assumed that the sender can only interpret *one* of these elements - namely, the message. In this subsection, we propose a richer

notion of dictionaries that allows the sender to interpret the use of dictionaries as well. This is a self-referential dictionary - it contains a description of the frequency with which different types of dictionaries are employed in the various states.

To illustrate this notion, suppose that $K = 1$, such that the model of Section 2 collapses into the rational-expectations benchmark of Kamenica and Gentzkow (2011) - i.e., the maximal probability of persuasion is 2π with commitment and zero without. The set of available dictionaries is $\mathcal{D} = \{D_{mes}, D_{dic}\}$. The sender's strategy is a function $\sigma : \Theta \to \Delta(\{0,1\} \times \mathcal{D})$. When the receiver obtains a report (m, D_{mes}), he gets access to the conditional distribution $(\sigma(m \mid \theta))_{\theta \in \Theta}$, and therefore draws inferences about θ exclusively on the basis of m. When he obtains a report (m, D_{dic}), he gets access to the conditional distribution $(\sigma(D \mid \theta))_{\theta \in \Theta}$, and therefore draws inferences about θ exclusively on the basis of D.

To see how a self-referential dictionary can help the sender, consider the following strategy. In state Y, the sender plays a single report $(1, D_{dic})$. In state N, he randomizes uniformly between $(1, D_{mes})$ and $(0, D_{dic})$. When the receiver gets the reports $(1, D_{dic})$ or $(0, D_{dic})$, he cannot interpret the message, yet the self-referential dictionary D_{dic} enables him to draw an inference from the fact that he received this particular dictionary. Specifically, the likelihood ratio of D_{dic} is $1/(1/2) = 2$. Similarly, when the receiver gets the report $(1, D_{mes})$, he is only able to draw inferences from the message. The likelihood ratio of $m = 1$ is $1/(1/2) = 2$. It follows that this strategy induces full persuasion whenever $\pi \geq \frac{1}{3}$.

As this example demonstrates, allowing for self-referential dictionaries is akin to adding a dimension to the message profile. Therefore, it increases the sender's ability to persuade the receiver. However, when $K > 1$, it does not enhance the probability of persuasion by the same amount as adding a regular message dimension. Also, utilizing it destroys the property that $D \perp \theta \mid m$.

5.3 Dictionaries as Collections of Marginal and Conditional Distributions

So far, we assumed that dictionaries provide data about the distribution of variables conditional on θ. However, statistical data can involve other combinations of marginal and conditional distributions. We build on the previous sub-section, setting $K = 1$ and allowing for self-referential dictionaries. Let p denote the joint distribution over θ, m, D induced by the prior over θ and the sender's strategy.

Consider the following four *primitive* dictionaries, R_1, R_2, R_3, and R_4, where: R_1 only gives access to the marginal distribution ($p(m)$); R_2 only gives access to the conditional distribution ($p(m \mid \theta)$); R_3 only gives access to the conditional distribution ($p(D \mid \theta)$); and R_4 only gives access to the conditional distribution ($p(D \mid \theta, m)$). The set $\mathcal{D}$ of feasible dictionaries consists of all the subsets of $\{R_1, R_2, R_3, R_4\}$.

As elsewhere in the paper (and as made explicit in Section 5.1), assume that the receiver extrapolates an unconditional belief from the data he receives using the maximum-entropy principle, and he uses this extrapolated belief to draw inferences from the realized report. In particular, we will calculate the unconditional subjective belief induced by four specific dictionaries. The dictionary $\{R_1\}$ induces the belief $p(\theta)p(m)$ over θ, m; the dictionary $\{R_2\}$ induces the belief $p(\theta, m)$ over θ, m (in both these cases, the receiver receives no data about D and therefore ignores this variable); the dictionary $\{R_3\}$ induces the belief $p(\theta, D)$ over θ, D (in this case, the receiver receives no data about m and therefore ignores this variable); finally, the dictionary $\{R_1, R_4\}$ induces the belief $p(\theta)p(m)p(D \mid \theta, m)$ over θ, m, D. In each of these cases, the receiver updates the extrapolated unconditional belief according to the realized report.

