Persuasion with Rational Inattention*

Preliminary and Incomplete

Alexander W. Bloedel† Ilya Segal‡

First draft: 16th April, 2018
This version: 1st June 2018

Abstract

We study a model of Bayesian persuasion in which Receiver has limited information-processing capacity, or *attention*, and must exert costly effort to process Sender's signals. Receiver is rationally inattentive (Sims (2003)): attention costs are proportional to the mutual information (expected entropy reduction) between Sender's signals and Receiver's "perceptions" of them. Information disclosure plays a dual role: in addition to the usual *persuasion* motive, Sender engages in *strategic attention manipulation*. When Receiver has a binary action choice, we characterize the optimal persuasion strategy using a first-order approach. At the optimum, "complex" signals are used to exploit Receiver's limited attention when interests are misaligned, and "simple and convincing" signals are used to focus attention when interests are aligned. When the persuasion motive is absent (preferences are aligned) we trace the attention manipulation motive to the *multi-dimensionality of information* and the *extensive margin of attention allocation*: if the state space is binary or Receiver faces a pure capacity constraint, full disclosure is uniquely optimal. Applications include advertising, information management in organizations, design of disclosure regulations, and dual-process theories of attention and choice. We also discuss formal connections to models of (i) persuasion with a privately informed Receiver and (ii) contracting with flexible information acquisition, as well as extensions to cheap talk communication.

Keywords: Bayesian persuasion, rational inattention, costly communication, information acquisition, information design, monotone partition, cheap talk

JEL Classification: D82, D83, D91

(∗) We thank seminar participants at Stanford Theory Lunch and Paul Milgrom's research group for useful comments.
(†) Stanford University, Department of Economics. <abloedel@stanford.edu>
(‡) Stanford University, Department of Economics. <isegal@stanford.edu>

1. Introduction

Communication — the transfer of information from one party to another — is among the most fundamental of economic transactions, and limited *attention* — the ability and willingness to process that information — is among the primary transaction costs that determines its form. Moreover, as Simon (1971, p. 41) elegantly wrote (emphasis added):

> "In an information-rich world, *most of the cost of information is the cost incurred by the recipient*. It is not enough to know how much it costs to produce and transmit information; *we must also know how much it costs, in terms of scarce attention, to receive it*."

What is the optimal way for a Sender to structure communications with a Receiver who has limited attention and who, in the service of economizing on transaction costs, allocates it optimally? How does this structure depend on the degree of preference alignment between these parties? On the degree of Sender's commitment power? These questions are of central importance in many settings of interest:

- A prosecutor, who aims to convict a defendant in a complicated white-collar case, must argue her case to a jury that, in addition to not being expert in the relevant financial law, must absorb information about a vast amount of (likely boring) evidence over the course of long days in the courtroom. Which pieces of evidence should the prosecutor emphasize? For which aspects of the case should she provide full detail, and which should she simplify into more easily-digestible narratives?

- Every day, an executive — e.g., the president, the CEO of a large company — faces a constant stream of complex decisions, often regarding diverse topics and under time constraints. Paying attention to any given task takes time and energy, and so is costly. An advisor, who has superior information about her area of expertise, needs to advise the executive on the optimal course of action. Even if their preferences over actions are perfectly aligned, is (complete) honesty the best policy? Or can omission of details be a virtue? How should the advisor structure her presentation to prevent the executive's attention from drifting off to other matters during a meeting?

- An online retailer — e.g., Amazon — collects data about a consumer through his purchase history, and has superior information about the likely match quality between this consumer and a sea of products he is unaware of. While online, the consumer is inundated with ads, emails, and social media, and therefore has limited time and energy to search for good product matches. Due to reputation effects, the retailer does not price discriminate (so prices are fixed) and will not lie about product quality. If the retailer wants to maximize profits, what information about products should it provide? Which ones should it explicitly recommend — e.g., send emails about, place prominently on the consumer's homepage? For those that it does not explicitly recommend, how much information should it provide through, e.g., easily-accessible customer reviews? Can it ever benefit from the fact that the consumer is inattentive?

We study these questions in a model of Bayesian persuasion in which Receiver has limited

information-processing capacity — or, more informally, *attention* — and must exert costly effort to process Sender's signals. Following the literature on rational inattention (RI) initiated by Sims (2003), Receiver's cost of attention is proportional to the mutual information — i.e., the expected reduction of Shannon entropy — between Sender's signals and his "perceptions" of these signals, which will generally be "inaccurate" and "noisy." Despite his limited ability to process information, Receiver is sophisticated and allocates his attention optimally given the disclosure policy announced by Sender.

The main innovations of our model are twofold. First, relative to the literature on persuasion, information disclosure in our model plays a dual role. Taking Receiver's attention strategy as given, Sender, as usual, uses information as a tool for *persuasion*, i.e., to convince Receiver to take Sender-preferred actions. But, crucially, Sender does not directly control which pieces of the disclosed information Receiver attends to, and must incentivize him to direct his attention appropriately. Indeed, Receiver's attention strategy — and, hence, his entire "best response function" — is *endogenous*, and it depends on the *entire signal structure* that Sender proposes. Our Sender thus also uses information disclosure in a new way, namely, as a tool for *strategic attention manipulation*. The RI model allows Receiver's attention strategy to be completely flexible: he chooses not only whether and how much to pay attention but, importantly, what aspects of Sender's information to pay attention to. By choosing the shape of her signal structure, Sender is able to manipulate Receiver's attention allocation along each of these dimensions. Second, relative to the RI literature, our Receiver pays attention to an *endogenous* and *strategically chosen* variable. Strategic applications of the RI model are still in their infancy, and ours is among the first to apply it in a setting of strategic communication.

More specifically, we study a setting in which Receiver has a binary action choice — whether or not to act — and both Sender and Receiver have preferences that are affine in the state. For the majority of the paper, we assume there are a continuum of states, though we also provide a full analysis when the state space is binary. Receiver's RI problem is inherently infinite-dimensional, but it can be reduced to a single first-order condition for a scalar *activity parameter*. This allows us to solve Sender's persuasion problem using a first-order approach, which appears to be new to the RI literature.[1] Signals are identified with the posterior means they induce, so that signal structures can be identified with mean-preserving contractions of the prior. For a fixed activity parameter, Sender's problem is an infinite-dimensional linear program with a single linear constraint, and can be solved via LP duality as in Dworczak and Martini (2018).

Given the reduction to the first-order approach, we present three sets of results. First, we provide general characterizations that apply to all preferences in the affine class we consider. Theorem 1 shows that, when the prior is continuous, the optimal persuasion strategy always induces a monotone partition of the state space of a particularly simple form. Theorem 2 fully characterizes the set of implementable attention strategies. To elucidate the role and form of optimal attention manipulation, the latter two sets of results pertain to full solutions and comparative statics in two special cases of the model.

In the first special case, studied in Section 5.2, Sender has state-independent preferences and

(1) Closest in this respect is Yang (2017), but he directly jointly optimizes over contractual and attention variables. We, on the other hand, first optimize "contractual" variables to implement a fixed set of attention variables and, in a second stage, optimize over attention variables. Thus, our approach is closer in spirit to the two-step procedure familiar from moral hazard models.

merely wants to maximize the probability that Receiver acts. At the optimum, she pools together high states into one signal, and fully discloses the state when it is low. Conditional on the high signal, unlike the standard persuasion model, Receiver *strictly* prefers to act. The high signal is thus *simple* (it pools many states together) and *convincing* (Receiver has a strict preference to act), and is used to *attract and focus Receiver's attention*. The low signals — which provide unfiltered information about the state, and are thus *complex* — are used to *exploit Receiver's inattention*. Receiver's inattention causes him to make mistakes; formally, his action is stochastic conditional on the realized signal. When the state is low, complex signals maximize the likelihood that Receiver makes a mistake by acting against his own interest but in favor of Sender's. As Receiver's cost of attention increases, the optimal signal structure becomes more informative, Sender's utility decreases, and Receiver's utility is non-monotone, strictly decreasing when costs are near zero or very large and increasing when they are intermediate.

In the second special case, studied in Section 5.3, Sender and Receiver's preferences over actions are perfectly aligned, but Sender does not internalize Receiver's attention cost. We provide a complete solution when the prior is symmetric. Perhaps surprisingly, neither full disclosure nor direct action recommendations are generally optimal, though the optimal persuasion strategy combines useful elements of both. In particular, sufficiently high states are pooled together, sufficiently low states are pooled together, and intermediate states are fully revealed. Thus, simple and convincing signals play the role of "strong action recommendations" when the stakes are high, and information processing is fully delegated to Receiver when the stakes are low. This is optimal because the RI cost function gives Receiver a strong incentive to *smooth* his attention across the state space and, by pre-filtering information, Sender is able to relax this attention-smoothing motive. In other words, Sender *focuses* Receiver's attention on more important parts of the state space, inducing him to distinguish more sharply between extremal and intermediate states than he would under full disclosure. As Receiver's cost of attention increases, the optimal signal structure becomes more informative and both parties' utility decreases.

Even under aligned preferences, Sender would, in a sense, like to *exaggerate* the state to catch Receiver's attention. This exaggeration incentive is never counterproductive in the Bayesian persuasion paradigm, which endows Sender with full commitment power. The commitment assumption can be justified in many contexts: the prosecutor is required to present hard evidence, and both the advisor and Amazon are plausibly bound by reputational concerns given that they interact repeatedly with their respective audiences. But it is equally important to understand how limited attention shapes communication when Sender lacks commitment, e.g., when Sender is the advertiser of a new, unfamiliar product. Thus, we extend our model to cheap talk communication in Section 6 and show that the incentive to exaggerate is always hindrance to communication: equilibria are weakly less informative, and both parties' utilities strictly lower, than under full attention. Interestingly, the monotone partitions induced by informative equilibria, when they exist, are identical to those under full attention.

The rest of the paper proceeds as follows. After discussing related literature, in Section 2 we walk through a simple binary-state example to highlight the main intuitions behind our model. Section 3 lays out the full model, and the first-order approach to Sender's problem is developed in Section 4. Our main results on optimal persuasion strategies are developed in Section 5, where we also discuss several economic applications. The extension to cheap talk communication is developed in Section 6. Section 7

contains a discussion of the model and our solution methods, including possible generalizations and more detailed connections to the literature. Section **??** concludes.

1.1. Related Literature

Bayesian persuasion and information design: We contribute to the literature on Bayesian persuasion initiated by Kamenica and Gentzkow (2011) and Rayo and Segal (2010). Closest is the independent and contemporaneous work of Lipnowski, Mathevet and Wei (2018), who study a model of delegated information processing; we defer a detailed comparison in Section 7.4. In short, they restrict attention to aligned preferences and, in that special case of our model, their problem is formally a relaxation of ours. At a formal level, our model has a close connection to the recent work of Kolotilin et al. (2017) and Kolotilin (2017), who study persuasion when the Receiver has private, payoff-relevant information but is fully attentive.[2] As discussed in Section 4.4, differences in interpretation aside, our model can be viewed as an generalization in which the distribution of private information is endogenous to the mechanism. The moral hazard aspect of Receiver's attention allocation problem also connects our paper to Boleslavsky and Kim (2018).[3] In independent work, they develop of model of persuasion subject to a moral hazard constraint, but there are numerous differences: for example, in their model effort is (i) exerted, and privately observed, by an Agent who is a distinct player from Receiver, (ii) one-dimensional by assumption, and (iii) leads to first-order shifts in the state distribution. As a result, the analyses and underlying economics are very different. Finally, from a technical perspective, we follow Dworczak and Martini (2018) in using LP duality to characterize optimal persuasion strategies.

Several other papers are thematically related to ours. Gentzkow and Kamenica (2014) study persuasion when communication is costly for *Sender*. Our model nests Receiver's costly *self*-persuasion problem[4] inside of Sender's (costless) persuasion problem, significantly enriching the analysis. Matysková (2018), in independent and contemporaneous work, studies a model in which Receiver observes Sender's signals for free but, conditional on the realized signal, can subsequently acquire *additional* costly information about the state subject to an RI cost. Her Sender aims to *prevent* information acquisition, whereas our Sender *encourages* Receiver to pay attention. In the engineering literature, Akyol, Langbort and Basar (2017) take an information-theoretic perspective on persuasion with noisy transmission, but there is no moral hazard and they assume a one-dimensional Gaussian-Quadratic structure, which trivializes the qualitative properties of the optimal signal — namely, Sender always transmits the true state plus an additive Gaussian noise term.

None of these papers have the feature that Receiver's best response function is endogenous to the entire signal structure, which is an essential aspect of our model. In that respect, our paper relates to to the more general *information design* literature.[5] For example, in Roesler and Szentes (2017) a buyer

(2) Guo and Shmaya (2017) also study persuasion with a privately informed Receiver, but their model and its solution, which involves non-monotone partitions, have significantly different structures from ours.

(3) Li and Yang (2017) and Georgiadis and Szentes (2018) study related models of costly monitoring in moral hazard problems that touch on RI and information design.

(4) The RI problem is a special case of costly persuasion in which material preferences of Sender and Receiver are perfectly aligned. See, e.g., Caplin, Dean and Leahy (2018) for the belief-based approach to RI.

(5) See Bergemann and Morris (2017) for a comprehensive survey.

chooses how much to learn (though at no cost) about his private value for a good, and his learning strategy determines the price charged by a monopolistic seller through the induced demand curve. In both the first-order relaxation of our problem and their full problem, Sender (the buyer, in their model) optimizes over CDFs of posterior means subject to (i) a mean-preserving spread constraint and (ii) an incentive constraint for a one-dimensional activity parameter (the monopoly price, in their model).

Rational inattention: We build directly on the RI model introduced by Sims (1998) and Sims (2003) and, in particular, the "Logit" characterization for general discrete-choice problems in Matějka and McKay (2015).[6] Substantively, our paper is most closely related to recent extensions of the RI model to strategic settings, in which agents pay attention to endogenous, strategically-chosen variables. Closest is Matějka (2015), who studies monopolistic pricing when the seller has commitment power and the buyer is inattentive to the offered price. Beyond obvious differences in formulation (e.g., prices vs. signals, capacity constraint vs. linear attention cost), the main tradeoffs in his model are driven by the buyer's risk aversion to the (stochastic) price, whereas our Receiver is risk neutral by construction of the signal space. This leads to starkly different comparative statics: in Matějka (2015) prices become *less variable* (more "rigid") as attention costs increase,[7] while in our model signals become more informative and thus *more variable*. Ravid (2018) studies a different model of pricing in which the seller lacks commitment power and, in equilibrium, takes the buyer's attention strategy as given. This bears some resemblance to our cheap talk extension in Section 6 where attention is determined as part of a (perfect Bayesian) Nash equilibrium, but is very different from our main model where Sender is a Stackelberg leader and actively manipulates Receiver's attention.[8]

Also related is the literature on information acquisition in contracts/mechanisms. While information acquisition is typically modeled as either a binary or one-dimensional effort choice in this literature, the recent work of Yang (2017) allows for flexible information choice.[9] He studies optimal security design in a model where the buyer can flexibly learn about a random cashflow (the state) subject to a generalized RI cost function. The buyer directly learns about the exogenous state while being perfectly attentive to the (deterministic) offered security, which is different from our model (and those above) where Receiver is inattentive to Sender's endogenous, stochastic signals. Despite these differences, as discussed in Section 7.3, there turns out to be a close formal connection to our model.

Costly and noisy communication: Costly and noisy communication (without the full flexibility inherent in the persuasion and RI paradigms) is important to the team-theoretic perspective on organizational

(6) The Logit solution of Matějka and McKay (2015) has long been known in the branch of information theory concerned with the rate-distortion problem; to our knowledge, their necessary and sufficient conditions first appeared as Theorem 9.4.1 of Gallager (1968). See Chapter 9.5 of Gray (2011) for an intellectual history of the Logit solution.

(7) More precisely, as Receiver's attention capacity decreases.

(8) Though, as discussed in Sections 5.2 and 5.3, important special cases of our model endogenously result in solutions that correspond to Nash equilibria. See also Matějka and McKay (2012) and Martin (2017) for other models of pricing with an RI buyer in which which the price-setters take attention as given.

(9) See also the follow-up work of Yang and Zeng (2017), and Bergemann and Välimäki (2007) for a survey of the early literature.

economics, which abstracts from incentives.[10] A pertinent recent example is Dessein, Galeotti and Santos (2016), whose benchmark model is equivalent to one with a Gaussian-Quadratic structure and an RI constraint, but importantly without the moral hazard problem that is central to our model. Theirs is a model of *who* in an organization should speak, while ours is a model of *how* to speak once the communication structure has been determined.

Thus far, there has been little work studying the interaction between "technological frictions" (as in team theory) and "incentive frictions" (as in cheap talk, disclosure, and persuasion) in communication, despite numerous calls for the development of such models (Dewatripont (2006), Sobel (2010), Sobel (2012), and Kamenica (2017)). In this respect, the closest precedent to our work is Dewatripont and Tirole (2005), whose model is related to the special case of ours in which Sender's preferences are state-independent.[11] However, they focus on a highly reduced-form version of the problem: the state is binary, Sender is required to send fully-informative signals, Receiver's attention choice is one-dimensional, and signals are either perfectly understood or completely ignored. Our contribution is to relax all of these assumptions and study the rich implications for the form of optimal communication.

In Section 6 we extend our model to cheap talk communication, and show that the *endogenous* noise generated by an inattentive Receiver always hurts both parties when actions are binary. Myerson (1991, pp. 285-288), Board, Blume and Kawamura (2007), and Hernández and Stengel (2014) show that cheap talk with *exogenous* noise can either improve or hinder communication, depending on assumptions. Steiner and Stewart (2016) and Gossner and Steiner (2017) study a model of noisy cheap talk with the interpretation that Sender and Receiver are different "selves" of the same agent. A special case of our model admits a similar interpretation, which we develop through an application to dual-process theories of attention and choice in Section 5.3. Finally, there is a literature building on Glazer and Rubinstein (2004) that studies cheap talk when Receiver has the ability to *partially* verify the validity of Sender's messages. Partial verification can be interpreted as a kind of limited attention, albeit one that is very different from information-theoretic constraint considered here.

2. A Binary-State Example

To develop intuition before delving into the full model, we consider the simplest possible example that shows how limited attention can lead to departures from the standard model. A prosecutor (Sender) wants to convince a jury (Receiver) to convict the defendant in a trial.[12] The defendant is either *g*uilty

(10) See Dessein and Prat (2016) and Garicano and Prat (2011) for surveys of recent developments and Marschak and Radner (1972) for the original treatment. Cremer, Garicano and Prat (2007), Sobel (2015), and Dilmé (2017) study optimal languages in organizational settings, also using the team theoretic approach. Their Senders are constrained to coarse messages spaces, but their Receivers are fully attentive.

(11) Closest is the variant of their model in which the Sender is a Stackelberg leader, communication is costly only for Receiver, and decision-making is of a "supervisory" nature. Calvó-Armengol, Martí and Prat (2015) extend the Dewatripont and Tirole (2005) model to bilateral communication over a network, while Persson (2018) studies a different extension in which Receiver is subject to "information overload" when Sender "complexifies" her messages. Her model of information processing is very different from the information-theoretic constraint we consider.

(12) This is a minor variant on the leading example from Kamenica and Gentzkow (2011); the only difference being that we replace their judge with a jury, which is more plausibly attention constrained.