It turns out that with this richer notion of dictionaries, the sender is able to outperform the optimal strategy for $K = 2$ described in Section 3. Consider the following strategy. In state Y, the sender randomizes between *two* reports: With probability ε, he sends the report $(1, \{R_1, R_4\})$, and with the remaining probability $1 - \varepsilon$, he sends the report $(0, \{R_3\})$. In state N, the sender mixes between *three* reports. With probability α, he sends the report $(1, \{R_1, R_4\})$; with probability β,

he sends the report $(1, \{R_3\})$; and with the remaining probability $1-\alpha-\beta$, he sends the report $(0, \{R_2\})$.

Claim 5 *For every $\pi > \frac{1}{10}\left(5-\sqrt{5}\right)$, there exist $\alpha, \beta, \varepsilon \in (0,1)$ such that the sender attains full persuasion with the above strategy.*

Proof. Consider the realized report $(1, \{R_1, R_4\})$. Recall that the receiver's unconditional subjective probability of any θ, m, D that is induced by the dictionary $\{R_1, R_4\}$ is $p(\theta)p(m)p(D \mid \theta, m)$. This subjective belief induces the following likelihood ratio of the realized report $(1, \{R_1, R_4\})$:

$$\frac{p(D = \{R_1, R_4\} \mid m = 1, \theta = Y)}{p(D = \{R_1, R_4\} \mid m = 1, \theta = N)} = \frac{1}{\frac{\alpha}{\alpha+\beta}} = 1 + \frac{\beta}{\alpha}$$

Now consider the realized report $(1, \{R_3\})$. The dictionary R_3 impels the receiver to draw inferences from D only. Therefore, the subjective likelihood ratio of this report is

$$\frac{p(D = \{R_3\} \mid \theta = Y)}{p(D = \{R_3\} \mid \theta = N)} = \frac{1-\varepsilon}{\beta}$$

Finally, consider the realized report $(0, \{R_2\})$. The dictionary R_2 impels the receiver to draw inferences from m only. Therefore, the subjective likelihood ratio of this report is

$$\frac{p(m = 0 \mid \theta = Y)}{p(m = 0 \mid \theta = N)} = \frac{1-\varepsilon}{1-\alpha-\beta}$$

In order to attain full persuasion, the three subjective likelihood ratios must all be weakly greater than $(1-\pi)/\pi$. A straightforward calculation establishes that whenever $\pi > \frac{1}{10}\left(5-\sqrt{5}\right)$, we can find $\alpha, \beta, \varepsilon$ that will satisfy these three inequalities. In particular, ε will be arbitrarily small. ■

The lower bound on π that ensures full persuasion in this example is strictly below $\frac{1}{3}$, which was the cutoff in the previous sub-section. This example shows that (despite the impression that our conjecture at the end of Section 5.1 might create), the basic notion of a dictionary given in Section 2 does entail a loss of generality. Also, note that in contrast to the other sender strategies employed in this paper,

here full persuasion requires sending *two* distinct messages in state Y (although one of them is played with arbitrarily small probability).

6 Another "Non-Bayesian Persuasion" Model[5]

An alternative approach to incorporating boundedly rational expectations in a persuasion game is to allow the sender to commit to a *non-partitional information structure* for the receiver. Non-partitional information structures violate the introspection axioms that characterize the standard epistemic model of possibility correspondences that underlies Harsanyi's model of games with incomplete information (see Rubinstein (1998, Ch. 3)). In the present context, they mean that the receiver draws correct statistical inferences from learning that a particular event has occurred, but makes no inference from the fact that other events have not occurred.

This alternative framework captures situations in which a third party - a biased interpreter whose preferences are aligned with the sender's - provides the receiver with *coarse* datasets: When the sender sends a message m, the interpreter does not explain to the receiver how likely *that* particular message is in each state; instead, the interpreter pools m with other messages and provides the receiver with the conditional probability of receiving one of those messages. That is, the interpreter strategically adds ambiguity to the meaning of a message. Because the receiver cannot make inferences from the data he receives, the interpreter can manipulate the receiver's beliefs by pooling two distinct messages m and m' with the *same* set of messages (which induces a non-partitional information structure).

Formally, let M be a finite set consisting of n feasible messages. For every $m \in M$, let $\mathcal{I}(m)$ be a collection of subsets $I \subseteq M$ such that $m \in I$. The sender's feasible action set (independently of the state) is $\mathcal{A} = \{(m, I) \mid m \in M, I \in \mathcal{I}(m)\}$. The meaning of an action (m, I) is that m is the sender's actual message and I is the receiver's information set - i.e. he only learns that $m \in I$. The sender commits to a strategy $\sigma : \Theta \to \Delta(\mathcal{A})$. Let $\sigma(m, I \mid \theta)$ denote the probability that the strategy

[5]The model presented in this section was inspired by insightful comments by Xiaosheng Mu. In particular, he conjectured the optimal strategy that we describe in Proposition 1.

assigns to the action (m, I) in state θ. As before, $\sigma(m \mid \theta) = \sum_I \sigma(m, I \mid \theta)$ is the probability that the message m is played in θ. The receiver uses naive Bayesian updating to form his posterior belief. That is, given the sender's strategy σ, when the action (m, I) is realized, the receiver's posterior belief about the likelihood that $\theta = Y$ is given by:

$$P_\sigma(m, I) = \frac{\pi \sum_{m' \in I} \sigma(m' \mid \theta = Y)}{\pi \sum_{m' \in I} \sigma(m' \mid \theta = Y) + (1 - \pi) \sum_{m' \in I} \sigma(m' \mid \theta = N)}$$

The receiver's subjective likelihood ratio of (m, I) is therefore

$$\rho_\sigma(m, I) = \frac{\sum_{m' \in I} \sigma(m' \mid \theta = Y)}{\sum_{m' \in I} \sigma(m' \mid \theta = N)}$$

The sender's objective is to choose σ that maximizes the probability that the agent chooses Y, subject to the constraint that the receiver's action is optimal given his posterior belief.

Our original model of Section 2 can be equivalently described in terms of this alternative model. Set $M = \{0, 1\}^K$ and define $\mathcal{I}(m)$ as the collection of all sets $I = \{m' \mid m'_D = m_D\}$ for some $D \subseteq \{1, ..., K\}$. In contrast, the more elaborate specifications of our model discussed in Sections 5.1 and 5.3 *cannot* be described in this manner. The reason is that they make more complex use of maximum-entropy extrapolation and cannot be described as naive Bayesian updating with respect to an information set. This means that although the two models are equivalent for our basic notion of dictionaries, they diverge when we consider more complex notions.

To better understand the restriction that our original formulation imposes on $\mathcal{I}(m)$, we now solve for the maximal probability of persuasion when $\mathcal{I}(m)$ is unrestricted - i.e. it consists of *all* subsets $I \subseteq M$ that include m.

Proposition 1 *Assume $\mathcal{I}(m) = \{I \subseteq M \mid m \in I\}$ for every $m \in M$.*

(i) Let $\pi \geq 1/n$. Then, the following strategy attains full persuasion. In state Y, the sender plays $(m_1, \{m_1\})$ with probability one. In state N, he uniformly randomizes over the $n - 1$ actions $(m_2, \{m_1, m_2\}), ..., (m_n, \{m_1, m_n\})$.