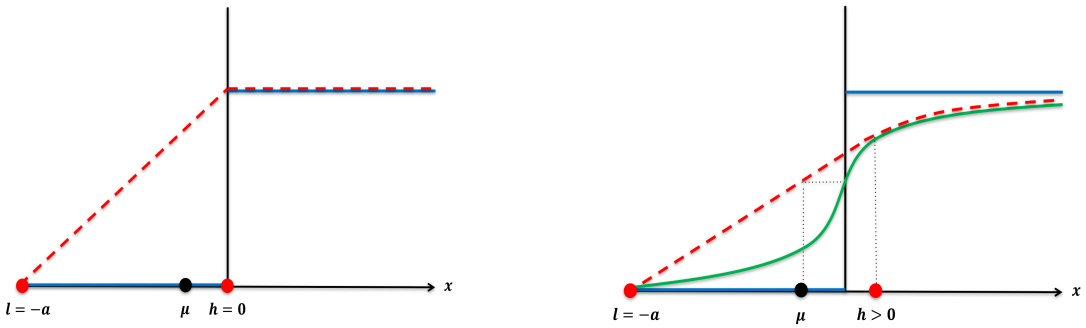

Figure 1: *The optimal signal structure when attention is free (left) and when it is costly (right).*

or *i*nnocent; the states of the world are $s \in \{g, i\}$. The prosecutor gets payoff 1 from conviction and 0 from acquittal, due to, e.g., career concerns. The jury — who, for simplicity, we model as a single agent — gets payoff $-a < 0$ from conviction when the defendant is innocent ($s = i$) and payoff $a > 0$ from conviction when the defendant is guilty ($s = g$). The jury's payoff from acquittal is zero in both states.

Denote by $\mu \in [-a, a]$ the jury's expected payoff from conviction given only prior information about the case, e.g., media coverage about the defendant.[13] The only tool at the prosecutor's disposal is the choice of what information to provide to the jury — through, e.g., the process of evidence production and argumentation in the courtroom. The prosecutor sends signals of the form $x = \mathbb{E}[s|x]$, so that signal x is intended to give Receiver a posterior mean of x. Under full attention, the jury convicts conditional on signal x if and only if $x \geqslant 0$. Importantly, Receiver's best-response *function* is *independent of the distribution of signals*.

Turning to optimal information disclosure, it is clear that when $\mu \geqslant 0$ Sender can do no better than stay silent, as the jury is willing to convict even in the absence of additional information. When $\mu < 0$, some information must be provided, and the optimal signal structure can determined graphically. The left-hand panel of Figure 1 shows the usual picture. The blue step function plots the probability that the jury convicts (and thus the prosecutor's payoff) given signal x, and the red dashed curve is its concavification. Two signals are sent: a *l*ow signal that perfectly reveals the defendant's innocence (and therefore induces posterior expectation $l = -a$, and a *h*igh signal that makes the jury indifferent between acquitting and convicting (and therefore induces posterior expectation $h = 0$). These signals are chosen to maximize the probability of conviction subject to Bayes' rule; the value function is linear on the interval $[-a, 0)$ because increasing the prior mean μ mechanically leads to a linear increase in the probability of sending the high signal. Note that the low (high) signal can be interpreted as a direct recommendation to acquit (convict).

In reality, jurors do not perfectly process the information provided to them during trial, and it is costly for them to pay attention — e.g., understand potentially complex legal instructions, internalize and recall detailed and numerous pieces of evidence, or merely stay focused after long days in the courtroom. In response, prosecutors choose carefully how to present their arguments by selecting and

(13) If the prior belief is that the defendant is guilty with probability p, we have $\mu = p \cdot a - (1 - p) \cdot a$. With a binary state space, it is equivalent to work in terms of probabilities and expectations. Though it is standard to work with probabilities in these simple examples, we work with expectations in anticipation of the message space used in the full model.

emphasizing the most important pieces of evidence, synthesizing complicated timelines into more easily-digestible narratives, and the like. Our model formalizes these effects by assuming that, if the prosecutor sends a signal intended to induce posterior mean x, the jury "sees" a noisy version of this signal. Thus, even if $x \geqslant 0$ so that it is optimal to convict, the jury will "make a mistake" and acquit with positive probability; similarly, if $x < 0$, the jury will convict with positive probability. Moreover, and this is the most important point, we assume that the jury is sophisticated in the sense that it optimally chooses the correlation between these mistakes and the prosecutor's signals subject to a particular form of attention costs. The process of optimal attention allocation will imply that the jury's best response *function* is *endogenous* and *depends on the entire distribution of signals*.

Suppose that the prosecutor uses the optimal signal structure derived under the assumption of full attention. If paying attention is costly, the jury will simply ignore these signals and acquit for certain: conditional on the low signal acquittal is optimal, so it gets payoff zero; at the high signal, it is indifferent, so must also get payoff zero. Thus, there is no point in sinking any costs to process these signals.

Instead, the prosecutor must incentivize the jury to pay attention by revealing enough information so that, conditional on some signals, conviction is *strictly* preferred. What is the optimal way to resolve this tradeoff? As before, we can solve for the optimal policy by concavification, though the argument is more subtle. Consider the right-hand panel of Figure 1. The blue step function, as before, is the jury's full-attention best-response function. The green curve is the jury's *stochastic choice rule*. Because the jury does not fully process signals, the probability of taking the correct action is always less than one; this is represented by the green curve being bounded by the blue step function. But the jury allocates its limited attention optimally, meaning it is more likely to convict when the signal x is higher — i.e., is more *convincing* of the defendant's guilt. This is represented by the fact that the green curve is strictly increasing. The fact that it is S-shaped follows from the form of the RI cost function.

The red dashed curve is the concavification of the stochastic choice rule. By analogy to the full attention case, it is natural to conjecture that we can read off the optimal signals from the concavification. This procedure results in the binary signal structure illustrated in the figure, where the only difference from the full attention case is that the high signal induces posterior mean $h > 0$, as we've argued must be the case. But this logic is not quite correct: we took the best response function (green curve) as given when reading off this signal structure but, as argued above, the best response function is itself endogenous to the signal structure! Somewhat surprisingly, it turns out that, at the optimum, taking the best response function as given is indeed without loss. Given a distribution of signals, a necessary condition for optimality of the jury's attention strategy is that the probability of conviction conditional on signal $x = 0$ (where the jury is indifferent between conviction and acquittal) equals the unconditional probability of conviction (i.e., averaged over all signal realizations). Intuitively, the RI cost function penalizes deviations from the average action, so the optimal stochastic choice rule must be "centered" around the average action. This is represented graphically by the fact that the y-intercept of the green curve equals the value of the concavification at the prior. This argument will be formalized through our first-order approach.

The endogeneity of the best-response function also leads to comparative statics that are absent from the full-attention model. What if, for example, the prosecutor faces a more "pessimistic" jury —

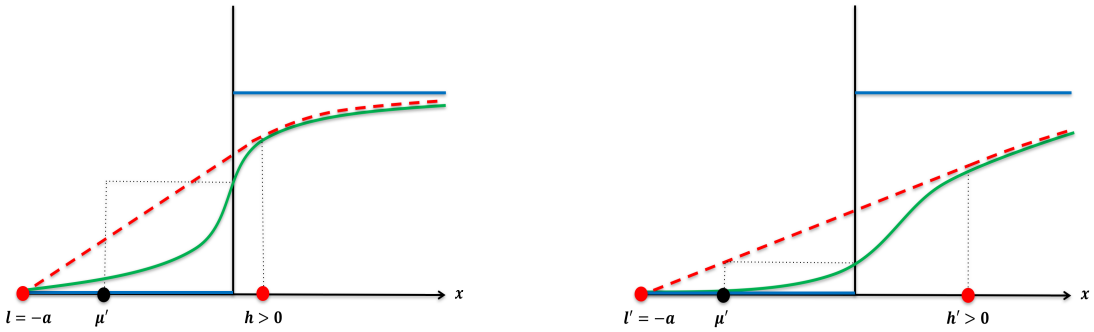

Figure 2: *Comparative statics of optimal signal structure with respect to the prior mean.*

i.e., one whose prior belief puts more weight on the defendant's innocence? This is depicted in Figure 2. Two things happen. First, the jury believes that it is less likely that a high signal will be sent. Because high signals — the only ones that would change the optimal "status quo" action of acquittal — are less likely, to economize on attention costs the jury will pay less attention. Thus, it relies more heavily on its prior belief and is more likely to acquit given any signal realization. This is illustrated in the left-hand panel of Figure 2: with the green curve held fixed, the necessary optimality condition is not satisfied; it is optimal for the jury to shift the green curve downward, and thus acquit more frequently. If the prosecutor stuck with the high signal $h > 0$, the jury would pay too little attention (the green curve would shift down too far). This leads to the second effect: the optimal way for the prosecutor to counteract this loss of attention is to make the high signal *more convincing* (represented as $h' > h$). The net effect, illustrated in the right-hand panel of Figure 2, is that the high signal increases to $h' > h$ (and therefore becomes less likely) and that the stochastic (the green curve) shifts down to the point where the jury's optimality condition holds. Again, the optimal signals can be read off of the concavification (the red dashed curve) but, as before, that this is true is not obvious and requires additional argument.

A full analysis of the binary-state model, including this example, can be found in Appendix C. The main model, to which we now turn, generalizes this example to a wider class of preferences and continuous prior distributions, where the effects of inattention are much more subtle.

3. Model

3.1. Set-up

The state of nature, s, is the realization of a real-valued random variable S with distribution function G, which we assume has a non-singleton, compact support $\mathcal{S} := \text{supp}(G)$. We assume further that $\underline{s} < 0 < \overline{s}$, where $\underline{s}$ and $\overline{s}$ are, respectively, the smallest and largest elements of $\mathcal{S}$. With a slight abuse of terminology, we say that G has *full support* if $\text{supp}(G) = [\underline{s}, \overline{s}]$. The expectation of S is denoted $\mu := \mathbb{E}_G[S]$.

There are two players, Sender and Receiver. Both are expected utility maximizers. Receiver ultimately makes a choice $a \in \mathcal{A} := \{0, 1\}$, where we interpret $a = 0$ as "inaction" and $a = 1$ as "action."

The gross utility functions for Sender and Receiver are given, respectively, by

$$v(a, s) = a \cdot (\alpha + \beta \cdot s)$$

and

$$u(a, s) = a \cdot s$$

The payoff to inaction is normalized to zero for both players. Receiver's payoff from action is equal to the state, and Sender's payoff from action is an affine transformation of Receiver's. When $\alpha = 0$ and $\beta = 1$, so that $v(\cdot) = u(\cdot)$, we say that preferences are *aligned*. When $\alpha = 1$ and $\beta = 0$, we say that Sender's preferences are *state independent*. We assume throughout that $\beta \geqslant 0$,[14] and when $\beta > 0$ we say that preferences are *strictly co-monotone*.

Given a domain $\mathfrak{X}$ (a measure space, with the σ-algebra suppressed), an *information structure* $(\gamma, \mathcal{Y})$ consists of a measurable space $\mathcal{Y}$ and a Markov kernel γ from $\mathfrak{X}$ to $\mathcal{Y}$.[15] Both Sender and Receiver will choose information structures.

To influence Receiver's action choice, Sender commits to a *persuasion strategy*, denoted by $(\pi, \mathcal{X})$, which is an information structure with domain $\mathcal{S}$. Elements $x \in \mathcal{X}$ are called *signals*, so that $\pi : s \mapsto \pi(s, \cdot) \in \Delta(\mathcal{X})$ maps states into distributions over signals. Receiver perfectly observes Sender's choice of persuasion strategy, which we will refer to interchangeably as the *mechanism* or *signal structure*.

The key feature of the model is that Receiver has limited information-processing capacity, or *attention*, and must exert costly effort to process Sender's signals. We model this as follows. After observing $(\pi, \mathcal{X})$, but before a signal x is realized or an action choice is made, Receiver selects an *attention strategy*. An attention strategy, denoted by $(\mu, \mathscr{M})$, is an information structure with domain $\mathcal{X}$. Elements $m \in \mathscr{M}$ are called *perceptions*. Thus, $\mu : x \mapsto \mu(x, \cdot) \in \Delta(\mathscr{M})$ maps signals into distributions over perceptions.

Intuitively, Receiver only observes Sender's signals with noise, and an attention strategy describes how much and what kind of noise Receiver chooses to add to Sender's chosen information structure. Examples of attention strategies include:

- *Finite categorization:* Receiver selects a partition $P = \{P_1, \ldots, P_n\}$ of $\mathcal{X}$ and perceptions m indicate which cell of the partition x belongs to. Formally, $M = P$ and $\mu(x, P_i) = \mathbb{1}(x \in P_i)$.

- *Noisy mistakes:* Suppose Sender transmits a binary signal, $\mathcal{X} = \{x_1, x_2\}$. Receiver's perception is drawn from $\mathscr{M} = \{m_1, m_2\}$ according to $\mu(x_i, m_i) = p \in (\frac{1}{2}, 1)$.

- *Additive noise:* Suppose Sender transmits real-valued signals, $\mathcal{X} = \mathbb{R}$. Receiver observes Sender's signal with additive Gaussian noise, i.e., $\mathscr{M} = \mathbb{R}$ and $m = x + \epsilon$ where $\epsilon \perp x$ and $\epsilon \sim N(0, 1)$.

(14) This for expositional simplicity only. Our results easily extend to the case $\beta < 0$.

(15) Recall that a Markov kernel is a measurable mapping $\gamma : \mathfrak{X} \times \mathcal{F}_{\mathcal{Y}} \to [0, 1]$ such that, for every $z \in \mathfrak{X}$, $\gamma(z, \cdot)$ is a probability measure on $\mathcal{Y}$, and where $\mathcal{F}_{\mathcal{Y}}$ denotes the σ-algebra on $\mathcal{Y}$. In our solution, all spaces will be compact metric and endowed with the Borel σ-algebras.

We assume that *all* attention strategies are feasible. Thus, Receiver has full flexibility in choosing (i) *whether* to pay attention, (ii) *how much* to pay attention, and (iii) *what to to pay attention to*, without being restricted to any of the parametric classes described above. The cost of paying attention is described by a cost function

$$C : \mathscr{P} \times \mathscr{A} \to \overline{\mathbb{R}}$$

where $\mathscr{P}$ is the set of all persuasion strategies and $\mathscr{A}$, the set of all attention strategies.[16] Observe that attention costs may *directly* depend on both on Receiver's attention strategy and Sender's persuasion strategy.

While Receiver has limited attention in the above sense, he is fully rational and allocates it optimally. Given a persuasion strategy, Receiver chooses an attention strategy and, conditional on the realized perception m, an action to maximize his expected utility net of attention costs. Formally, given the prior, a persuasion strategy $(\pi, \mathfrak{X})$, and an attention strategy $(\mu, \mathscr{M})$, every realized m induces a posterior belief about the state, denoted by $\nu(m, \cdot) \in \Delta(\mathcal{S})$. For example, if G admits a density g, this posterior can be written as

$$\nu(m, s) = \frac{g(s) \cdot \int_{\mathfrak{X}} \mu(dx, m) d\pi(s, dx)}{\int_{\mathcal{S} \times \mathfrak{X}} \mu(dx, m) \pi(ds, dx) dG(s)}$$

Thus, given realized perception m, Receiver optimally chooses an action in the set

$$A(m) := \arg \max_{a \in \{0,1\}} \int u(a, s) \nu(m, ds)$$

As usual, we assume that any ties are broken in Sender's favor.[17] Denote a Sender-optimal element of $A(m)$ by $\hat{a}(m)$.

Receiver's optimal choice of attention strategy, $(\mu^*, \mathscr{M}^*)$, is then given by the solution to *Receiver's problem*

[RP] $$\max_{(\mu, \mathscr{M})} \int_{\mathcal{S} \times \mathfrak{X} \times \mathscr{M}} u(\hat{a}(m), s) \mu(dx, dm) \pi(ds, dx) dG(s) - C\left((\pi, \mathfrak{X}); (\mu, \mathscr{M})\right)$$

Similarly, Sender's optimal choice of persuasion strategy, $(\pi^*, \mathfrak{X}^*)$, is then given by the solution to *Sender's problem* [SP], which is subject to Receiver's *obedience constraint* [Ob]:

[SP] $$\tilde{V}(G) := \max_{(\pi, \mathfrak{X}), (\mu, \mathscr{M})} \int_{\mathcal{S} \times \mathfrak{X} \times \mathscr{M}} v(\hat{a}(m), s) \mu(dx, dm) \pi(ds, dx) dG(s)$$

[Ob] $$\text{s.t. } (\mu, \mathscr{M}) \text{ solves [RP]}$$

In summary, the timing of the game is as follows (see Figure 3). First, Sender publicly commits to a persuasion strategy. Second, Receiver chooses an attention strategy. Third, the state-signal-perception triple (s, x, m) is realized, and Receiver only observes m. Finally, Receiver makes a binary action choice and payoffs are realized for both parties. As usual, the solution concept is Sender-preferred subgame perfect Nash equilibrium.

(16) This is a slight abuse of notation, as $C(\cdot)$ is only well-defined for persuasion-attention strategies pairs that are *consistent* in the sense that the domain of the attention strategy and range of persuasion strategy coincide.

(17) It will follow from the solution to Receiver's RI problem that he is in fact never indifferent between actions on the equilibrium path.

Figure 3: *Model timeline.*

3.2. RI Cost Function and Interpretation

We follow the literature on rational inattention and posit an attention cost function that is proportional to the mutual information between Sender's signals and Receiver's perceptions.

Definition 1. Given two random variables Y and Z defined on the same probability space with joint distribution $P_{(Y,Z)}$ and marginals P_Y and P_Z, the *mutual information* between Y and Z is given by

$$I(Y;Z) := \begin{cases} \int \log\left(\frac{dP_{(Y,Z)}}{dP_Y \times dP_Z}\right) dP_{(Y,Z)}, & \text{if } P_{(Y,Z)} \ll P_Y \times P_Z \\ +\infty, & \text{else} \end{cases}$$

Note that every persuasion strategy $(\pi, \mathcal{X})$ and attention strategy $(\mu, \mathscr{M})$ together induce a pair of random variables X and M on the same space with joint distribution

$$P_{(X,M)}(A,B) = \int_{\mathcal{S}\times A\times B} \mu(dx,dm)\pi(ds,x)dF(s)$$

where $A \subseteq \mathcal{X}$ and $B \subseteq \mathscr{M}$ are measurable sets.

Assumption 1. Receiver's attention cost function is proportional to the mutual information between X and M, namely,

$$C((\pi,\mathcal{X});(\mu,\mathscr{M})) = \lambda I(X;M)$$

for some $\lambda > 0$.

A Receiver who solves problem [RP] with the cost function in Assumption 1 is said to be *rationally inattentive* (RI), and [RP] is said to be his *rational inattention (RI) problem given* $(\pi, \mathcal{X})$. Assumption 1 will stand for the remainder of the paper.

We view our assumptions as capturing the steady-state behavior arising from a series of repeated interactions between Sender and Receiver. As usual, we think of Sender's commitment power as driven by reputational considerations inherent in such a relationship. On Receiver's end, Assumption 1 can be interpreted as a reduced-form representation of a process of *optimal coding* that would plausibly arise after repeated exposure to Sender's signals. Two examples that map quite closely on to this story are as follows:

- Sender is Amazon, which makes a sequence of product recommendations to Receiver, who is a consumer deciding which ones to purchase. While some recommendations are explicit, many are

implicit, e.g., the placement of products on the consumer's homepage. Through repeated exposure, the consumer learns which type of recommendations are typically most valuable, where to locate them on the site, and how to optimize his search process for them. Amazon, in turn, designs its recommendation system taking into account the consumer's response.

- Sender is a specialized advisor to Receiver, an executive with decision-making power. They interact daily in a morning briefing and, over time, the executive learns how to ask the sequence of yes/no questions that most efficiently elicits from the advisor the information needed for her daily decisions. Informally, the advisor's design problem can be thought of choosing the optimal way to structure her presentation at the morning briefing.

More formally, we think of Sender and Receiver as operating a jointly-controlled *communication channel* in the sense of information theory (Cover and Thomas (2006)), with Sender choosing the *source* and Receiver choosing the *encoding* and *noise structure*. It is well-known that, when the random variables are discrete, mutual information can be written as

$$I(X;M) = H(X) - H(X|M) \tag{3.1}$$

$$= H(M) - H(M|X) \tag{3.2}$$

where, e.g., $H(X)$ is the entropy of X and $H(X|M)$ is the conditional entropy of X given M.[18] Entropy is a measure of the uncertainty inherent in a random variable, and mutual information is a measure of the expected reduction of uncertainty. For example, $H(X) - H(X|M)$ measures the extent to which observing M, averaged over realizations m, reduces one's uncertainty about the realization of X.[19] In particular, the Source Coding Theorem states that $H(X)$ is equal to the "amortized communication cost" of X, i.e., the minimum number of bits (answers to binary yes/no questions) required to determine the value of n independent draws of X with zero error, on average over realizations and size of the sample.[20] Mutual information similarly represents the minimum number of bits needed, on average (in the same amortized sense), to transmit without error a (potentially noisy) message M about X, given their joint distribution. Thus, the RI cost function $\lambda I(M;X)$ represents the long-run average cost to Receiver of operating a communication channel with Sender, who chooses the source X, given that (i) he chooses a particular joint distribution for his perceptions and Sender's signals and (ii) chooses the optimal sequence of binary yes/no questions to ask about X at marginal cost λ per question (i.e., per bit).[21]

(18) When X and M have finite support, *entropy* is defined by $H(X) := -\sum_{x \in \text{supp}(P_X)} P_X(x) \log\left(P_X(x)\right)$ and *conditional entropy* is defined by $H(X|M) := -\sum_{m \in \text{supp}(P_M)} P_M(m) \sum_{x \in \text{supp}(P_{X|M=m})} P_{X|M=m}(x) \log\left(P_{X|M=m}(x)\right)$. If X admits a density (with respect to Lebesgue measure), then we can replace entropy with *differential entropy* in the formula for mutual information, but the interpretation becomes more subtle. The formula in Definition 1 is always valid. See, e.g., Chapter 8 of Cover and Thomas (2006).

(19) Mutual information is symmetric in the sense stated in the display. The expression [3.1] represents reduction of uncertainty from Receiver's standpoint — namely, how informative his perception is about the signal. Expression [3.2], on the other hand, represents reduction of uncertainty from Sender's standpoint — namely, how informative his signal is about the random perception that Receiver will draw from it.

(20) "Bits" are associated with the base-2 logarithm, while we work with natural logarithms and the corresponding unit, "nats." We slightly abuse terminology and refer to bits throughout, as this is more intuitive and equivalent up to a multiplicative constant that could be absorbed into the marginal cost λ.

(21) Looking at the amortized costs means that we abstract from the fixed costs of choosing the optimal encoding itself, and

The fact that Receiver's marginal cost λ is (i) taken as given and (ii) strictly positive is a manifestation of the idea that Receiver operates in an "information-rich world" (Simon (1971)). A direct action recommendation only requires sending a single bit (as the action choice is binary), which should be easy to transmit without noise when taken in isolation. But, as in examples of the online consumer and corporate executive, we view the action choice in question as being merely one of many that Receiver must attend to; while he could easily pay full attention in any given decision problem, his attention becomes a scarce resource when spread among many. In the background, we think of Receiver as having a convex cost function over his aggregate attention allocation, e.g., a capacity constraint as in the original formulation of Sims (2003). The marginal cost λ represents the Lagrange multiplier from this overall problem; we take it as given because the decision in question requires attention that is negligible in the aggregate.

While this fairly literal information-theoretic interpretation is especially natural in our setting of bilateral communication, it is important to note that there are other justifications for the RI cost function. For example, Hébert and Woodford (2017) and Morris and Strack (2017) have shown that the RI cost function arises naturally (though not universally) in models of sequential sampling in which the agent observes a diffusion process and decides when to stop and take an action. In our setting, this corresponds to a story in which Receiver chooses a strategy for contemplating about what Sender has said, at some subjective mental cost. This foundation has the arguable advantage that it does not rely on repeated interactions, which are needed to justify mutual information as an amortized communication measure.

4. First-Order Approach to Sender's Problem

4.1. Overview

In this section, we simplify Sender's problem using a *first-order approach (FOA)*. This will be done in two steps. First, we present the solution to Receiver's RI problem. A crucial feature of the solution is that, despite the inherently infinite-dimensional nature of the RI problem, in our binary-action setting the obedience constraint [Ob] can be reduced to a single first-order condition for a scalar *activity parameter*. This allows us, in the second step, to formulate Sender's problem in a way that formally resembles a standard moral hazard problem in which the agent (in this case, Receiver) has a one-dimensional effort choice.

In the process, we answer the critical question: *what is the appropriate signal space?* There are at least three standard — and, under appropriate assumptions, equivalent — answers in the literature. The *belief-based* approach identifies signals with the posteriors they induce, the *Revelation Principle* approach identifies them with the induced optimal actions, and the *posterior mean* approach identifies them with the induced posterior expectations. The main complication in our setting is that Receiver does not directly observe the signals, only his perceptions of them. Moreover, he makes a decision — his choice of attention strategy — after observing Sender's persuasion strategy but before any signals are

instead focus on the "implementation" costs of operating the channel.

realized. His incentives at the attention-choice stage are determined by the *entire persuasion strategy*, not just on the beliefs that *individual signals* are intended to induce. Thus, the standard reductions do not immediately apply. However, we will show that our setting is in fact amenable to the posterior mean approach. Later, in Section 7.1, we discuss the limits of our approach and appropriate generalizations.

4.2. Solution to Receiver's RI Problem

The first critical observation is that rationally inattentive Receiver will never acquire more information than is necessary to make an action choice. To see this, consider an attention strategy featuring two perceptions $m_1 \neq m_2$ such that $\hat{a}(m_1) = \hat{a}(m_2)$. Receiver can always construct a new attention strategy in which these messages are pooled, i.e., he receives a new message m_3 whenever the old attention strategy would have sent m_1 or m_2, and the remainder of the strategy is unaffected. This will not change the chosen action — and thus Receiver's expected utility gross of attention costs — but is less informative about Sender's signal. Because mutual information is strictly Blackwell monotone, this pooled attention strategy strictly lowers Receiver's attention costs.

More formally, an attention strategy is called a *direct recommendation* if $\mathscr{M} = \mathscr{A}$ and $\hat{a}(m) = m$. The above argument shows that any solution to the RI problem must be a direct recommendation strategy. Restricting attention to direct recommendations is a substantial simplification. In particular, we may interpret Receiver's problem as one of choosing a *stochastic choice rule*, i.e., a measurable mapping $p : \mathfrak{X} \to [0, 1]$, where $p(x)$ denotes the probability of action (choosing $a = 1$) when the realization of Sender's signal is x. Receiver's problem may then be written as

$$\max_{p(\cdot)} \mathbb{E}_{\pi} \left[\hat{u}(x)p(x)\right] - \lambda I(X; A) \tag{4.1}$$

where the maximization is over all measurable function $p : \mathfrak{X} \to [0, 1]$ and $\hat{u}(x) := \mathbb{E}_{G,\pi}[s|x]$ is the expectation of Receiver's utility from action (the state) given signal realization x. It is convenient to solve this problem in two steps. First, fix a number $k \in \overline{\mathbb{R}}_+$, which we will refer to as the *activity parameter*. It represents a "target" for the unconditional (on the signal realization) likelihood ratio of action to inaction, $\Pr(a = 1)/\Pr(a = 0)$, and is thus a scalar measure for how "active" Receiver is. For example, $k = 0$ corresponds to inaction ($a = 0$) — and $k = +\infty$, to action ($a = 1$) —regardless of the signal realization. Taking k as given, variational arguments can be used to pin down the shape of a candidate optimal stochastic rule. This gives us a family of candidate solutions parametrized by the one-dimensional parameter k, denoted by $p(\cdot; k)$. The second, and final, step is then to optimize over the activity parameter k. The following result summarizes the above discussion and fully characterizes the solution.

Definition 2. A persuasion strategy $(\pi, \mathfrak{X})$ is *degenerate* if $\hat{u}(x) = 0$ π-almost surely. Otherwise, it is said to be *non-degenerate*.

Proposition 1. *Given any non-degenerate persuasion strategy $(\pi, \mathfrak{X})$, Receiver's RI problem has an essentially unique*[22] *solution. It satisfies the following properties:*

(22) That is, the solution is uniquely determined up to π-null sets.

1. *It is a direct recommendation attention strategy.*
2. *For fixed activity parameter $k \in \overline{\mathbb{R}}_+$, the optimal stochastic choice rule is given by*

$$p(x;k) = \frac{k \exp\left(\frac{\hat{u}(x)}{\lambda}\right)}{1 + k \exp\left(\frac{\hat{u}(x)}{\lambda}\right)}$$

   π-almost surely.
3. *The first-order optimality condition for k is*

[4.2] $$P(k) := \frac{k}{k+1} = \mathbb{E}_\pi\left[p(x;k)\right]$$

   It has at most one interior *solution $k \in \mathbb{R}_{++}$ while $k \in \{0, +\infty\}$ are always solutions.*
4. *The optimal choice of k, call it k^*, is determined as follows:*
   (*a*) *$k^* = 0$ if and only if*

[4.3] $$\mathbb{E}_\pi\left[e^{\hat{u}(x)/\lambda}\right] \leqslant 1$$

   (*b*) *$k^* = +\infty$ if and only if*

[4.4] $$\mathbb{E}_\pi\left[e^{-\hat{u}(x)/\lambda}\right] \leqslant 1$$

   (*c*) *$k^* \in (0, \infty)$ if and only if*

[4.5] $$\mathbb{E}_\pi\left[e^{\hat{u}(x)/\lambda}\right] > 1 \quad \textit{and} \quad \mathbb{E}_\pi\left[e^{-\hat{u}(x)/\lambda}\right] > 1$$

   in which case it is the unique interior solution to [4.2]*. Conversely, if* [4.2] *is solved by some $k \in \mathbb{R}_{++}$, then $k = k^*$ and, in particular,* [4.5] *is satisfied.*

Proof. See Appendix A.1 for a proof sketch.[23] □

The solution has several noteworthy features. Most immediately, Receiver's action choice is generally random from Sender's point of view, even contingent on the realization of the signal x generated by her persuasion strategy. That is, Receiver *makes mistakes*. More technically, Receiver either completely ignores the state (parts (a) and (b) of Proposition 1) or follows an interior choice rule such that $p(\cdot;k) \in (0,1)$ (part (c) of the proposition). In each case, he never *perfectly* observes *any* state. Intuitively, this is because the RI cost function incentivizes *attention smoothing* over the state space: it is very costly to perfectly distinguish between states (e.g., if $p(\cdot;k)$ were not constant but $p(x;k) \in \{0,1\}$ for some x) or allocate a "discrete" amount of attention toward any particular event (e.g., if $p(\cdot;k)$ had jump discontinuities, and thus infinite slope).

(23) Proposition 1 is essentially a restatement of Lemma 1 and Proposition 2 of Yang (2017); alternatively, it follows from Lemma 1, Theorem 1, and Corollary 2 of Matějka and McKay (2015). However, both papers assume that the RI agent's prior is either discrete or admits a density with respect to Lebesgue measure, and the way they define the cost function directly in terms of (differential) entropy reduction reflect this assumption. We must allow for Receiver's "prior" (the distribution of X) to be completely general to avoid placing arbitrary restrictions on Sender's persuasion strategies. As shown in the appendix, it is easy to incorporate this additional level of generality.

When the solution is interior, the stochastic choice rule described in part 2 of Proposition 1 takes the form of a *shifted Logit rule*.[24] The activity parameter k induces first-order shifts in the choice distribution; the marginal attention cost λ induces second-order shifts. As $\lambda \to 0$, $p(\cdot; k)$ converges to a step function.[25] This has two important consequences. First, $p(\cdot; k)$ is strictly increasing. Thus, *stakes matter*: even though Receiver should act given both $\hat{u}(x') > \hat{u}(x) \geqslant 0$, he is more likely to act (less likely to make a mistake) given x'. This is a natural feature for a model of rational attention allocation. Second, $p(\cdot; k)$ is S-shaped. Thus, there is *differential sensitivity* in Receiver's attention allocation. When $p(\cdot; k)$ is steeper, Receiver is effectively distinguishing more finely between adjacent states. Indeed, we may view

$$\Delta(x; k) := \frac{\partial p(x; k)}{\partial x} \tag{LAI}$$

as a measure of *local attention intensity at* x.[26] Both the *level effects* related to the size of the stakes and the *marginal effects* related to local attention intensity will driving forces in determining Sender's optimal persuasion strategy.

It is somewhat curious that the first-order condition in part (c) is identical to the consistency condition that k is, in fact, the unconditional likelihood ratio of action. Given the important role it plays in our analysis, it is worthwhile to develop additional intuition. Using part 2 of the proposition, the optimality condition in part 3 may be equivalently written as $p(0; k) = \mathbb{E}_\pi [p(x; k)]$. This is perhaps more intuitive: there is always a tendency to keep $p(x; k)$ near its average value $\mathbb{E}_\pi [p(x; k)]$ because the RI cost function penalizes "variation" of the stochastic choice rule and, when $x = 0$, Receiver is indifferent between actions, so any deviation of $p(0; k)$ from the average has no benefit but adds to attention costs. When there is an $x \in \text{supp}(\pi)$ with $\hat{u}(x) = 0$, a variational argument implies that $p(0; k)$ and the average must coincide at any optimum; for general distributions, $p(0; k)$ needn't be pinned down by the Logit rule, but the first-order condition still holds.

Finally, the converse statement in part 4(c) is important for the tractability of the first-order approach. The inequalities in part 4 correspond to second-order conditions that an optimal solution must satisfy and, in general, would need to either be incorporated directly as constraints in Sender's problem or verified ex post. Fortunately, that is not needed: though the first-order condition [4.2] always has at least two solutions, it has at most one *interior* solution that, if it exists, corresponds to a global optimum.

4.3. Sender's Persuasion Problem

With Proposition 1 in hand, we are now in a position to simplify Sender's persuasion problem. First, we must identify the appropriate signal space, $\mathfrak{X}$. For purposes of determining Receiver's incentives, it is

(24) The solution is only pinned down on the support of π, but it is without loss to assume it takes the Logit form globally. Because Sender is a Stackelberg leader, even if she shifts the support of her persuasion strategy, Receiver will optimally respond by "filling in" the Logit curve at appropriate points.

(25) In the binary-state example of Section 2, $p(x; k)$ was represented by the green curves in Figure 2. As $\lambda \to 0$, it converges to the blue step function.

(26) This intuition plays an important role in both Yang (2017) and Morris and Yang (2016).

without loss to identify signals with the conditional expectations about the state, or *posterior means*, that they induce. This follows from the structure of the solution to the RI problem alone: all "higher moments" of the information structure are irrelevant for computing expected payoffs — hence, optimal actions — so an RI agent will rationally ignore them. Because we have assumed that Sender's utility is affine in the state, the posterior mean distribution is all that matters for computer her payoffs, as well.

More formally, we say that two persuasion strategies are *outcome equivalent* if they induce the same joint distribution over state-action pairs. They are *Sender-payoff equivalent* if they induce the same joint distribution over pairs $(s, \mathbb{E}[v(a,s)|x])$ of states and Sender conditional expected payoffs. They are *equivalent* if they are both outcome and Sender-payoff equivalent. A persuasion strategy $(\pi, \mathcal{X})$ is said to be *straightforward* if $\mathcal{X} = \text{CH}(\mathcal{S})$ and $x = \mathbb{E}_\pi[s|x]$. Given two CDFs H and H', we say that H *is a mean-preserving contraction of* H', denoted $H \leqslant^{MPS} H'$, if H' is a mean-preserving spread of H.

Lemma 1. *For every persuasion strategy, there exists an equivalent straightforward strategy. Moreover,* $(\pi, \mathcal{X})$ *is straightforward if and only if* $X \sim F$ *for some* $F \leqslant^{MPS} G$.

Proof. The solution to Receiver's RI problem characterized in Proposition 1 has the feature that the stochastic choice rule $p(\cdot)$ is a function of x only through $\hat{u}(x) := \mathbb{E}_\pi[s|x]$. Moreover, from the affine form of Sender's utility function, her expected payoff conditional on signal x is, by the law of iterated expectations, simply $(\alpha + \beta \cdot \hat{u}(x)) \cdot p(x)$, and thus is also only a function of $\hat{u}(x)$. This establishes the existence of an equivalent straightforward strategy. The characterization of straightforward strategies in terms of the mean-preserving contraction constraint is familiar from, e.g., Blackwell (1953). □

Given Lemma 1, we henceforth identify signals with their induced posterior means, and identify the Markov kernel π with its induced posterior mean distribution (i.e., the marginal distribution $\text{marg}_{\mathcal{X}}\pi$). That is, $\mathcal{X} \equiv \text{CH}(\mathcal{S}) = [\underline{s}, \overline{s}]$ and $\pi \equiv F \leqslant^{MPS} G$. It remains to incorporate Receiver's obedience constraint [Ob] in a tractable way. Parts 1 and 2 of Proposition 1 pin down Receiver's optimal attention strategy to a one-parameter family of stochastic choice rules. At any interior solution, the activity parameter k is in turn pinned down by the first-order condition [4.2] in part 3 of the proposition. We take a *first-order approach* (FOA) by replacing [Ob] with this first-order condition. Our solution method thus follows the two-step procedure typical of moral hazard problems, with the activity parameter k playing the role of the usual one-dimensional "effort" variable. In the first step, we take k as given and solve for the optimal posterior-mean distribution that implements k. Sender's *component problem* is given by

$$\text{[CP]} \qquad V(k|G) := \sup_{F \leqslant^{MPS} G} \mathbb{E}_F[(\alpha + \beta \cdot x) \cdot p(x;k)]$$

$$\text{[A]} \qquad \text{subject to} \qquad \mathbb{E}_F[p(x;k)] = P(k)$$

and any solution is denoted $F^*(k)$.[27] It is notable that the component problem [CP] is a linear program, albeit an infinite-dimensional one: both the objective function and constraint are linear functionals of F, and the mean-preserving contraction constraint can also be represented by a convex, compact feasible set determined by a linear functional. This will allow us to use LP duality to characterize the solution.

(27) The feasible set is compact and the objective continuous, so a solution exists whenever the feasible set is nonempty. If the feasible set is empty at k, we set $V(k|G) = -\infty$.