(ii) Let $\pi < 1/n$. Then, the maximal probability of persuasion is $\pi(n-1)$, implemented by the following strategy. In state Y, the sender plays $(m_1, \{m_1\})$ with probability one. In state N, he assigns probability $\pi/(1-\pi)$ to each of the the $n-2$ actions $(m_2, \{m_1, m_2\}), ..., (m_{n-1}, \{m_1, m_{n-1}\})$, and probability $1-(n-2)\pi/(1-\pi)$ to the action $(m_n, \{m_1, m_n\})$.

Note that in this model, full persuasion is attained at significantly lower priors than in the original model of Section 2, where $n = 2^K$ and yet full persuasion is only attained for

$$\pi \geq \frac{1}{1+\binom{K}{\lfloor K/2 \rfloor}} > \frac{1}{2^K}$$

However, the optimal strategy given by Proposition 1 lacks a natural interpretation, whereas the optimal strategy in our original model can be described in terms of selective interpretation of "good news".

6.1 Proof of Proposition 1

We first argue that the strategy σ outlined in the Proposition persuades the receiver with probability one when $\pi \geq 1/n$ and with probability $(n-1)\pi$ when $\pi < 1/n$. To see this, note that when $\pi \geq 1/n$, then $P_\sigma(m_1, \{m_1\}) = 1$;

$$P_\sigma(m_i, \{m_1, m_i\}) = \frac{\pi(n-1)}{\pi(n-2)+1} \geq 1/2$$

for $i = 2, \ldots, n$. When $\pi < 1/n$, then $P_\sigma(m_1, \{m_1\}) = 1$; $P_\sigma(m_i, \{m_1, m_i\}) = 1/2$ for $i = 2, \ldots, n-1$ and

$$P_\sigma(m_n, \{m_1, m_n\}) = \frac{\pi}{1-(n-2)\pi} < 1/2.$$

We now proceed to show that no other strategy achieves a higher probability of persuasion. Let σ be an optimal sender strategy.

Observation 3 *Without loss of generality, we can restrict attention to strategies that accompany each message m with a unique information set $I(m)$.*

Proof. Suppose that the pairs $(m, I), (m, I')$ are both played with positive probability under σ, such that $I \neq I'$ and $\rho_\sigma(m, I) \geq \rho_\sigma(m, I')$. Let $\hat{\sigma}$ be a strategy that differs from σ only by replacing every occurrence of (m, I') with (m, I). Since the deviation does not change the distribution of messages conditional on each state, it leaves $\rho_\sigma(m, I)$ and $\rho_\sigma(m, I')$ unchanged, and it does not affect the likelihood ratio of any other report. Therefore, the deviation weakly raises the probability of persuasion. ■

Henceforth, we restrict attention to strategies in which each $m \in M$ that is sent with positive probability is paired with a unique information structure $I(m)$. Define J as the set of messages m for which

$$\frac{\sigma(m \mid \theta = Y)}{\sigma(m \mid \theta = N)} > \frac{1-\pi}{\pi}$$

By the definition, the receiver would be persuaded by the pair $(m, \{m\})$ for every $m \in J$. Since the sender can always select $I(m) = \{m\}$, it follows that $\rho_\sigma(m, I(m)) \geq (1-\pi)/\pi$ for every $m \in J$.

Observation 4 *Without loss of generality, we can set $I(m) = J \cup \{m\}$ for every message m that is sent with positive probability.*

Proof. For any action $(m, I(m))$ that is played with positive probability and does *not* persuade the receiver, it must be the case that $\rho_\sigma(m, I) < (1-\pi)/\pi$ for *all* $I \in \mathcal{I}(m)$. Therefore, we can select $I(m) = J \cup \{m\}$ without loss of generality in this case. It remains to show that $\rho_\sigma(m, J \cup \{m\}) \geq (1-\pi)/\pi$ for every message m for which $\rho_\sigma(m, I(m)) \geq (1-\pi)/\pi$.

Let $(m, I(m))$ be an action that is played with positive probability and persuades the receiver. Then,

$$\pi \sum_{m' \in I(m)} \sigma(m', I(m') \mid \theta = Y) \geq (1-\pi) \sum_{m' \in I(m)} \sigma(m', I(m') \mid \theta = N) \qquad (12)$$

Suppose, in contradiction to the claim, that $I(m) \neq J \cup \{m\}$. In particular, suppose there is a message $\tilde{m} \in J - I(m)$. By the definition of J,

$$\pi\sigma(\tilde{m}, I(\tilde{m}) \mid \theta = Y) > (1-\pi)\sigma(\tilde{m}, I(\tilde{m}) \mid \theta = N)$$

Adding these two inequalities, we get

$$\frac{\sum_{m' \in I(m) \cup \{\tilde{m}\}} \sigma(m', I(m') \mid \theta = Y)}{\sum_{m' \in I(m) \cup \{\tilde{m}\}} \sigma(m', I(m') \mid \theta = N)} > \frac{1-\pi}{\pi}$$

Therefore, we can add $\tilde{m}$ to $I(m)$ and the action $(m, I(m) \cup \{\tilde{m}\})$ will still persuade the receiver.

Now suppose there is a message $\hat{m} \in I(m) - J$. By the definition of J,

$$\pi\sigma(\hat{m}, I(\hat{m}) \mid \theta = Y) \leq (1-\pi)\sigma(\hat{m}, I(\hat{m}) \mid \theta = N)$$

Subtracting this inequality from (12) and rearranging, we get

$$\frac{\sum_{m' \in I(m) - \{\hat{m}\}} \sigma(m', I(m') \mid \theta = Y)}{\sum_{m' \in I(m) - \{\hat{m}\}} \sigma(m', I(m') \mid \theta = N)} \geq \frac{1-\pi}{\pi}$$

Therefore, the receiver would also be persuaded by the action $(m, I(m) - \{\hat{m}\})$.

We can repeat this process of adding or eliminating elements, until $(m, I(m))$ is replaced with $(m, J \cup \{m\})$ and the probability of persuasion is unchanged. ■

Observation 5 *If σ is an optimal strategy, then J is non-empty.*

Proof. If $J = \emptyset$, then by Observation 4 it is without loss to assume that each message m is sent with $\{m\}$. Since such a strategy endows the receiver with rational expectations, it implies that the maximal probability of persuasion is 2π. But then σ cannot be optimal, since we have already identified a strategy that achieves a higher probability of persuasion. ■

We are now ready to derive the upper bound on persuasion. Define H as the set of messages $m \notin J$ that are played with positive probability such that $(m, I\{m\})$

persuades the receiver. Assume that $I\{m\} = J \cup \{m\}$ persuades the receiver. By Observation 4 this is without loss. Then, for every $m \in H$, we have

$$\begin{aligned} \rho_\sigma(m, J \cup \{m\}) &= \frac{\sum_{m' \in J \cup \{m\}} \sigma(m', J \cup \{m'\} \mid \theta = Y)}{\sum_{m' \in J \cup \{m\}} \sigma(m', J \cup \{m\} \mid \theta = N)} \\ &= \frac{\sum_{m' \in J \cup \{m\}} \sigma(m', J \cup \{m'\} \mid \theta = Y)}{\sum_{m' \in J} \sigma(m', J \mid \theta = N) + \sigma(m, J \cup \{m\} \mid \theta = N)} \geq \frac{1-\pi}{\pi} \end{aligned} \tag{13}$$

Rearranging the inequality, we obtain an upper bound on the probability that the action $(m, J \cup \{m\})$ is played in state N:

$$\begin{aligned} \sigma(m, J \cup \{m\} \mid \theta = N) &\leq \frac{\pi}{1-\pi} \sum_{m' \in J \cup \{m\}} \sigma(m', J \cup \{m'\} \mid \theta = Y) - \sum_{m' \in J} \sigma(m', J \mid \theta = N) \\ &\leq \frac{\pi}{1-\pi} - \sum_{m' \in J} \sigma(m', J \mid \theta = N). \end{aligned} \tag{14}$$