In the second step, we optimize over the implementable activity levels. Sender's *relaxed problem* is thus given by

[RP] $$V(G) := \sup_{k \in \mathbb{R}_{++}} V(k|G)$$

and any solution is denoted k^*. We call $F^* := F^*(k^*)$ a *FOA-optimal persuasion strategy*, while (F^*, k^*) is referred to as a *FOA-optimal pair*. Henceforth, our goal is to characterize FOA-optimal pairs. This would be the end of the story, save for one small but important detail. Notice that [RP] only optimizes over interior activity levels. This is because the FOA is *almost valid*: by Proposition 1 Receiver's first-order condition is sufficient (for optimality in his RI problem) only when k is interior. To complete the characterization of Sender's optimal persuasion strategy, we need to separately check extremal activity activity levels, taking into account Receiver's second-order conditions. Define[28]

[4.6] $$\overline{V}(G) := \begin{cases} v(1,\mu), & \text{if } \overline{D}(G) \neq \emptyset \\ -\infty, & \text{else} \end{cases}$$

where $\overline{D}(G) := \left\{F \leqslant^{MPS} G : \mathbb{E}_F\left[e^{-x/\lambda}\right] \leqslant 1\right\}$ and

[4.7] $$\underline{V}(G) := \begin{cases} 0, & \text{if } \underline{D}(G) \neq \emptyset \\ -\infty, & \text{else} \end{cases}$$

where $\underline{D}(G) := \left\{F \leqslant^{MPS} G : \mathbb{E}_F\left[e^{x/\lambda}\right] \leqslant 1\right\}$. Clearly $\hat{V}(G) := \max\left\{V(G), \overline{V}(G), \underline{V}(G)\right\}$ is an upper bound for the value $\tilde{V}(G)$ of Sender's (full) problem [SP]. The following proposition shows that to solve [SP], it suffices to (i) find any FOA-optimal pair, should one exist and (ii) if none exist, determine which extremal activity level is optimal; any persuasion strategy F^* generated by this procedure solves [SP].

Definition 3. Activity parameter $k \in \overline{\mathbb{R}}_+$ is *implementable* at prior G if there exists some $F \leqslant^{MPS} G$ such that $p(\cdot; k)$ solves Receiver's RI problem given F. In that case, we say that k *is implemented by* F. The set of implementable activity parameters at G is denoted $K(G)$.

Proposition 2. *A solution to Sender's problem* [SP] *exists, and is characterized as follows:*

1. *If there exist an FOA-optimal pair* (F^*, k^*)*, then it solves Sender's problem. That is,* $V(G) \geqslant \hat{V}(G)$ *and* k^* *is implemented by* F^**;*
2. *If an FOA-optimal pair does not exist, then Sender's problem is solved either by* $(F, 0)$ *for some* $F \in \underline{D}(G)$ *or* (F, ∞) *for some* $F \in \overline{D}(G)$*.*[29]

Proof. See Appendix A.2. □

(28) Recall, $v(1,\mu) := \alpha + \beta \cdot \mu$ is Sender's expected payoff from action given that the state is μ; by linearity, this is also her expected utility if the prior mean is μ.

(29) By construction, any $F \in \overline{D}(G)$ implements $k = +\infty$ and any $F \in \underline{D}(G)$ implements $k = 0$.

4.4. Connection to Privately-Informed Receiver

We pause to note the formal connection between Sender's component problem [CP] and the recent models of persuasion with a privately-informed, but fully attentive, Receiver of Kolotilin et al. (2017) and Kolotilin (2017). Both papers study models, like ours, in which Receiver has a binary action choice, both Sender and Receiver have utility that is affine in a scalar state of the world, and Receiver's action is stochastic from Sender's point of view, even conditional on the realized signal.[30] For example, Kolotilin et al. (2017) assume that Receiver privately observes a random variable $R \sim H$ with support on $[\underline{r}, \overline{r}]$, which is distributed independently of the state S, and that his utility function is given by $u(a, s, r) = a \cdot (s - r)$. Thus, R is Receiver's *reservation value*. Suppose Sender's persuasion strategy generates signal $x = \mathbb{E}[s|x]$. Then, from Sender's point of view, the conditional probability that Receiver acts is $H(x) = \Pr(r \leqslant x)$. *Ignoring the activity constraint*, Sender's component problem [CP] from our model can be nested as a special case of the private information model. Namely, for fixed k simply define the CDF H by $H(\cdot) := p(\cdot; k)$. Then both models induce the same stochastic choice rule and, hence, identical solutions to Sender's (component) problem. This observation can be understood by noting that the inattentive Receiver in our model effectively *generates* private information through his attention strategy. Namely, Receiver's perception is privately observed, random conditional on the realized signal, and is the ultimate determinant of his action choice — just like Receiver's reservation value in the private information model.

The critical difference is that, in those papers, the stochasticity of Receiver's choice is driven by *exogenous* private information about *his preferences*, while in our model it is driven by his *endogenous* noisy perception of *Sender's signal*. This has three immediate consequences:

1. The stochastic choice rule in our setting is optimally chosen by Receiver to be of the shifted Logit form described in Proposition 1, while in the private information models, the implied stochastic choice rule can an arbitrary CDF. Thus, the noise structure in our model is more specific, but as a consequence it generates sharper predictions. For example, our Theorem 1 shows that optimal mechanisms correspond to monotone partitions of a very particular form, while the private information models easily generate more complicated partitions.

2. More importantly, in our model the stochastic choice rule is itself a function of Sender's persuasion strategy. In this sense, our model effectively generalizes the private-information models by making the distribution of private information endogenous to the mechanism. In Sections 5.2 and 5.3, we give two important examples where, surprisingly, the endogeneity of k does *not* matter in the following sense: if (F^*, k^*) is optimal in our model, then F^* is optimal in the private information model when $H(\cdot) := p(\cdot; k^*)$. In Sections 5.2, where preferences are state-independent, the activity constraint [A] is binding, but only affects feasibility of solutions. In Sections 5.3, where preferences are aligned and the prior is symmetric, the activity constraint does not bind. However, endogeneity of k alters comparative statics in the state-independent case (Proposition 5) and, in general, induces distortions that cannot be replicated in the private-information model for intermediate degrees of

(30) Much of Kolotilin (2017) relies on weaker assumptions, though his Propositions 2 and 3 (the main characterizations of optimal persuasion strategies) do require affine payoffs.

bias (or when preferences are aligned but the prior is not symmetric).[31]

3. Sender's signals should be interpreted differently. In the private-information models, Sender's messages are type-contingent direct recommendations and Receiver is obedient. That is, signals of the form "$x = \mathbb{E}[s|x]$" are equivalent to recommendation signals of the form "$x = $ *all types* $r \leqslant x$ *should act*," and all types of Receiver follow the recommendations. In our model, on the other hand, signals do *not* admit such a direct recommendation interpretation. It is easy to see why: an inattentive agent does not always "see" direct recommendations, and hence could not hope to always follow them. Indeed, although Receiver in our model has no exogenous private information and a binary action choice — so that the Revelation Principle approach would suggest a message space with cardinality at most two — we will see that optimal persuasion strategies typically involve sending *infinitely-many* signals.

5. The Optimal Mechanism

As noted in Section 4.3, our FOA is analogous to the procedure in standard moral hazard problems whereby one first finds the optimal way to implement a given effort level, and then optimizes over implementable effort levels. This suggests organizing the analysis around four questions:

1. What is the optimal way to implement a given activity level k? That is, what is the solution to the component problem [CP]?
2. Which k are implementable? That is, what is a characterization of the set $K(G)$?
3. Which $k \in K(G)$ is optimal for Sender?
4. When is Receiver's moral hazard problem binding (i.e., when is the multiplier $\eta \neq 0$?), and how does this distort the optimal persuasion strategy?

In Section 5.1, we answer questions 1 and 2 through Theorems 1 and 2. These results are general, in that they hold for all preferences within the affine class we consider. In Sections 5.2 and 5.3, we analyze in more detail the optimal mechanisms for two important special cases, state-independent and aligned preferences. In these more specialized settings, we address questions 3 and 4 and conduct comparative statics.[32] Our main economic applications are discussed along the way.

5.1. General Characterization

We begin by characterizing the solution to Sender's component problem [CP], for fixed activity parameter k. Sender's problem is always solved by simple persuasion strategies that only pool adjacent states;

(31) A characterization of these distortions and associated comparative statics are topics of ongoing work not reported in this draft.

(32) In general, both questions 3 and 4 are challenging. As is familiar from one-dimensional moral hazard models, the interim value function $V(k|G)$ in [RP] need not be quasi-concave in k, making optimization over k not entirely straightforward. Moreover, it is not a priori clear how to determine the sign of the multiplier η. Sender cares about the correlation between the state and Receiver's action, and, due to non-linearities in the RI problem, the first-order shift in Receiver's stochastic choice rule $p(\cdot; k)$ induced by changes in k changes this correlation in complicated ways.

moreover, the structure of the pooling and separating regions is tightly pinned down.

Definition 4. A distribution of posterior means $F \leqslant^{MPS} G$ is *monotone partitional* if there exists a finite set of points $\{x_i\}_{i=0}^n$ such that $\underline{s} = x_0 < x_1 < \cdots < x_n = \overline{s}$ and for each interval $[x_i, x_{i+1}]$ one of the following holds:

(i) *Pooling:* F puts the full mass $G(x_{i+1}) - G(x_i)$ on $\mathbb{E}_G[s|s \in [x_i, x_{i+1}]]$.
(ii) *Separation:* $F = G$.

Moreover, such an F is said to have n *regions*. F is said to be *maximally informative* if $n = 1$ and $F = G$. F is said to *minimally informative* if $n = 1$ and F puts full weight on μ. F is said to be *extremal* if it either maximally or minimally informative.

Theorem 1. *Assume that G is continuous and has full support. The solution F^* to Sender's component problem* [CP] *is essentially unique and is an n-region monotone partition with $n \leqslant 3$. In particular, assuming that F^* is not extremal:*

1. *If Sender has state-independent preferences ($\alpha > 0$ and $\beta = 0$), then F^* has two regions with separation on the bottom region and pooling in the top region.*
2. *If preferences are strictly co-monotone ($\beta > 0$), then F^* has at most three regions with pooling in the bottom and top regions and separation on the middle region.*[33]

Proof. See Appendix B.1. □

The main argument proceeds in two steps. First, we write Sender's component problem [CP] in Lagrangian form

$$\mathscr{L} = \mathbb{E}_F[(\alpha + \beta \cdot x) \cdot p(x;k)] + \eta \cdot \mathbb{E}_F[p(x;k)]$$

where $\eta \in \mathbb{R}$ is a multiplier on the activity constraint [A]. Importantly, the addition of a multiplier simply changes Sender's preferences to $\tilde{v}(s) := \tilde{\alpha} + \beta \cdot s$ where $\tilde{\alpha} := \alpha + \eta$, preserving the affine structure and co-monotonicity properties. Thus, for a fixed value of the multiplier, the Sender's Lagrangian problem is a standard linear persuasion problem. Second, we characterize the optimal signal structure for this unconstrained persuasion problem. To do so, we first characterize the curvature properties of the function $\tilde{v}(x) \cdot p(x;k)$ and show that, due to the shifted Logit form of Receiver's stochastic choice rule, this function is always convex-concave-convex when $\beta > 0$ and is concave-convex when $\beta = 0$. This allows us to guess that the optimal mechanism will feature pooling on the concave regions and separation on the convex regions, and we verify this guess using an appropriate version of complementary slackness for our linear program.[34]

Theorem 1 substantially simplifies the search for a solution to Sender's relaxed problem [RP]. Namely, the fact that $n \leqslant 3$ means we need only optimize over (at most) three scalar variables: the

(33) This encompasses the case where the top or bottom region may be empty, so that the monotone partition is "separating-pooling" or "pooling-separating."

(34) In particular, we invoke Theorem 1 of Dworczak and Martini (2018), which also allows us to give a straightforward proof of uniqueness. Other duality results, such as Proposition 3 of Kolotilin (2017), could also be applied using essentially the same proof. Figures 4 and 6 illustrate the multipliers used to certify optimality in the proofs of Propositions 3 and 6, respectively.

activity parameter and the boundaries of the pooling/separating regions of the monotone partition. It also delivers an equivalent, and useful, way to interpret Sender's problem. In particular, because monotone partitions specify deterministic mappings from states to signals, we may view Sender as designing not information, but a *payoff schedule* $s \mapsto x(s)$ for Receiver. Suppose, for example, that Sender's strategy involves pooling on the interval $[a, b] \subseteq \mathcal{S}$. From Receiver's perspective, it is as if his payoff, as a function of s, is constant and equal to $x(s) = \mathbb{E}_G[s | a \leqslant s \leqslant b]$ on this interval. Thus, *even if* he were able to *learn directly about the state*, he would not be willing to pay the cost to distinguish among different states $s, s' \in [a, b]$. Formally, Sender's component problem [CP] is equivalent to

$$[\text{CP}_x] \qquad \sup_{x(\cdot)} \mathbb{E}_G\left[(\alpha + \beta \cdot s) \cdot p(x(s); k)\right]$$

$$[\text{A}_x] \qquad \text{s.t.} \qquad \mathbb{E}_G\left[p(x(s); k)\right] = P(k)$$

where we require that $x : \mathcal{S} \to \mathcal{S}$ and $x(S) \sim F \leqslant^{MPS} G$. This formulation will be useful in subsequent sections to understand where (on which regions of the state space) Sender wants to focus Receiver's attention.[35] To that end, for any feasible payoff schedule $x(\cdot)$, we define the *composite stochastic choice rule*

$$\begin{aligned} \hat{p}_x : \mathcal{S} \times \overline{\mathbb{R}}_+ &\to [0, 1] \\ (s, k) &\mapsto p(x(s); k) \end{aligned}$$

and, similarly, the *composite local attention intensity*

$$[\text{LAI}_x] \qquad \begin{aligned} \hat{\Delta}_x : \mathcal{S} \times \overline{\mathbb{R}}_+ &\to \overline{\mathbb{R}}_+ \\ (s, k) &\mapsto \Delta(x(s); k) \end{aligned}$$

Next, we turn to characterizing the set of implementable activity parameters. We require two new definitions, which will be explained after the statement of the characterization theorem.

Definition 5. Activity parameter k is *strictly implementable* at prior G if $k \in K^*(G)$, where

$$K^*(G) := \text{cl}\left(\left\{k \in \overline{\mathbb{R}}_+ : \ k \text{ is implemented by some non-degenerate } F \leqslant^{MPS} G\right\}\right)$$

and where, recalling Definition 2, any $F \neq \delta_0$ is non-degenerate.

Definition 6. The prior G is *non-trivial* (given λ) if $\mathbb{E}_G\left[e^{s/\lambda}\right] > 1$ and $\mathbb{E}_G\left[e^{-s/\lambda}\right] > 1$. Otherwise, G is said to be *trivial* (given λ). Let $\Lambda(G)$ denote the set of $\lambda \geqslant 0$ such that G is non-trivial given λ.

Theorem 2. *The prior G is trivial if and only if exactly one of the following is true: $K(G) = K^*(G) = \{0\}$ or $K(G) = K^*(G) = \{\infty\}$. For any non-trivial prior G, the sets of (strictly) implementable activity parameters, $K(G)$ and $K^*(G)$, are a compact[36] intervals with the following properties:*

1. *If $\mu < 0$, then $K(G) = K^*(G) = [0, \overline{k}]$ for some $\overline{k} \in \mathbb{R}_{++}$.*
2. *If $\mu > 0$, then $K(G) = K^*(G) = [\underline{k}, \infty]$ for some $\underline{k} \in \mathbb{R}_{++}$.*

(35) In Section 7.3, it will also be used to draw a connection to models of contracting with flexible information acquisition.
(36) When viewing $\overline{\mathbb{R}}_+$ as the one-point compactification of $\mathbb{R}_+$ with the associated topology.

3. *If $\mu = 0$, then $K^*(G) = \left[\underline{k}, \overline{k}\right] \subseteq K(G) = \overline{\mathbb{R}}_+$ for some $\underline{k} < \overline{k}$. In addition, $\overline{k} \in \mathbb{R}_{++}$ if and only if $\lim_{n\to\infty} \overline{k}_n < \infty$ for all sequences of priors G_n such that $\mu_n < 0$ and $G_n \to^{w^*} G$; similarly, $\underline{k} \in \mathbb{R}_{++}$ if and only if $\lim_{n\to\infty} \underline{k}_n > 0$ for all sequences of priors G_n such that $\mu_n > 0$ and $G_n \to^{w^*} G$.*
4. *The correspondence $\lambda \mapsto K^*(G)$ is continuous and strictly decreasing (in the set inclusion order) on $\Lambda(G)$.*

Proof. See Appendix B.2. □

[INCOMPLETE]

The main content of the theorem is in parts 1 and 2, namely, the nontrivial upper/lower bounds on $K(G)$ and its convexity. Part 3 is trivial, as when $\mu = 0$ the completely uninformative persuasion strategy ($F = \delta_\mu$) makes Receiver indifferent between actions and any mixing probability can be implemented. In applications, it is typically obvious whether extremal k are optimal, and the solution is uninteresting in those cases. Thus, we wish to rule out the trivial cases in which only extremal k are feasible; in light of Theorem 2, Assumption 2 below stands for the remainder of the paper.

Assumption 2. The prior, G, is nontrivial.

5.2. State-Independent Preferences

When preferences are state-independent, Sender merely wants to maximize the probability that Receiver acts. In addition to Assumptions 1 and 2, throughout Section 5.2 we assume:

Assumption 3. The prior G has strictly negative mean, $\mu < 0$.

If $\mu \geqslant 0$, it is clearly optimal to provide no information, in which case Receiver will act with probability one based on his prior information (see part (b) of Proposition 1). Together, Assumptions 2 and 3 ensure that persuasion is partially but imperfectly effective at getting Receiver to act, and that some information must be provided at the optimum, i.e., $F = \delta_\mu$ is suboptimal.

An easy observation that distinguishes our problem from its full attention analogue[37] is that, if it is possible to induce Receiver to act with positive probability, then Receiver must earn *strictly* positive rents under the optimal mechanism.

Lemma 2. *Suppose there exists some implementable $k > 0$, i.e., Receiver can be induced to act with positive probability. Then $F^*(0) < 1$. That is, with strictly positive probability under the optimal F^*, Receiver* strictly *prefers to act.*

Proof. If not, then $x \leqslant 0$, and thus $e^{x/\lambda} \leqslant 1$, F-a.s. Part (a) of Proposition 1 implies that $V(G) = 0$, a contradiction. □

(37) See, e.g., Kolotilin (2015).

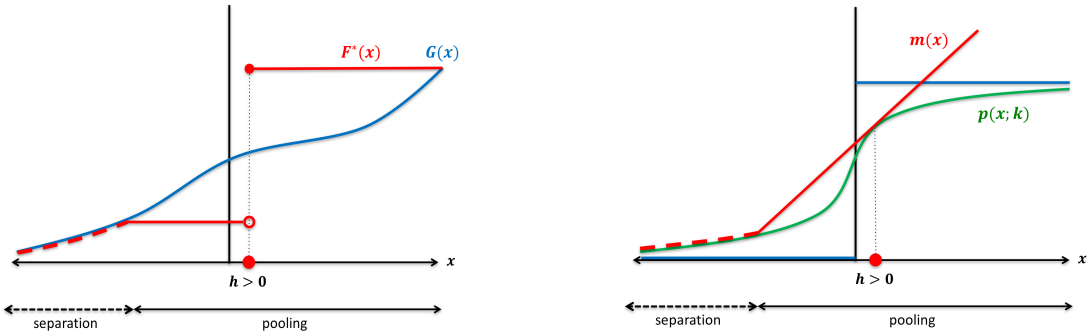

Figure 4: *The optimal F^* (left) and proof of its optimality (right) for state-independent preferences.*

This observation formalizes the intuition discussed in the context of the binary-state example in Section 2: if Receiver is willing at the interim stage (before signals are realized) to sink the cost of paying attention to a particular signal, he must anticipate that doing so is strictly beneficial ex post (after signals are realized). Thus, he must accrue some rents, unlike the full attention problem in which Receiver is typically made indifferent between multiple actions (see Proposition 5 of Kamenica and Gentzkow (2011)). We may refer to signals that make Receiver strictly prefer one action over the other as *convincing*.[38]

From Part 1 of Theorem 1, we know that the optimal F is either extremal or is a two-region monotone partition with separation at the bottom and pooling at the top — what Kolotilin et al. (2017) refer to as *upper censorship*. Since pooling regions correspond to atoms of F^*, it remains to characterize the location of this atom and the optimal activity parameter k.

Proposition 3. *Assume that G is continuous and has full support, preferences are state-independent, and Assumption 3 is satisfied. Then the optimal F^* is characterized by a single number $h \in (0, \overline{s}]$:*

1. *There is full separation on the interval $[\underline{s}, b)$, where $b \leqslant h$ is defined by $h = \mathbb{E}_G[s | s \geqslant b]$.*
2. *There is pooling on the interval $[b, \overline{s}]$.*

That is, the optimal F^* is given by

$$F^*(x) = \begin{cases} G(x), & \text{for } x \in [\underline{s}, b) \\ G(b), & \text{for } x \in [b, h) \\ 1, & \text{for } x \in [h, \overline{s}] \end{cases}$$

Proof. Immediate from Part 1 of Theorem 1 and Lemma 2. □

Note that the case of $b = h = \overline{s}$ corresponds to a fully informative mechanism. The optimal F^* is illustrated in the left-hand panel of Figure 4, where $h > 0$ denotes the position of the *upper atom*. Intuitively, one may think of the upper atom as being a *simple and convincing* signal. It is convincing in the sense described above; it is simple in the sense it corresponds to many states pooled together, so

(38) This insight that convincing signals are necessary to induce action, while simple, appears to be very robust, and we expect it to hold in much more general models of costly attention.

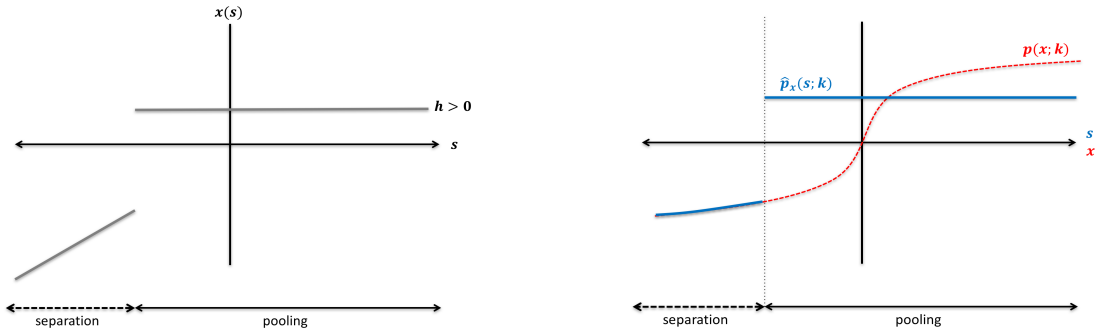

Figure 5: *Induced payoff schedule (left) and composite stochastic choice rule (right) for state-independent preferences.*

that Sender is effectively pre-processing the information that Receiver has access to. On the other hand, one may intuitively think of the separating region at the bottom as a collection of *complex* signals, as Sender provides Receiver with raw, unfiltered information about the state. Roughly speaking, the optimal mechanism involves sending (i) complex signals at states where Sender and Receiver's preferences over actions are opposed, and (ii) a single simple and convincing signal at an interval of states where their preferences are (more closely) aligned.

To understand why this must be the case, we must consider Receiver's stochastic choice rule $p(x;k)$, which is illustrated in the right-hand panel of Figure 4 along with the multiplier $m(\cdot)$ constructed in the proof of Theorem 1 that certifies optimality of F^*. Intuitively, when the state $s<0$ so that Sender and Receiver preferences are opposed, Sender wants to *exploit Receiver's inattention* — namely, the fact that he may mistakenly choose to act — and this makes Sender *risk-seeking*. This is a force towards full revelation. When the state $s>0$ so that Sender and Receiver preferences are aligned, Sender wants to *avoid Receiver mistakes* — namely, the fact that he may mistakenly choose not to act — and this makes Sender *risk-averse*. This is a force towards pooling. This intuition is, however, incomplete: Sender's risk-seeking and -aversion hinges on curvature properties of $p(x;k)$ which, in turn, are driven by the shape of the RI cost function.

Formulating Sender's problem in terms of the induced payoff schedule, illustrated in Figure 5, helps to understand where Sender wants Receiver's attention to be allocated. With a slight abuse of notation, by $p(s;k)$ we mean the composition of $p(x;k)$ with $x(s)$. The composite local attention intensity, $\hat{\Delta}_{x^*}(\cdot;k^*)$ is infinite at the boundary of the separation and pooling regions. This corresponds to the idea that Sender wants to provide strong incentives for Receiver to *focus his attention* on differentiating between low signals and the high atom. Put another way, by pre-filtering information, Sender is able to *relax Receiver's attention-smoothing motive*.

We now turn to the question of how the binding activity constraint leads to distortions in information disclosure. When preferences are state-independent, note that in Sender's component problem [CP], the objective function and left-hand side of the activity constraint [A] coincide. This observation suggests a simple procedure for finding the FOA-optimal pair (F^*,k^*). First, for fixed k, solve the *unconstrained* persuasion problem

$$\max_{F \leqslant^{MPS} G} \mathbb{E}_F\left[p(x;k)\right].$$

Denote the solution by F_k^*.[39] Second, check the sign of $S(k) := \mathbb{E}_{F_k^*}[p(x;k)] - P(k)$.[40] If $S(k) < 0$, Receiver will deviate downward to some $k' < k$ and, moreover, k cannot be implemented: by construction of F_k^*, the activity constraint [A] cannot be satisfied at k by any $F \leqslant^{MPS} G$. If $S(k) > 0$, then Receiver has an incentive to deviate upward to some $k' > k$. This is strictly better for Sender, as it implies that she can implement a higher activity level (perhaps, but not necessarily, in the interval $(k, k']$). If $S(k) = 0$, and assuming $k \in \mathbb{R}_{++}$, then (F_k^*, k) forms a *Nash equilibrium* of the persuasion/attention allocation game. That is, F_k^* is optimal for Sender taking k as given, and k is optimal for Receiver taking F_k^* as given. As this discussion suggests, the Nash equilibrium with the largest k corresponds to a FOA-optimal pair.

Proposition 4. *The (essentially unique) FOA-optimal pair (F^*, k^*) is a Nash equilibrium of the simultaneous-move game so that, in particular, $F^* \in \arg\max_{F \leqslant^{MPS} G} \mathbb{E}_F[p(x;k^*)]$. Moreover, $k^* = \max K(G) < \infty$.*

Proof. See Appendix B. □

In short, the Stackelberg and Nash solutions to Sender's problem coincide. The fact that (F^*, k^*) constitutes a Nash equilibrium means that the activity constraint [A] only affects the optimal persuasion strategy by restricting the set of implementable activity levels. Namely, it does *not* lead to distortions relative to the case in which Receiver's stochastic choice rule $p(\cdot; k^*)$ is taken as given. This was already demonstrated informally in the binary-state example from Section 2, and it relies critically on the form of [A], i.e., Receiver's first-order condition for optimal attention allocation.

We conclude the analysis with a discussion of comparative statics. Recall that the solution to the full-attention version of Sender's problem with state-independent preferences consists of (i) an atom at $h = 0$ and (ii) any amount of information disclosure on the interval $[\underline{s}, b]$ where b satisfies $\mathbb{E}_G[s | s \geqslant b] = 0$. In particular, pooling, separation, or anything in between is optimal on the lower region. The solution in which there is full separation on the low region is called the *most informative* solution to the full attention problem.[41]

Proposition 5. *Assume that G is continuous and has full support. As a function of the attention cost $\lambda > 0$:*

1. *The optimal persuasion strategy $F^*(\lambda)$ is increasing in the Blackwell order.*
2. *As $\lambda \to 0$, the high atom $h(\lambda) \to 0$ and $F^*(\lambda)$ converges weakly to the* most informative *solution to full attention problem.*
3. *Receiver's expected welfare is maximized, and strictly positive, at some $\lambda^* > 0$. His welfare is minimized as $\lambda \to 0$ or $\lambda \to \infty$.*
4. *Sender's expected welfare is decreasing in λ, and strictly so whenever $k_\lambda^* > 0$.*

Proof. See Appendix B.3. □

(39) Note that this is different than the FOA-optimal persuasion strategy $F^*(k)$ defined in Section 4.3.

(40) $S(k)$ stands for *s*lack in the activity constraint [A]. From part 3 of Proposition 1, this captures Receiver's incentive to deviate from k.

(41) See, e.g., Kolotilin (2015).

Part 1 is intuitive: as attention becomes more costly, Sender must provide more information to make paying attention worthwhile for Receiver. While many persuasion strategies are optimal under full attention, part 2 shows that small perturbations away from full attention uniquely select the most informative one.[42] These perturbations break Sender's indifference in favor of more complex signals which, as previously discussed, are effectively "gambles" that Receiver will make a mistake and choose to act. Finally, part 3 implies that Receiver may strictly prefer to be (moderately) attention constrained for strategic reasons. In essence, limited attention acts as a *commitment device* for Receiver to ignore insufficiently convincing signals, so he can extract more information from Sender.[43]

Part 4 shows that Sender always prefers to face a more attentive Receiver. It is obvious that Sender prefers a fully attentive Receiver ($\lambda = 0$) to an inattentive one ($\lambda > 0$), as when $\lambda = 0$ Sender can always replicate herself the additional garbling that the inattentive Receiver would have added. It is less obvious, and perhaps surprising, that this ranking extends to intermediate attention costs. Indeed, it is easy to see from the form of the F^* characterized in Proposition 3 and the subsequent discussion that (i) conditional on $s < 0$ and (ii) holding k fixed, Sender actually prefers *less* attentive (higher λ) Receivers, as they are more prone to making mistakes. But such Receivers also make more mistakes when $s > 0$, which hurts Sender. As the proof shows, even for fixed k, this latter effect outweighs Sender's ability to exploit Receiver's inattention at states $s < 0$. But there is a second, reinforcing, channel. As λ increases, Receiver pays less attention and relies more heavily on his prior; since $\mu < 0$ by Assumption 3, this means that k endogenously decreases.

Application: Advertising. **[TO BE ADDED]**

5.3. Aligned Preferences

When preferences are aligned, Sender's goal is to minimize losses from Receiver's inattention-induced mistakes.[44] If Receiver were fully attentive, many persuasion strategies would be optimal. Sender could simply provide a direct recommendation to Sender — i.e., tell him to act whenever $s \geqslant 0$ and not act whenever $s < 0$. This is clearly the minimal amount of information required to induce Receiver to act optimally in every state. Any more informative persuasion strategy, and in particular full disclosure, $F = G$, is also optimal under full attention.

Under limited attention, there are at least two natural conjectures regarding the optimal persuasion strategy. First, one might guess that the direct recommendation mechanism is uniquely optimal, relying on the intuition that direct recommendations are the "simplest" form of information disclosure, and so help Receiver economize on costly attention. Alternatively, it might be conjectured that full disclosure

(42) Of course, by continuity, all of the full attention optima will be approximately optimal when attention costs are small.

(43) A similar intuition arises in Aghion and Tirole (1997), where a manager may prefer to operate in a regime of information overload as a commitment device not to over-ride an agent's favored decision, thereby providing stronger incentivizes for information acquisition.

(44) Clearly, the full-attention outcome could be implemented if Sender were granted decision-making authority, leaving both parties strictly better off. The underlying assumption is that Receiver must maintain formal authority for reasons outside the model.

is uniquely optimal, relying on the intuition that more information gives rise to greater incentives to pay attention. Perhaps surprisingly, both of these natural conjectures are incorrect: in general, neither full disclosure nor a direct recommendation is optimal. However, the naive intuitions behind the conjectures are not misplaced, and the optimal persuasion strategy is constructed by appropriately combining them: it leverages the simplicity of direct recommendations by pooling together very low states into one low message l and pooling together very high states into one high message h, while fully disclosing intermediate states.

For simplicity, we will assume that G is symmetric around $x = 0$, i.e., $G(-x) = 1 - G(x)$ for all $x \in [0, \overline{x}]$ and $CH\,(\text{supp}(G)) = [-\overline{s}, \overline{s}]$. Such distributions will simply be called *symmetric*.[45] An FOA-optimal pair (F^*, k^*) is called *symmetric* if F^* is symmetric and $k^* = 1$ or, equivalently, the unconditional probability of acting is $P(k^*) = 1/2$.

Proposition 6. *Assume that G is continuous, has full support, and is symmetric. The optimal* symmetric *mechanism is characterized by one number, $h \in (0, \overline{s}]$:*

1. *There is full separation on the interval $(-b, b)$, where $b > 0$ is defined by $h = \mathbb{E}_G\,[s | s \geqslant b]$.*
2. *There pooling on the intervals $[\underline{s}, -b]$ and $[b, \overline{s}]$.*

That is, the optimal symmetric F^* is given by

$$F^*(x) = \begin{cases} 0, & \text{for } x \in [\underline{s}, -h) \\ G(-b), & \text{for } x \in [-h, -b] \\ G(x), & \text{for } x \in (-b, b) \\ G(b), & \text{for } x \in [b, h) \\ 1, & \text{for } x \in [h, \overline{s}] \end{cases}$$

Proof. See Appendix B.4. □

Remark 5.1. *Note that Proposition 6, as stated, characterizes the (unique) optimal mechanism within the class of symmetric mechanisms, but does not state that symmetric mechanisms are optimal. We are quite confident in the conjecture that the stated mechanism is in fact optimal — this is true in every numerical example we have computed — though we do not have a completed proof at the time of writing this draft. The difficulty lies in characterizing the optimal k, which is not straightforward, as Sender's objective function is not quasi-concave. With this caveat in mind, we will, with some loss of precision, refer to the F^* specified in the proposition merely as the optimal mechanism.*

Note that the case of $b = h = \overline{s}$ corresponds to the fully informative mechanism. (The case $b = 0$ corresponds to the direct recommendation mechanism, but the proposition states that this is *never* optimal.) The optimal F^* is depicted in the left-hand panel of Figure 6, where we denote the low atom by $l := -h < 0$.

(45) Note that symmetry implies that the prior mean $\mu = 0$ and, if G admits a density g, is equivalent to the condition that $g(-x) = g(x)$.

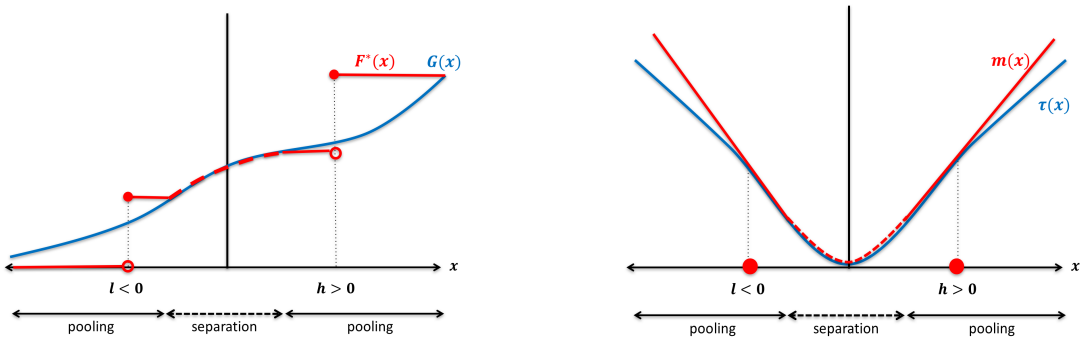

Figure 6: *The optimal F (left) and proof of its optimality (right) for aligned preferences when G is symmetric.*

As in Section 5.2, we may interpret the atoms as *simple and convincing* signals used to draw Receiver's attention. To provide additional intuition, it is useful to consider Sender's *unconstrained problem*, in which we fix $k = 1$ and drop the activity constraint [A] from [RP]. This is equivalent to solving

$$\max_{F \leqslant^{MPS} G} \mathbb{E}_F [\tau(x)] \tag{5.1}$$

where

$$\tau(x) := x \cdot p(x; 1) - \frac{x}{2} \tag{5.2}$$

is a mean-preserving transformation of Sender's objective function, and represents the gain from following $p(x; 1)$ over the strategy of choosing to act uniformly at random and independently of the signal.[46] This new objective is symmetric around $x = 0$, i.e., $\tau(-x) = \tau(x)$, which allows us to more easily illustrate the symmetry of the problem (see the right-hand panel of Figure Figure 6, which also displays the multiplier $m(\cdot)$, constructed in the proof, that certifies optimality of F^*).

Importantly, $\tau(\cdot)$ is strictly increasing in $|x|$, which corresponds to the fact that taking the correct action is increasingly important when the stakes are higher. If Sender were not restricted by the mean-preserving contraction constraint, this *level effect* would lead him to put as much weight as possible on extremal states. That is, Sender would want to *exaggerate* the state by claiming that $x = \overline{s}$ whenever $s \geqslant 0$ and that $x = \underline{s}$ whenever $s < 0$, as this maximizes the likelihood of catching Receiver's attention and minimizes mistakes. When the state is binary, this strategy is consistent with Bayes' rule.

Proposition 7. *Suppose that G has binary support.*[47] *Then the uniquely optimal persuasion strategy is full disclosure, i.e., $F^* = G$.*

Proof. See Appendix C. □

In essence, when the state is binary, there is no conflict in the agents' preferences between the *quantity* and *quality* of the information that Receiver acquires. When the state space is rich, as when

(46) Because $\mu = 0$ when G is symmetric, the expectation of the second term drops out.
(47) Note that the proposition does *not* require that G is symmetric; it is true for all binary-support priors.

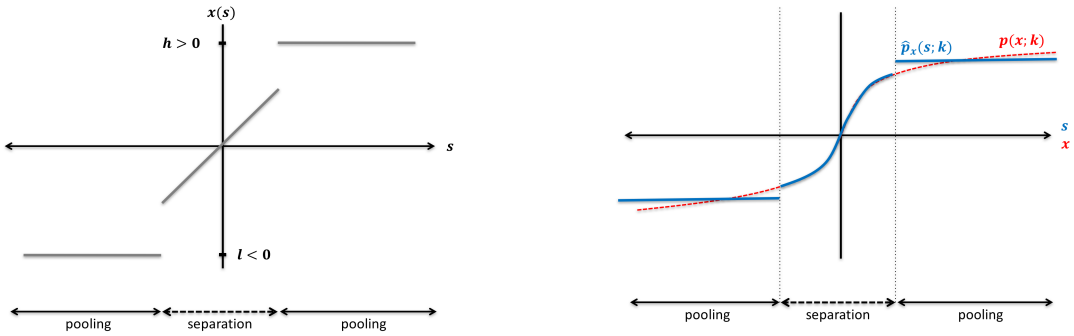

Figure 7: *Induced payoff schedule (left) and composite stochastic choice rule (right) for aligned preferences when G is symmetric.*

G is continuous, this kind of extreme exaggeration is not feasible, and the optimal mechanism needs to be more carefully tailored to take into account *marginal effects* on Receiver's attention allocation. In particular, the agents have different preferences over *where* Receiver should focus his attention or, equivalently, the *type* of information he acquires.