Note, that (14) is the same for every $m \in H$. Therefore, the probability of persuasion in state N is bounded above as follows

$$\begin{aligned} &\sum_{m \in H \cup J} \sigma(m, J \cup \{m\} \mid \theta = N) \\ &\qquad \leq |H| \left[\frac{\pi}{1-\pi} - \sum_{m' \mid m' \in J} \sigma(m', J \mid \theta = N) \right] + \sum_{m' \in J} \sigma(m', J \mid \theta = N) \\ &\qquad \leq |H| \frac{\pi}{1-\pi} - (|H| - 1) \sum_{m' \mid m' \in J} \sigma(m', J \mid \theta = N). \end{aligned} \tag{15}$$

By definition of J we have $\sigma(m', J \mid \theta = N) < \pi/(1-\pi)$. This implies that the upper bound on persuasion is increasing in $|H|$ and thus decreasing in $\sigma(m', J \mid \theta = N)$ for every $m' \in J$ when $n > 2$. By Observation 5, $|H| \leq n-1$ since $J \neq \emptyset$. Furthermore, if there exists an action that does not persuade the receiver, then $|H| \leq n-2$. This will always be the case when $(n-1)\pi/(1-\pi) < 1$.

It follows that the upper bound on the overall probability is

$$\begin{array}{ll} \pi + (1-\pi)(n-1)\frac{\pi}{1-\pi} & when \quad \pi \geq \frac{1}{n} \\ \pi + (1-\pi)(n-2)\frac{\pi}{1-\pi} & when \quad \pi < \frac{1}{n} \end{array}$$

This completes the proof.

7 Conclusion

Conventional models of strategic communication focus on the role of selective transmission of information. And yet, real-life communication also involves strategic *interpretation* of information. This paper formalized this aspect as selective provision of *statistical data* regarding the long-run mapping between messages and the underlying state. In a pure persuasion model, we showed that strategic interpretation significantly enhances the sender's ability to persuade the receiver - to the point that *full* persuasion is sometimes possible, in sharp contrast to the standard rational-expectations benchmark. From a broader perspective, the modeling innovation in this paper is the idea that one player can influence another player's understanding of equilibrium regularities, by affecting the statistical data regarding the equilibrium distribution that the latter player has at his disposal. To use the terminology of Spiegler (2018), one player can influence another player's "archival information" - just as in a standard model, one player's information set is determined by the prior moves of other players. Exploring this idea outside the context of strategic communication is an interesting problem for future research.

Related literature

Our paper joins a recent small literature on persuasion that departs from the standard paradigm of rational expectations under a common prior. De Barreda et al. (2018) study a sender-receiver model in which the receiver exhibits "correlation neglect". Specifically, the sender submits multiple simultaneous signals and the receiver erroneously treats them as being conditionally independent. In Section 5, we show how to map this setting into the language of our model, such that the correlation-

neglect effect arises from the use of a particular notion of dictionaries. Schwartzstein and Sunderam (2018) examine a game of persuasion in which both parties observe the realization of a signal that is drawn from a distribution that depends on an unobserved state of Nature. The receiver's non-rational expectations are captured by the assumption that the sender knows the signal distribution, while the receiver believes in whatever signal distribution the sender reports. Galperti (2018) analyses a model of persuasion with non-common priors, where the sender can influence the prior beliefs held by the receiver. In particular, when the receiver observes a message realization that has zero probability according to his prior belief, he abandons his original prior in favor of a new one. We, on the other hand, maintain the common prior assumption but allow the sender to strategically determine the receiver's understanding of the equilibrium distribution.