To understand the use of simple and convincing signals near the extremes of the state space, recall from Section 4.2 that Receiver has a strong motive to *smooth* his attention allocation over the state space. Because, in choosing whether to act, it ultimately matters only whether $s \geqslant 0$ or not, his local attention intensity [LAI] is greatest near the threshold $s = 0$ and decreases as $|s|$ increases.[48] That is, the value of information is highest for Receiver at the *pivotal* states $s \approx 0$. Sender, who does not internalize this attention smoothing motive, instead wants to minimize the likelihood of mistakes when they are most costly (when $|s|$ is large) which is precisely where Receiver, given full information, would pay the least attention. At the optimum, by providing simple and convincing signals near the extremes, relaxes the attention-smoothing motive: Receiver no longer distinguishes between nearby states when they both are very high or very low, but sharply distinguishes between *intermediate* and *extreme* states. Indeed, the composite local attention intensity$\hat{\Delta}_{x^*}(\cdot; 1)$ is infinite at $x = \pm b$, the upper and lower boundaries of the separation region.[49] This is illustrated in Figure 7, which shows the payoff schedule induced by the optimal F^* (left) and the induced composite stochastic choice rule (right).

Why does Sender not also pool on the middle region? As noted above, Receiver naturally pays most attention to these *pivotal* states. Thus, in this region the agents' preferences over information structures are aligned; more is unambiguously better and Sender need not try to manipulate Receiver's attention allocation.

To better understand the role of moral hazard in generating misaligned preferences over information structures, it is useful to compare Proposition 6 to two benchmarks. The first benchmark case is when Sender *fully* internalizes Receiver's attention costs, so it is as if Receiver is choosing his own

(48) This mechanically means that the choice rule $p(\cdot; 1)$ is steepest in a neighborhood of $s = 0$, and translates to convexity of the function $\tau(\cdot; 1)$ in this neighborhood.

(49) This intuition is in line with that of Szalay (2005) who shows that, in a delegation setting, restricting the agent to choose extreme actions is the optimal way to incentive costly information acquisition. However, the models are not directly comparable, as his agent has a continuum of possible actions and a quadratic loss function.

information. Formally, we say that (F^R, k^R) is *Receiver-optimal* if it solves

$$\max_{F \leqslant^{MPS} G,\ k \in \overline{\mathbb{R}}_+} \mathbb{E}_F \left[x \cdot p(x;k) \right] - \lambda I(X;A) \tag{5.3}$$

where $A|X \sim p(\cdot;k)$. The second benchmark case is when Receiver does not have an explicit attention cost, but is *capacity constrained* as in the original formulation of Sims (2003). For a given capacity constraint $C > 0$ and persuasion strategy $X \sim F$, Receiver chooses a stochastic choice rule $p^*_{F,C}(\cdot)$ such that

$$\begin{aligned} p^*_{F,C}(\cdot) \in \arg\max_{p(\cdot)} & \mathbb{E}_F \left[x p(x) \right] \\ \text{s.t.} \quad & I(X;A) \leqslant C \end{aligned} \tag{5.4}$$

where $A|X \sim p(\cdot)$.[50] Sender's problem may then be written

$$\max_{F \leqslant^{MPS} G} \mathbb{E}_F \left[x \cdot p^*_{F,C}(x) \right] \tag{5.5}$$

Let F^C denote a solution to Sender's problem when the capacity is C.

Proposition 8. *Let G be any prior. Then:*

1. *If (F^R, k^R) is Receiver-optimal and $p(\cdot;k^R)$ is not F^R-a.s. constant, then $F^R = G$.*
2. *If $I(X^{direct}; \mathbb{1}_{X \geqslant 0}) \leqslant C$, then $X \sim F^C$ is optimal if and only if $F^{direct} \leqslant^{MPS} F^C \leqslant^{MPS} G$ and $I(X; \mathbb{1}_{X \geqslant 0}) \leqslant C$.*
3. *If $I(X^{direct}; \mathbb{1}_{X \geqslant 0}) > C$ and there exists some $F \leqslant^{MPS} G$ such that $p^*_{F,C}(\cdot)$ is not F-a.s. constant, then the unique solution to Sender's problem is $F^C = G$.*
4. *Let F^* be FOA-optimal and A^* the induced solution to Receiver's problem. Assume A^* is not F^*-a.s. constant. Let $\hat{A}$ denote the solution to Receiver's problem under full information ($F = G$). Then, $I(X^*; A^*) \geqslant I(S; \hat{A})$ with a strict inequality if and only if it is not the case that $F(s) = G(s)$ G-a.s.*

Proof. See Appendix B.4. □

Intuitively, under full disclosure Receiver can ignore any unnecessary information, as in the RI model there is no cost to "sorting through" data to find the relevant pieces. Thus, partial disclosure merely restricts the feasible set of attention strategies and acts as an additional constraint, so an RI agent would never choose to limit his own information.

We close the analysis with a characterization of comparative statics:

Proposition 9. *Assume that G is continuous, has full support, and is symmetric. As a function of the attention cost $\lambda > 0$:*

(50) The choice rule $p^*_{F,C}(\cdot)$ takes the same Logit form familiar from Proposition 1, the only difference being that the "marginal cost" λ represents a Lagrange multiplier, the value of which depends on both F and C. Moreover, the solution is generically unique in the same sense as in the proposition.

1. *The optimal persuasion strategy F^*_λ is increasing in the Blackwell order.*
2. *There exists $\overline{\lambda}$ such that for $\lambda \geqslant \overline{\lambda}$, full disclosure is uniquely optimal.*
3. *As $\lambda \to 0$, F^*_λ converges weakly to the direct recommendation mechanism, $F^{direct} = \frac{1}{2}\delta_z + \frac{1}{2}\delta_{-z}$ where $z := \mathbb{E}_G[x|x \geqslant 0]$, which is the* least informative *solution to the full attention problem.*
4. *Both Sender's and Receiver's expected welfare is decreasing in λ.*

Proof. See Appendix B.4. □

Parts 1-3 speak to the naive intuitions in favor of direct recommendations and/or full disclosure given at the start of this section. Part 3 shows that, while many information structures are optimal under full attention, small perturbations away from full attention uniquely select the "simplest" one. But as paying attention must be more costly for Receiver, Sender must provide a greater value to doing so — i.e., disclose more information. Eventually, full disclosure becomes optimal. Finally, part 4 shows that costly attention hurts both parties under aligned preferences.

Application: Information Management in Organizations. The idea that organizational structure is largely determined by limited information processing and communication capabilities is an old and influential one (Marschak and Radner (1972), Garicano and Prat (2011)) that has arguably gained ever more relevance with the advent of new information technologies that, by making communication easier for Senders, can force Receivers into a regime of "information overload" (Edmunds and Morris (2000), Eppler and Mengis (2004)). The problem of information overload is particularly problematic for executives. A recent study by Bain & Company suggests the number of electronic communications that a typical executive receives each year has increased from around 1,000 in the 1970s to over 30,000 in the last decade,[51] while Bandiera et al. (2017) document that CEOs spend over 70% of their time with others and over 50% of their time in meetings. And these issues extend well beyond the private sector. As Simon (1971, p. 47) put it: "Attention is *generally* scarce in organizations, *particularly* scarce at the tops, and *desperately* scarce at the top of the organization called the United States government. There is only one President."

It is thus natural that subordinates (Sender) will actively filter the information they pass on to the executives (Receiver) for decision-making purposes; indeed, this is considered to be one of the primary and most important roles of the president's chief of staff.[52] Our model formalizes these ideas, and Proposition 6 sheds light on the optimal information-filtering policy. When the optimal decision is clear-cut, the subordinate should give a hard yes/no recommendation; when it is not, the subordinate should provide a detailed briefing and let the executive decide on his own.

Application: Disclosure Regulations. Many policies aimed at improving agents' decision-making rely on controlling the information they receive. One important and ubiquitous class of such policies

(51) And increased steadily each decade in between with the rise of voicemail, email, and, most recently, "virtual collaboration" technologies. See <https://hbr.org/2014/05/your-scarcest-resource>.

(52) See <https://www.politico.com/story/2017/08/24/john-kelly-trump-control-241967>. Of course, there are other reasons for this information filtering role aside from limited attention: ensuring that information is factual, limiting inefficient competition among advisors, etc.

consists of "mandatory disclosure" regulations, which require that businesses and service providers — e.g., credit card companies, insurance providers, medical professionals — provide detailed information about their goods and services to consumers. The main demand (consumer) side justification for these regulations is that additional information is always beneficial to expected-utility-maximizing consumers.[53] Ben-Shahar and Schneider (2011) and Ben-Shahar and Schneider (2014), among others in the legal literature, take issue with this view and give many reasons for why detailed disclosures may actually hurt consumers in practice, all of which (necessarily) rely on appeals to behavioral biases or bounded rationality. The theme of limited attention (what they call the "complexity" and "accumulation" problems) plays an especially central role in their arguments.

Our model helps to formalize aspects of this debate. In a world where consumers (Receiver) are idealized information processors well-described by the RI model, Proposition 6 indicates that mandated disclosure regulations are typically suboptimal if the policy-maker (Sender) does *not* internalize consumer's attention costs; Proposition 8 implies the converse statement. This elucidates four points. First, *even in* such an idealized world, there are coherent demand-side arguments against mandatory disclosure. Second, the optimal persuasion strategy characterized in Proposition 6 resembles suggestions to move away from full revelation toward "simple information" and "advice," i.e., direct recommendations (Ben-Shahar and Schneider (2011, pp.743-749)). In particular, it accords well with casual observation of, e.g., medical disclosures. When doctors assess that a surgery is either necessary or unacceptably risky for a patient, they give a hard yes/no recommendation; when they assess that it is a borderline case, they provide the patient with detailed information and let him make up his own mind. Third, given idealized information processing, the (sub)optimality of mandatory disclosure depends on the policy-maker's objectives, namely, whether she is concerned with consumer decision quality or takes a more wholistic view that includes consumer's attention costs. Fourth, any argument against mandated disclosure based on consumer's attention costs must rely on additional deficiencies in their information processing abilities.[54]

Application: Dual-Process Theories. When material preferences are aligned, it is natural to take a dual-process perspective in which Sender and Receiver represent different steps in some cognitive process (of a single agent) that may physically be housed in different parts of the brain. Optimality of the persuasion and attention strategies can be viewed as the outcome of evolutionary forces: Sender maximizes long-term fitness as captured by decision quality, but does not internalize the subsequent processing or choice-implementation costs borne by Receiver.[55] We provide two such interpretations that speak to ideas and evidence from neuroscience and cognitive psychology.

In the first interpretation, Sender and Receiver represent two stages in the cognitive implementation of the *choice* process. Sender is an "valuation center" in the brain that encodes payoff-relevant information and transmits it to Receiver, a "decision center," for implementation. Viewed in this light,

(53) There are of course important supply (firm) side rationales, but we abstract from them here.

(54) Namely, it requires a model in which consumers face a cost of "searching" for valuable information or "filtering out" irrelevant information. The RI model abstracts from both features, which is not to say they are unimportant in reality.

(55) Platt and Plassman (2013) and Johnson et al., 2013 emphasize this evolutionary viewpoint to constrained-optimal cognition and choice. That Sender doesn't internalize Receiver's attention costs is natural when they represent distinct regions of the brain, each of which expends energy firing neurons for a particular, specialized task.

Proposition 6 suggests that there are evolutionary gains for the "valuation center" to encode values using *coarse categorizations*. There is neuroeconomic support for such a two-step process involving communication between the centers. Platt and Plassman (2013) emphasize that different parts of the brain encode "predicted valuation signals" and "action valuation signals" that correspond, respectively, to Sender's signals and Receiver's perceptions in our model.[56] When the agent is called upon to make a decision (i.e., compare valuations of different alternatives and physically implement the solution), the "valuation center" transmits the "predicted valuation signals" to the "decision center," which then converts them to "action valuation signals" that are compared to make a final decision. There is evidence that this communication step takes place before the "decision center" becomes active, reinforcing the hierarchical relationship between Sender and Receiver in our model.[57] There is also evidence that construction of the "predicted valuation signals" does not require substantial cognitive resources (Platt and Plassman (2013, pp. 241-242)), which corresponds to the fact that Sender faces no information acquisition or communication costs.

In the second interpretation, Sender and Receiver represent two stages in the cognitive implementation of the *selective attention* process. Sender is a "perceptual gate" that filters out stimuli before they are completely perceived, and Receiver is a "processor" that more completely analyzes stimuli that have passed the gate (for, e.g., later use in a decision problem). This formalizes ideas from the cognitive psychology literature on attention which, since Broadbent (1958), has been at least implicitly based on information-theoretic concepts.[58] The classical dichotomy is between theories of *early selection* and *late selection*: both posit that inattention is the product of a binding capacity constraint, but differ in their views about *when* in the perceptual process the capacity constraint binds. Early selection theories posit that the main bottleneck is at the stage of "identifying" stimuli, and that there is a "perceptual gate" that prevents certain stimuli from proceeding to the identification and processing stages (Pashler (1998, p. 14)). Late selection theories, on the other hand, posit that all stimuli are identified without cost (so there is no perceptual gate), and that the main bottleneck is at the subsequent stage of "processing" stimuli (Pashler (1998, p. 17)).

The standard RI model is essentially a theory of late selection: an RI agent has limited capacity to process information for use in decision problems, but is assumed to have access to all conceivable information about the state (i.e., is assumed to have identified all available stimuli at the time of choice). Our model can be viewed as a hybrid theory containing the essential elements of both the early- and late-selection accounts: Sender acts as a "perceptual gate" that filters out stimuli at no cost, after which Receiver undertakes costly "selective processing" of (only) those stimuli that pass through the gate. As we have shown in Proposition 6, this kind of early selection is typically necessary to optimize

(56) As Platt and Plassman (2013, p. 249) emphasizes, these valuation signals are computed in a "common currency" that "is independent of the modality of the desired outcome," which corresponds to the fact that Sender's signals and Receiver's perceptions are based entirely on expected payoffs and not other, irrelevant details of the state space.

(57) This is true for decision problems of the type we consider here, while for pure perceptual tasks, the evidence suggests that the communication and choice steps take place concurrently and continuously. However, it is believed that the same brain regions are involved in both pure perceptual decisions and economic choices. See Glimcher (2013, p. 389).

(58) The modern line of psychology research on attention is commonly traced back to the "filter model" of Broadbent (1958) who was directly inspired by the notion of channel capacity developed in Shannon (1948). See Driver (2001) and Chapter 1 of Pashler (1998) for intellectual histories of this literature.

decision performance when there are constraints at the processing stage, even absent constraints at the identification stage. Moreover, we identify the optimal form of early selection: it is optimal to *filter out information about extreme states*, and simply identify them categorically as "high" or "low." This accords well with the argument of Pashler (1998, pp. 223-231) that such a hybrid model (which he dubs "controlled parallel processing") is empirically superior to accounts based on early- or late-selection alone.[59]

6. Cheap Talk Communication

To better understand the interaction between Sender's commitment power and Receiver's limited attention, in this section we extend the model of Section 3.1 to the case in which Sender communicates via cheap talk messages. The main message is that, absent commitment power, Sender's incentive to "exaggerate" is unambiguously detrimental to effective communication.

The model primitives of the model are exactly as in Section 3, with one small change in terminology. To reflect Sender's lack of commitment, we refer to his chosen information structure as a *communication strategy* and refer to realizations as *messages*. In Appendix D, we define strategies and equilibria at the level of generality of Section 3.1 and discuss how they admit descriptions in terms of stochastic choice rules in a manner similar to the FOA reduction in Section 4. For brevity, here we take these reductions for granted, and define equilibria directly in those simpler terms. Throughout this section, Assumption 1 stands and we assume that G is continuous and has full support (meaning that $\text{supp}\,(G) = [\underline{s}, \overline{s}]$).

Definition 7. An *equilibrium* of the cheap talk game consists of a communication strategy $(\pi, \mathfrak{X})$ for Sender and a stochastic choice rule $p(\cdot; k^*)$ for Receiver such that:

1. $\text{supp}(\pi(s, \cdot)) \subseteq \arg\max_{x \in \mathfrak{X}} v(1, s) \cdot p(x; k^*)$ for every $s \in \mathcal{S}$;
2. Receiver's stochastic choice rule solves his RI problem given π, as described in Proposition 1.

Moreover, the tuple (F^*_π, k^*) is referred to as the *equilibrium pair*, where F^*_π is the distribution over posterior means induced by $(\pi, \mathfrak{X})$ via Bayes' rule.

The equilibrium definition corresponds to the usual notion of perfect Bayesian equilibrium, and the notion of an "equilibrium pair" is intended to parallel the notion of an "FOA-optimal pair" from the full-commitment problem (see Section 4.3). Part 1 is Sender optimality, given Receiver's strategy. Part 2 corresponds to Receiver optimality, and also builds in belief consistency (i.e., consistency with Bayes' rule wherever possible) through the characterization of Receiver's optimal solution in Proposition 1, which implicitly states that Receiver forms correct posterior beliefs at each on-path signal. Two points are noteworthy. First, Sender's problem in part 3 is much simpler than her full-commitment problem [CP], in large part because here Sender does not internalize Receiver's obedience/activity constraint.

(59) There is some neuroscientific evidence for such a two-step process. In the context of visual perception tasks, Russo, Martínez and Hillyard (2003) suggest that the first stage of perception does not vary with measures of attention while later stages, carried out in a different region of the brain, does.

Second, unlike the case of Sender commitment,[60] it is now *not* without loss of generality to assume that Receiver's stochastic choice rule takes the shifted Logit form globally. Assuming that the Logit form holds globally amounts to an equilibrium refinement but, fortunately, our equilibrium characterization does *not* require such restrictions on off-path behavior.[61] Thus, part 4 of the equilibrium definition only requires that the Logit form must hold on the support of F_π, as stated in part 2 of Proposition 1.

We say that an equilibrium is *monotone partitional* if F_π is monotone partitional (recall Definition 4). A monotone partitional equilibrium is *binary* if the partition has two elements, and it is *babbling* if the partition has one element. An equilibrium is *informative* if it is not babbling. Two equilibria are *equivalent* if they induce the same joint distributions over state-action pairs.

Proposition 10. *Assume that G is continuous and has full support. There always exists a babbling equilibrium, and every equilibrium is monotone partitional. Moreover:*

1. *Under full attention ($\lambda = 0$), every equilibrium is equivalent to either a binary or babbling equilibrium. Whenever an informative equilibrium exists, there exists an equivalent informative equilibrium with a continuum of messages.*
2. *Under limited attention ($\lambda > 0$), every equilibrium is either binary or babbling.*
3. *The partitions induced by binary equilibria, when they exist, are independent of $\lambda \geqslant 0$.*
4. *There exists $\overline{\lambda} > 0$ (which will depend on the preference parameters α, β) such that both agents' equilibrium utilities are strictly decreasing in λ on $[0, \overline{\lambda}]$ and constant on $(\overline{\lambda}, \infty)$. Moreover, the equilibrium set is constant on $(0, \overline{\lambda}]$ and consists only of a babbling equilibrium for $\lambda > \overline{\lambda}$.*

Proof. See Appendix D. □

The full-attention characterization in part 1, as well as the claims that all equilibria are monotone partitional and that babbling is always an equilibrium, are all standard. Under full attention, informative equilibria are determined up to a *cutoff* $c \in [\underline{s}, \overline{s}]$ such that Sender strictly prefers that Receiver (doesn't) act whenever $s > c$ ($s < c$) and is indifferent to Receiver's action when $s = c$. Starting with any such equilibrium, it is without loss to pool all messages $m \in [c, \overline{s}]$ into a direct recommendation to act, and similarly to pool all messages $m \in [\underline{s}, c)$ into a direct recommendation to not act. Similarly, starting from any binary equilibrium (which must take such a direct recommendation form), it is always possible to find a nonempty interval $[d, \overline{s}]$ with $d > c$ such that Sender recommends no action on $[\underline{s}, c)$, recommends action on $[c, d)$, and fully reveals the state when it lies in $[d, \overline{s}]$. For example, full revelation is an equilibrium under aligned preferences, as is the direct recommendation communication strategy with cutoff $c = 0$.

Part 2 is based on the intuition, hinted at in Section 5.3, that Sender has a strict incentive to *exaggerate* the stakes of the decision problem to an inattentive Receiver. This is driven entirely by *level effects*, namely, the fact that Receiver's stochastic choice rule is *strictly* increasing in his posterior

(60) Recall footnote 4.2.

(61) Matějka and McKay (2012) implicitly assume such a refinement, while Ravid (2018) takes a more explicit game-theoretic approach based on trembling-hand perfection.

mean. This is in contrast to the full commitment solution, which was determined by *marginal effects*, namely, the curvature properties of the stochastic choice rule. To illustrate, consider the case of aligned preferences and suppose there exists an equilibrium with at least three messages $\{x_i\}_{i=1}^3$ such that $\hat{u}(x_3) > \hat{u}(x_2) \geqslant 0 > \hat{u}(x_1)$.[62] For any state $s > 0$, Sender aims to maximize the probability that Receiver acts. Thus, when choosing between transmitting x_2 or x_3, she will always strictly benefit from sending x_3 because $p(x_3; k^*) > p(x_2; k^*)$. Of course, under full attention it is payoff-irrelevant whether Sender fully discloses the state or merely gives an action recommendation. But, as we saw in Section 5.3, under limited attention it is typically strictly optimal for Sender to disclose some intermediate amount of information and, in particular, it is *always strictly suboptimal* to give a direct action recommendation. Thus, even when material preferences are fully aligned, Receiver's moral hazard problem creates enough preference divergence to make Sender's lack of commitment power a substantive barrier to effective communication.

Parts 3 and 4 of Proposition 10 explain how the equilibrium set changes with attention costs. In particular, part 3 states that the cutoffs characterizing binary equilibria depend only on Sender's preference parameters and *not* on Receiver's attention cost. It is fairly easy to see why: the cutoff c is determined entirely by Sender's indifference condition. Though the forms of the binary equilibria do not change with λ, as stated in part 4, they cease to exist when attention costs become high enough because it is not possible to induce Receiver to pay attention. Moreover, both parties' utilities are strictly decreasing in λ on the range where binary equilibria exist. Intuitively, this is because the informativeness of Sender's communication strategy does not change, but Receiver's responses become noisier.

7. Discussion and Extension

7.1. Revisiting Basic Assumptions

A number of assumptions were made only for expositional simplicity. First, the state space can be essentially arbitrary and, in particular, need not be unidimensional. As long as the range of Receiver's utility function is compact and Sender's utility from action is an affine transformation of Receiver's, we may simply re-defined the "state" as Receiver's state-contingent utility and let G be the induced distribution over these utility levels. Second, we have focused on priors G that are either continuous or have binary support, but this can easily be relaxed. Continuity is only used in the proof of Theorem 1 and its corollaries in establishing that the optimal mechanism is monotone partitional. With atoms, randomization at endpoints of the partition cells may be necessary to satisfy the mean-preserving contraction constraint; this merely complicates notation without adding insight.

On the other hand, the affine-payoff[63] and binary-action assumptions are necessary for our analysis in its present form. It is well understood that without affine payoffs, Lemma 1 need not hold, i.e., it is not without loss of generality to identify messages with posterior means. If we maintain the affine payoff structure but allow Receiver to take n actions, Sender's messages can be identified with n-dimensional

(62) Recall from Section 4.2 that $\hat{u}(x) := \mathbb{E}_\pi[s|x]$ denotes the posterior mean induced by message x via Bayesian updating given $(\pi, \mathcal{X})$.

(63) Namely, the fact that Sender's utility is an affine function of Receiver's.

vectors consisting of the conditional expected utilities from taking each one of the n actions. More formally, the analogue of Lemma 1 is that it is without loss to consider messages x of the form

$$x = \left(\mathbb{E}_{G,\pi}\left[u(s,1)|x\right], \ldots, \mathbb{E}_{G,\pi}\left[u(s,n)|x\right]\right)$$

where $u(s,i)$ is Receiver's utility from taking action i in state s. We emphasize that, under full attention, no more than n distinct signals would be needed (using the Revelation Principle approach), whereas we generally would need infinitely-many n-dimensional signals. Dworczak and Martini (2018) show that their verification techniques extend to such multi-dimensional persuasion problems and solve perhaps the simplest possible example, but this is difficult even assuming full attention. One special case, which are in the process of studying, is when Receiver's payoffs to different actions are all affine transformations of the same scalar state, i.e., $u(s,i) = \alpha_i + \beta_i \cdot s$, and Sender's payoff is an affine transformation of Receiver's, in which case Lemma 1 goes through as stated.

Though our analysis relies on these assumptions, it is straightforward to formulate our model for essentially arbitrary action spaces and utility functions using the belief-based approach (recall Section 4.1).[64] Posterior beliefs will always be sufficient statistics in Receiver's RI problem, just as posterior means were sufficient statistics in our setting, so identifying Sender's signals with the posterior beliefs they induce is always without loss of generality. While formulating the model at this level of generality is easy, solving it is not. The main difficulty lies in characterizing the solution to Receiver's RI problem. With three or more actions, the first-order conditions to the RI problem are no longer sufficient and, in general, second-order conditions must be incorporated directly into Sender's problem. Even ignoring second-order conditions, with n actions Sender's component problem requires n linear constraints (corresponding to first-order conditions) and her relaxed problem involves optimizing over an n-vector of activity parameters.

7.2. Generalized Model

[INCOMPLETE]

The sets of states of nature $\mathcal{S}$, Sender signals $\mathcal{X}$, and Receiver actions $\mathcal{A}$ are assumed to compact metric and endowed with the Borel σ-algbras. The state S has prior distribution $\mathbb{P}_S$. For any compact metric space $\mathcal{Y}$, let $\Delta(\mathcal{Y})$ denote the space of Borel probability measures on $\mathcal{Y}$, endowed with the weak* topology. As before, the gross utility functions for Sender and Receiver, respectively, are denoted $v(a,s)$ and $u(a,s)$. We assume that Sender's utility $v(\cdot)$ is upper semi-continuous, and that Receiver's utility $u(\cdot)$ is continuous. Define Receiver's indirect (gross) utility function $\hat{u} : (\mathcal{S}) \to \mathbb{R}$ by $\hat{u}(q) := \max_{a \in A} \mathbb{E}_{s\sim q}\left[u(a,s)\right]$, and let $A : (\mathcal{S}) \rightrightarrows \mathcal{A}$ denote the associated argmax correspondence. Sender's indirect utility $\hat{v} : (\mathcal{S}) \to \mathbb{R}$ is then defined by $\hat{v}(q) := \max_{a \in A(q)} \mathbb{E}_{s\sim q}\left[v(a,s)\right]$. Persuasion and attention strategies are defined via Markov kernels as before. Assumption 1 holds, i.e., Receiver's attention cost function is given by mutual information.

(64) We need only impose mild regularity conditions (e.g., state and action spaces are compact metric, utility functions are continuous) to guarantee existence of solutions.

Proposition 11. *There exists a solution $(\pi^*, \mathfrak{X}^*)$ to Sender's problem in which $\mathfrak{X}^* = \Delta(\mathcal{S})$, so that signals x are defined with the posterior distributions they induce. Moreover, given any persuasion strategy $(\pi, \mathfrak{X})$ in which $\mathfrak{X} = \Delta(\mathcal{S})$, there exists a direct recommendation attention strategy that solves Receiver's problem.*

Proposition 12. *Suppose that $\mathbb{P}_S$ has binary support. Then the uniquely optimal persuasion strategy is full disclosure.*

Proof. See Appendix C. □

7.3. Contracting with Rational Inattention

[INCOMPLETE]

The solution technique adopted here — namely, the combined use of the FOA and LP duality as a verification tool — may also be useful for analyzing other models of mechanism/information design with rational inattentive agents. While such extensions are beyond the scope of this paper, as a proof of concept we point out that this technique can be used to easily solve the model of Yang (2017) (henceforth, Yang); his solution method, meanwhile, appears very difficult to apply in our setting. Here we provide a high-level discussion, with details relegated to Appendix E.

Yang considers the following contracting problem (formulated in our notation). There are two agents, Seller and Buyer. Seller owns an asset that will generate random cashflow $S \sim G$ tomorrow, but she discounts the future at rate $\delta \in (0, 1)$. There is a Buyer who does not discount the future; thus, it is efficient for the Buyer to purchase rights to (part of) the future cashflow. Seller designs a *security*, which is a function $x : \mathbb{R}_+ \to \mathbb{R}_+$ such that $x(s) \in [0, s]$ for all $s \in \mathbb{R}_{++}$. If cashflow s realizes, the Buyer receives $x(s)$ and Seller receives the residual $s - x(s)$. The constraints on $x(\cdot)$ reflect limited liability of both parties. Seller also sets a *price* $q \in \mathbb{R}_{++}$. In the joint design problem, $x(\cdot)$ and q are chosen together. Seller has no information about s beyond the prior. Buyer, who perfectly observes $(x(\cdot), q)$, can flexibly learn directly about the state by observing a correlated message M at cost proportional to $I(S; M)$.[65] This learning is purely rent-seeking and generates adverse selection for the seller. The optimal security therefore seeks, in a sense, to minimize learning. His main result is that *debt* is the uniquely optimal security (Propositions 4 and 5 in Yang), and he argues that the flexibility afforded by the RI model is essential for this result.

If we allow his Seller to offer "stochastic securities," it is possible to formulate Yang's problem (for a fixed price q and activity parameter k) as

(65) The latest version of Yang allows for a slightly more general class of cost functions.

$$\max_{F \leqslant^{FOSD} H} \mathbb{E}_F \left[(\alpha + \beta x) \cdot p(x;k)\right] \tag{7.1}$$

$$\text{s.t.} \quad \text{supp}(F) \subseteq \text{CH}\left(\text{supp}(H)\right) \tag{7.2}$$

$$\mathbb{E}_F \left[p(x;k)\right] = \frac{k}{k+1} \tag{7.3}$$

where $\alpha > 0$ and $\beta < 0$ are determined by model primitives, and the CDF H is a simple transformation of the prior G. Modulo the support constraint [E.11] (which is automatically satisfied in our setting with the MPS constraint), this is *identical* to our Sender's component problem [CP] with a different type of stochastic dominance constraint. Using an appropriate complementary slackness condition for this LP, one can show that any optimal solution F^* takes the form

$$F^*(x) = \begin{cases} H(x), & \text{if } x < D \\ 1, & \text{if } x \geqslant D \end{cases}$$

for some appropriately chosen D. The induced "payoff schedule" is of the form $x(s) = \min\{s, D\}$. This is nothing but a debt security, and by construction solves the original security design problem. Moreover, the solution and the proof of optimality have a simple graphical representation analogous to our Figures 4 and 6.

7.4. Comparison to Lipnowski, Mathevet and Wei (2018)

We close our discussion with a comparison to the independent and contemporaneous work of Lipnowski, Mathevet and Wei (2018) (henceforth, LMW), which is the closest paper to ours. Both papers are motivated by similar questions and, as we describe below, their information design problem is formally a relaxation of ours. But there are important differences in modeling assumptions, analysis, and results, making the papers largely complementary both formally and conceptually.

Several differences are immediate. First, we allow for preference misalignment between Sender and Receiver, whereas LMW only study the case of aligned preferences. This allows us to study the interaction between Sender's persuasion and strategic attention manipulation motives, while they focus on only the latter. Second, all of their substantive results concern binary and ternary state spaces (with potentially rich action spaces), and their analysis relies on this low-dimensionality to directly characterize Receiver's posterior beliefs. Our main focus is on uncountable state spaces, where the space of posterior beliefs is infinite-dimensional and thus intractable. Instead, we assume a binary action space and directly characterize the induced stochastic choice rules. Finally, the functional forms of the attention cost functions are different. Most of their substantive results pertain to what they call the "quadratic model," which has a number of special properties — none of which are satisfied by any model with the RI cost function — that are essential for their analysis.[66]

(66) It substantially simplifies the description of Receiver's optimal attention strategies (Proposition 1 in LMW) and reduces Sender's preferences over persuasion strategies to preferences over the variance of Receiver's action distribution (Observation 1 in LMW). In our model, aside from the case of state-independent preferences (where only the mean of the action distribution matters), Sender's preferences depend on the entire joint state-action distribution.

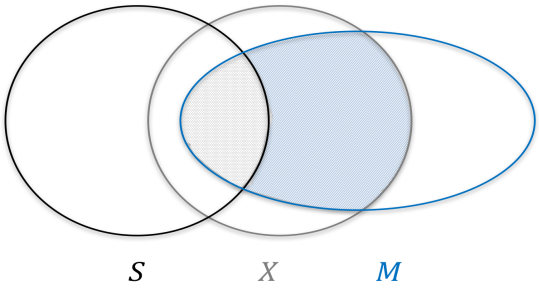

Figure 8: *Information costs in LMW (gray region) and here (gray plus blue regions).*

But the foremost differences pertain to the way information disclosure is conceptualized and attention costs are defined (beyond functional forms). For comparison, consider the special case of LMW with RI costs.[67] In our model, the cost function is $I(M;X)$, implying that Receiver learns about what Sender says and thus only indirectly about the state. LMW's cost function is $I(M;S)$, implying that Receiver learns directly about the state. This difference is consequential. In our model, Sender optimizes over *implementable* attention strategies, i.e., M for which there exists some $X \succsim_S^B S$ such that

$$\text{[BS]} \qquad M \in \arg\max_{M' \precsim_S^B X} U(M') - \lambda I(M';X)$$

where $\succsim_S^B$ denotes the Blackwell ordering over random variables, viewed as Blackwell experiments on S, and $U(M')$ denotes Receiver's indirect utility from attention strategy M'. In LMW, M is implementable if there exists some $X \precsim_S^B S$ such that

$$M \in \arg\max_{M' \precsim_S^B X} U(M') - \lambda I(M';S)$$

Theorem 1 in LMW shows that it is without loss to restrict attention to *incentive compatible* (IC) attention strategies, i.e., M such that

$$\text{[LMW]} \qquad M \in \arg\max_{M' \precsim_S^B M} U(M') - \lambda I(M';S)$$

and all of their subsequent results rely heavily on this fact. The idea is based on revelation-principle-type logic: Sender can "do the garbling for Receiver," so it is without loss to assume that Receiver "pays full attention." But this only works if Receiver's cost function doesn't depend on the signal structure; if our Sender tried to "do the garbling for Receiver" by setting $X = M$, Receiver would respond by further garbling M to some $M' \prec_S^B M$. Thus, in our model, Receiver will never pay full attention to what Sender says, and there is no useful analogue of IC attention strategies.[68]

(67) Their general model allows for any posterior separable cost function.

(68) Note that with binary actions and the RI cost function, part 1 of Proposition 1 implies that every IC attention strategy has binary support. In contrast, our results from Section 5 show that Sender typically needs to send uncountably-many signals. The closest analogue of IC attention strategies in our model can be gleaned from a generalized (and essentially tautological) version of the Revelation Principle: Sender "recommends" that Receiver induce a joint distribution over

On a conceptual level, this implies that the frameworks model different aspects of the problem. Our model captures a Receiver who does not perfectly attend to what Sender *says*, and is thus appropriate for understanding the role of inattention on *communication*. We find it difficult to interpret the LMW framework in this way: their Receiver *learns directly about the state*, while Sender simply places an *upper bound on what Receiver is allowed to learn*. Thus, LMW is a model of *delegated information acquisition*. In other words, our Sender first generates information and then conveys it to Receiver, while LMW's Receiver generates the information himself, subject to constraints chosen by Sender. For concreteness, consider the binary-state example involving the prosecutor and jury from Section 2. In our "communication" interpretation, the prosecutor collects evidence, signals represent what he says and puts on display in the courtroom, and the jury pays costly attention to what she says. In LMW's "delegation" interpretation, it is as if the prosecutor places restrictions on the investigation process (e.g., admissible types of evidence), and then the jury itself collects evidence subject to information acquisition costs.

On a formal level, it implies that LMW's information design problem is (often) a relaxation of ours. Of course, by restricting attention to IC policies, they focus entirely on Receiver's incentives to garble a target information structure, whereas in our model we need to fully solve Receiver's attention allocation problem given the "prior" induced by X. More importantly, LMW's Sender has a larger feasible set than our Sender, as described below.

Claim 7.1. *The following hold:*

1. *For any Markov chain $S \to X \to M$, we have*

$$I(X;M) = I(S;M) + I(X;M|S)$$

   Thus, $I(X;M) \geqslant I(S;M)$.
2. *Let X be a straightforward persuasion strategy and M a solution to Receiver's RI problem given X. Assume that M is not almost surely constant. Then, $I(X;M) = I(S;M)$ if and only if X is a deterministic function of S.*
3. *If M is implementable (by some X) in our model, then M is IC in LMW.*

Proof Sketch of Claim 7.1. Assume that all random variables have finite support, and let $\mathbb{P}(s,x,m)$ denote the joint PMF. (This is only for notational simplicity.) For part 1, the chain rule for mutual information implies

$$\begin{aligned} I(S,X;M) &= I(X;M) + I(S;M|X) \\ &= I(S;M) + I(X,M|S) \end{aligned}$$

states and actions (subject to some constraints), and Receiver is "obedient." More formally, any persuasion strategy $(\pi, \mathfrak{X})$ can be composed with the attention strategy $(\mu, \mathscr{M})$ that Receiver chooses as a best-response. This induces a stochastic choice rule $\sigma : \mathcal{S} \to \mathcal{A}$ defined by $\sigma(a|s) = \int_{\mathfrak{X}} \mu(dx,a)\pi(s,dx)$, which can itself be view as a direct recommendation. (By part 1 of Proposition 1, the set of implementable σ corresponds exactly to the set of implementable random posteriors.) But it is not clear that this approach is useful. Even in our simple setting, the set of implementable σ is typically not convex. Moreover, this approach does not directly reveal what Sender should actually *say* — the key question from the practical standpoint of indirect implementation.