The basic specification of dictionaries that we employ in Sections 2-4 forms a close link to Jehiel's (2005) notion of analogy-based expectations equilibrium (ABEE). According to this concept, players form coarse beliefs that are measurable with respect to an "analogy partition" of the possible states of the world. Our basic notion of a dictionary as a subset of components of multi-dimensional messages implies that a dictionary is essentially an analogy partition. Therefore, this version of the model can be viewed as an extensive game in which the sender chooses the message as well as the receiver's analogy partition, and the solution concept is ABEE. This tight link to ABEE breaks down when we consider more elaborate notions of dictionaries in Section 5. Relatedly, Jehiel (2011) analyzes an auction setting where the designer's choice of bidders' learning feedback regarding the distribution of past bids shapes their analogy partitions.

Jehiel and Koessler (2008) modify the Crawford-Sobel model by assuming that the receiver bundles states into analogy classes according to an interval analogy partition. They then show that for some given analogy partitions, there may exist an ABEE with partial information transmission even when under rational expectations the babbling equilibrium is the unique equilibrium. In contrast to our model, the analogy partitions are set exogenously and cannot be influenced by the sender. In a similar vein, Mullainathan et al. (2008) study a cheaptalk game where the receiver

uses coarse analogy partitions.

Spiegler (2018) introduces a general framework for static games, in which the description of players' types includes "archival information", defined as partial access to data about correlations among the variables that constitute the state of the world. Dictionaries in our model are a form of archival information. Indeed, our model is an example of how to extend the formalism of Spiegler (2018) to sequential games. Our approach to modeling the receiver's partial understanding of the equilibrium is also related to Glazer and Rubinstein's (in press) model of a "problem solver". In their model, a problem solver has partial understanding of the equilibrium: he observes some summary statistic of the other players' strategies, and then best-replies to a uniform belief over all the strategy profiles that are consistent with this statistic.

Finally, Glazer and Rubinstein (2012, 2014) study persuasion when the *sender* is boundedly rational in the sense of having limited ability to misrepresent the state. They show that a rational receiver can construct intricate disclosure mechanisms that take advantage of this element of the sender's bounded rationality Blume and Board (2013) study cheap talk when the receiver has uncertain ability to distinguish between distinct messages. In contrast to our framework, receivers in Blume and Board (2013) have rational expectations and the sender is unable to influence their interpretative abilities.

References

Ian Anderson. *Combinatorics of Finite Sets*. Oxford University Press, 1987.

Andreas Blume and Oliver Board. Language barriers. *Econometrica*, 81(2):781–812, 2013.

Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.

Ines De Barreda, Gilat Levy, and Ronny Razin. Persuasion with correlation neglect. 2018.

Simon Galperti. Persuasion: The art of changing worldviews. *American Economic Review*, 2018. In Press.

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

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics*, 1:395–410, 2006.

Jacob Glazer and Ariel Rubinstein. A model of persuasion with boundedly rational agents. *Journal of Political Economy*, 120(6):1057–1082, 2012.

Jacob Glazer and Ariel Rubinstein. Complex questionnaires. *Econometrica*, 82(4): 1529–1541, 2014.

Jacob Glazer and Ariel Rubinstein. Coordinating with a "problem solver". *Management Science*, Forthcoming.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106 (4):620, 1957.

Philippe Jehiel. Analogy-based expectation equilibrium. *Journal of Economic theory*, 123(2):81–104, 2005.

Philippe Jehiel. Manipulative auction design. *Theoretical economics*, 6(2):185–217, 2011.

Philippe Jehiel and Frédéric Koessler. Revisiting games of incomplete information with analogy-based expectations. *Games and Economic Behavior*, 62(2):533–557, 2008.

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

Patrick Kennedy and Andrea Prat. Where do people get their news? Columbia Business School Research Paper No. 17-65, 2017.

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

Andrea Prat. Media power. *Journal of Political Economy*, 126(4):1747–1783, 2018.

Ariel Rubinstein. *Modeling bounded rationality*. MIT press, 1998.

Joshua Schwartzstein and Adi Sunderam. Using models to persuade. December 2018.

Ran Spiegler. News and archival information in games. August 2018.