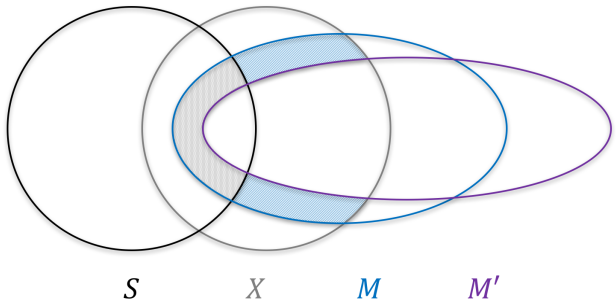

Figure 9: *Incentives to garble M to M'. Cost savings in LMW (gray region) smaller than here (gray plus blue regions).*

and $I(S;M|X)=0$ because $S \to X \to M$ form a Markov chain. For part 2, note that $I(X;M|S)=0$ if and only if $X \to S \to M$ forms a Markov chain. But $S \to X \to M$ forms a Markov chain by construction. Thus, $I(X;M|S)=0$ if and only if $\mathbb{P}(s)\mathbb{P}(x|s)\mathbb{P}(m|x) = \mathbb{P}(x)\mathbb{P}(s|x)\mathbb{P}(m|s)$ for all (s,x,m), which is equivalent to $\mathbb{P}(m|x)=\mathbb{P}(m|s)$ for all m and (s,x) such that $\mathbb{P}(x|s)>0$. But if M solves Receiver's RI problem and is not a.s. constant, this will be the case only if $\mathbb{E}[S|x] = \mathbb{E}[S|x']$ for all s and $x, x' \in \text{supp}(\mathbb{P}(\cdot|s))$. Because X is straightforward, this implies that it is a deterministic function of S. For part 3, define Receiver's indirect gross utility from direct attention strategy M by

$$U(M) := \int_{\mathcal{S}\times\mathcal{X}\times\mathcal{A}} u(m,s)\mu(x,dm)\pi(s,dx)dG(s)$$

If M solves Receiver's RI problem given X, then it must be that

[7.4] $$M \in \arg\max_{M' \succsim^B M} U(M) - \lambda I(X;M)$$

and, by definition, M is LMW-IC if and only if

[7.5] $$M \in \arg\max_{M' \succsim^B M} U(M) - \lambda I(S;M)$$

Let $M \succsim^B M'$. Define $\Delta(M,M') := [U(M) - \lambda I(S;M)] - [U(M') - \lambda I(S;M')]$. Then

$$\begin{aligned}\Delta(M,M') &= [U(M) - \lambda I(X;M)] - \left[U(M') - \lambda I(X;M')\right] + \lambda\left[I(X;M|S) - I(X;M'|S)\right] \\ &\geqslant [U(M) - \lambda I(X;M)] - \left[U(M') - \lambda I(X;M')\right] \\ &\geqslant 0\end{aligned}$$

where the first line follows from part 1 and the third line follows from [7.4]. The second line follows from the fact that, conditional on S, $X \to M \to M'$ forms a Markov chain (by the definition of the Blackwell order) and the data processing inequality. Thus, M satisfies [7.5]. □

The intuition for part 1 of 7.1 is that additional randomness in the signal makes it harder for an inattentive agent to "track" X than S, and $I(X;M|S)$ quantifies the cost of paying attention to this additional noise. The formula is derived from the "chain rule" for mutual information. This is represented in the Venn diagram Figure 8, standard in information theory, where sets represent random variables and set intersection represents mutual information. (That M only intersects S in $X \cap S$ represents

that $S \to X \to M$.) Similarly, the intuition for part 3 is that the cost savings from garbling M to M' in our model are always larger than in LMW, since our Receiver also saves the additional difference $I(M; X|S) - I(M'; X|S)$. This is represented in the Venn diagram in Figure 9.

An immediate corollary of part 3 is that, for the RI cost function, LMW's Sender has a weakly larger feasible set than our Sender. Indeed, this argument goes through for any "uniformly posterior separable" cost function (Caplin, Dean and Leahy (2018)), as this is exactly the class of attention cost functions for which the chain rule holds (see Theorem 5 of Zhong (2018)). Part 2 of Claim 7.1 makes a stronger statement: unless M can be implemented (in our model) by a "deterministic" X, then it will be strictly costlier for Receiver to pay attention in our model. Whenever the state space is finite, the optimal X will typically not be deterministic; this implies that Receiver will often have a *strict* incentive to garble M in our model whenever, in LMW, *(i)* M is IC and *(ii)* indifferent between paying full attention and garbling to some M'. Thus, at least in some cases, our Sender's feasible set is *strictly* smaller.

It is important to note, however, that the "quadratic model" in LMW has an attention cost function that is *not* uniformly posterior separable (it is merely "posterior separable"), so that many of their results are not comparable to ours. However, we may immediately deduce the following:

Corollary 7.1. *Consider the general version of our model, as described in Section 7.2. Assume that the prior has binary support. Then full disclosure is the uniquely optimal persuasion strategy.*

Proof. Optimality of full disclosure is immediate from part 2 of Claim 7.1 and Theorem 2 in Lipnowski, Mathevet and Wei (2018). **Need to add uniqueness.]** □

References

Aghion, Philippe and Jean Tirole (1997). 'Formal and real authority in organizations'. In: *Journal of Political Economy* 105.1, pp. 1–29 (cit. on p. 29).

Akyol, Emrah, Cedric Langbort and Tamer Basar (2017). 'Information-theoretic approach to strategic communication as a hierarchical game'. In: *Proceedings of the IEEE* 105.2, pp. 205–218 (cit. on p. 5).

Bandiera, Oriana et al. (2017). *CEO Behavior and Firm Performance*. Tech. rep. (cit. on p. 34).

Ben-Shahar, Omri and Carl E. Schneider (2011). 'The Failure of Mandated Disclosure'. In: *University of Pennsylvania Law Review* 159.3, pp. 647–749 (cit. on p. 35).

– (2014). *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton University Press (cit. on p. 35).

Bergemann, Dirk and Stephen Morris (2017). *Information Design: A Unified Perspective*. Tech. rep. Princeton University (cit. on p. 5).

Bergemann, Dirk and Juuso Välimäki (2007). 'Information in Mechanism Design'. In: *Proceedings of the 9th World Congress of the Econometric Society*. Cambridge University Press, pp. 186–221 (cit. on p. 6).

Blackwell, David (1953). 'Equivalent Comparison of Experiments'. In: *Annals of Mathematical Statistics* 24, pp. 265–272 (cit. on p. 19).

Board, Oliver J., Andreas Blume and Kohei Kawamura (2007). 'Noisy talk'. In: *Theoretical Economics* 2.4, pp. 395–440 (cit. on p. 7).
Boleslavsky, Raphael and Kyungmin Kim (2018). *Bayesian Persuasion and Moral Hazard*. Tech. rep. University of Miami (cit. on p. 5).
Broadbent, Donald E. (1958). *Perception and Communication*. Pergamon Press (cit. on p. 36).
Calvó-Armengol, Antoni, Joan De Martí and Andrea Prat (2015). 'Communication and influence'. In: *Theoretical Economics* 10.649-690 (cit. on p. 7).
Caplin, Andrew, Mark Dean and John Leahy (2018). *Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy*. Tech. rep. Columbia University (cit. on pp. 5, 46).
Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. 2nd edition. Wiley-Interscience (cit. on pp. 14, 51, 66, 67).
Cremer, Jacques, Luis Garicano and Andrea Prat (2007). 'Language and the Theory of the Firm'. In: *Quarterly Journal of Economics* 122.1, pp. 373–407 (cit. on p. 7).
Dessein, Wouter, Andrea Galeotti and Tano Santos (2016). 'Rational Inattention and Organizational Focus'. In: *American Economic Review* 106.6, pp. 1522–1536 (cit. on p. 7).
Dessein, Wouter and Andrea Prat (2016). 'The Oxford Handbook of the Economics of Networks'. In: ed. by Yann Bramoullé, Andrea Galeotti and Brian Rogers. Oxford University Press. Chap. Attention Networks (cit. on p. 7).
Dewatripont, Mathias (2006). '"Presidential Address" Costly Communication and Incentives'. In: *Journal of the European Economic Association* 4, pp. 253–268 (cit. on p. 7).
Dewatripont, Mathias and Jean Tirole (2005). 'Modes of Communication'. In: *Journal of Political Economy* 113.6, pp. 1217–1238 (cit. on p. 7).
Dilmé, Francesc (2017). *Optimal Languages*. Tech. rep. University of Bonn (cit. on p. 7).
Driver, Jon (2001). 'A selective review of selective attention research from the past century'. In: *British Journal of Psychology* 92, pp. 53–78 (cit. on p. 36).
Dworczak, Piotr and Giorgio Martini (2018). 'The Simple Economics of Optimal Persuasion'. In: *Journal of Political Economy* forthcoming (cit. on pp. 3, 5, 23, 40, 52, 53).
Edmunds, Angela and Anne Morris (2000). 'The problem of information overload in business organisations: a review of the literature'. In: *International Journal of Informaton Management* 20, pp. 17–28 (cit. on p. 34).
Eppler, Martin J. and Jeanne Mengis (2004). 'The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines'. In: *The Information Society* 20, pp. 325–344 (cit. on p. 34).
Gallager, Robert G. (1968). *Information Theory and Reliable Communication*. Vol. 2. Wiley (cit. on p. 6).
Garicano, Luis and Andrea Prat (2011). *Organizational Economics with Cognitive Costs*. Tech. rep. (cit. on pp. 7, 34).
Gentzkow, Matthew and Emir Kamenica (2014). 'Costly Persuasion'. In: *American Economic Review, Papers and Proceedings* 104.5, pp. 457–462 (cit. on p. 5).
Georgiadis, George and Balász Szentes (2018). *Optimal Monitoring Design*. Tech. rep. (cit. on p. 5).
Glazer, Jacob and Ariel Rubinstein (2004). 'On optimal rules of persuasion'. In: *Econometrica* 72.6, pp. 1715–1736 (cit. on p. 7).

Glimcher, Paul W. (2013). 'Value-Based Decision Making'. In: *Neuroeconomics*. Ed. by Paul W. Glimcher and Ernst Fehr. Academic Press. Chap. 20, pp. 373–391 (cit. on p. 36).

Gossner, Olivier and Jakub Steiner (2017). *On the Cost of Misperception: General Results and Related Behavioural Biases*. Tech. rep. (cit. on p. 7).

Gray, Robert M. (1988). *Probability, Random Processes, and Ergodic Properties*. Springer-Verlag (cit. on p. 51).

– (2011). *Entropy and Information Theory*. Second. Springer (cit. on pp. 6, 51).

Guo, Yingni and Eran Shmaya (2017). *The Interval Structure of Optimal Disclosure*. Tech. rep. Northwestern University (cit. on p. 5).

Hébert, Benjamin and Michael Woodford (2017). *Rational Inattention with Sequential Information Sampling*. Tech. rep. Stanford University GSB (cit. on p. 15).

Hernández, Penélope and Bernhard von Stengel (2014). 'Nash Codes for Noisy Channels'. In: *Operations Research* 62.6, pp. 1221–1235 (cit. on p. 7).

Johnson, Dominic D. et al. (2013). 'The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases'. In: *Trends in Ecology and Evolution* 28.8, pp. 474–481 (cit. on p. 35).

Kamenica, Emir (2017). 'Information Economics'. In: *Journal of Political Economy* 125.6, pp. 1885–1890 (cit. on p. 7).

Kamenica, Emir and Matthew Gentzkow (2011). 'Bayesian Persuasion'. In: *American Economic Review* 101, pp. 2590–2615 (cit. on pp. 5, 7, 26).

Kolotilin, Anton (2015). 'Experimental Design to Persuade'. In: *Games and Economic Behavior* 90, pp. 215–226 (cit. on pp. 25, 28, 60).

– (2017). 'Optimal Information Disclosure: A Linear Programming Approach'. In: *Theoretical Economics* forthcoming (cit. on pp. 5, 21, 23).

Kolotilin, Anton et al. (2017). 'Persuasion of a Privately Informed Receiver'. In: *Econometrica* 85.6, pp. 1949–1964 (cit. on pp. 5, 21, 26).

Li, Anqi and Ming Yang (2017). *Optimal Incentive Contract with Endogenous Monitoring Technology*. Tech. rep. (cit. on p. 5).

Lipnowski, Elliot, Laurent Mathevet and Dong Wei (2018). *Attention Management*. Tech. rep. (cit. on pp. 5, 42, 46).

Marschak, Jacob and Roy Radner (1972). *Economic Theory of Teams*. Yale University Press (cit. on pp. 7, 34).

Martin, Daniel (2017). 'Strategic pricing with rational inattention to quality'. In: *Games and Economic Behavior* 104, pp. 131–145 (cit. on p. 6).

Matějka, Filip (2015). 'Rigid Pricing and Rationally Inattentive Consumer'. In: *Journal of Economic Theory* 158, pp. 656–678 (cit. on p. 6).

Matějka, Filip and Alisdair McKay (2012). 'Simple market equilibria with rationally inattentive consumers'. In: *American Economic Review, Papers and Proceedings* 102.3, pp. 24–29 (cit. on pp. 6, 38, 68).

– (2015). 'Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model'. In: *American Economic Review* 105.1, pp. 272–298 (cit. on pp. 6, 17, 60, 64).

Matysková, Ludmila (2018). *Bayesian Persuasion with Costly Information Acquisition*. Tech. rep. CERGE-EI (cit. on p. 5).
Milgrom, Paul and Ilya Segal (2002). 'Envelope Theorems for Arbitrary Choice Sets'. In: *Econometrica* 70.2, pp. 583–601 (cit. on p. 61).
Morris, Stephen and Philipp Strack (2017). *The Wald Problem and the Equivalence of Sequential Sampling and Static Information Costs*. Tech. rep. Princeton University (cit. on p. 15).
Morris, Stephen and Ming Yang (2016). *Coordination and Continuous Choice*. Tech. rep. Princeton University (cit. on p. 18).
Myerson, Roger B. (1991). *Game Theory: Analysis of Conflict*. Harvard University Press (cit. on p. 7).
Pashler, Harold E. (1998). *The Psychology of Attention*. MIT Press (cit. on pp. 36, 37).
Persson, Petra (2018). 'Attention Manipulation and Information Overload'. In: *Behavioral Public Policy* 2.1, pp. 78–106 (cit. on p. 7).
Pirolli, Peter (2007). 'Cognitive Models of Human-Information Interaction'. In: *Handbook of Applied Cognition*. Ed. by Francis T. Durso. Wiley.
Platt, Michael L. and Hilke Plassman (2013). 'Multistage Valuation Signals and Common Neural Currencies'. In: *Neuroeconomics*. Ed. by Paul W. Glimcher and Ernst Fehr. Second. Academic Press. Chap. 13, pp. 237–258 (cit. on pp. 35, 36).
Ravid, Doron (2018). *Bargaining with Rational Inattention*. Tech. rep. University of Chicago (cit. on pp. 6, 38, 68).
Rayo, Luis and Ilya Segal (2010). 'Optimal Information Disclosure'. In: *Journal of Political Economy* 118.5, pp. 949–987 (cit. on p. 5).
Roesler, Anne-Katrin and Balász Szentes (2017). 'Buyer-Optimal Learning and Monopoly Pricing'. In: *American Economic Review* 107.7, pp. 2072–2080 (cit. on p. 5).
Russo, Francesco Di, Antigona Martínez and Steven A. Hillyard (2003). 'Source Analysis of Event-related Cortical Activity during Visuio-spacial Attention'. In: *Cerebral Cortex* 13.5, pp. 486–499 (cit. on p. 37).
Shannon, Claude E. (1948). 'A Mathematical Theory of Communication'. In: *The Bell System Technical Journal* 27, pp. 379–423 (cit. on p. 36).
Simon, Herbert A. (1971). 'Designing Organizations for an Information-Rich World'. In: *Computers, communications, and the public interest*. Ed. by Martin Greenberger, pp. 37–72 (cit. on pp. 2, 15, 34).
Sims, Christopher A. (1998). 'Stickiness'. In: *Carnegie-Rochester Conference Series on Public Policy*. Vol. 49, pp. 317–356 (cit. on p. 6).
– (2003). 'Implications of rational inattention'. In: *Journal of Monetary Economics* 50.3, pp. 665–690 (cit. on pp. 1, 3, 6, 15, 33).
Sobel, Joel (2010). 'Giving and receiving advice'. In: *Econometric Society 10th World Congress* (cit. on p. 7).
– (2012). 'Complexity versus conflict in communication'. In: *2012 46th Annual Conference on Information Sciences and Systems*. IEEE (cit. on p. 7).
– (2015). *Broad Terms and Organizational Codes*. Tech. rep. UCSD (cit. on p. 7).
Steiner, Jakub and Colin Stewart (2016). 'Perceiving Prospects Properly'. In: *American Economic Review* 106, pp. 1601–1631 (cit. on p. 7).

Szalay, Dezsö (2005). 'The Economics of Clear Advice and Extreme Options'. In: *Review of Economic Studies* 72, pp. 1173 –1198 (cit. on p. 32).

Yang, Ming (2017). *Optimality of Debt under Flexible Information Acquisition*. Tech. rep. Duke University (cit. on pp. 3, 6, 17, 18, 41, 51, 69).

Yang, Ming and Yao Zeng (2017). 'Financing Entreprenerial Production: Security Design with Flexible Information Acquisition'. In: *Review of Financial Studies* (cit. on p. 6).

Zhong, Weijie (2018). *Optimal Dynamic Information Acquisition*. Tech. rep. Columbia University (cit. on p. 46).

Appendices

[Several appendices are being actively updated.]

A. Proofs from Section 4

A.1. Proof Sketch of Proposition 1

Lemma A.1. *Fix a prior G and any persuasion strategy $(\pi, \mathfrak{X})$. Suppose that Receiver follows a direct recommendation attention strategy A. Then there exists a conditional probability system, denoted $p(\cdot)$, such that $\Pr(A = 1|X = x) = p(x)$. Moreover,*

$$\begin{aligned} I(X;A) &= H(A) - H(X|A) \\ &= h\left(\mathbb{E}_\pi\left[p(x)\right]\right) - \mathbb{E}_\pi\left[h\left(p(x)\right)\right] \end{aligned}$$

where $h : [0,1] \to \overline{\mathbb{R}}$ defined by $h(z) := -z\log(z) - (1-z)\log(1-z)$ is the "binary entropy function."

Proof of Lemma A.1. Under any direct recommendation attention strategy, A is a discrete random variable. Thus, a conditional probability system $p(\cdot)$ exists by, e.g., Corollary 5.8.1 of Gray (1988). The lemma then follows from Lemma 7.20 of Gray (2011). □

Existence follows from standard compactness and continuity arguments. Any optimal attention strategy must be a direct recommendation because $I(X;\cdot)$ is strictly increasing in the Blackwell order. (This is a consequence of the data processing inequality; see Theorem 2.8.1 of Cover and Thomas (2006).) Thus, in light of Lemma A.1, Proposition 1 follows exactly from the variational arguments used in the proof of Proposition 2 of Yang (2017). Here, we simply show why it is justified to refer to the condition in part 3 of the proposition as a first-order condition. Given the optimal shape of the stochastic choice rule $p(x;k)$ in part 2 of the proposition, Receiver's indirect utility is given by

$$V(k) := \mathbb{E}_\pi\left[\hat{u}(x)p(x;k) + \lambda h\left(p(x;k)\right)\right] - \lambda h\left(\mathbb{E}_\pi\left[p(x;k)\right]\right) \tag{A.1}$$

and his problem is to solve

$$\sup_{k \in \overline{\mathbb{R}}_+} V(k) \tag{A.2}$$

Note that a maximizer exists, as $V(\cdot)$ is continuous and $\overline{\mathbb{R}}_+$ is compact (when viewed as the one-point compactification of $\mathbb{R}_+$).

Lemma A.2. *Receiver's indirect utility* [A.1] *satisfies*

$$V'(k) =^{sign} \mathbb{E}_\pi\left[p(x;k)\right] - P(k) \tag{A.3}$$

Proof. Define $Q(k) := \mathbb{E}_\pi\left[p(x;k)\right]$. It is easy to verify that the derivative $Q'(\cdot)$ exists and is strictly positive on

$\overline{\mathbb{R}}_+$ (at endpoints, $Q'(\cdot)$ refers to the appropriate one-sided derivative). By direct computation, we have

$$\begin{aligned}
V'(k) &= \mathbb{E}_\pi\left[\left(\hat{u}(x) + \lambda h'(p(x;k))\right) \cdot \frac{\partial p(x;k)}{\partial k}\right] - \lambda h'\left(Q(k)\right) Q'(k) \\
&= \mathbb{E}_\pi\left[\left(\hat{u}(x) - \lambda \log\left(k\right) - \lambda \cdot \frac{\hat{u}(x)}{\lambda}\right) \cdot \frac{\partial p(x;k)}{\partial k}\right] - \lambda \log\left(\frac{1-Q(k)}{Q(k)}\right) Q'(k) \\
&= -\lambda \log(k) \cdot \mathbb{E}_\pi\left[\frac{\partial p(x;k)}{\partial k}\right] - \lambda \log\left(\frac{1-Q(k)}{Q(k)}\right) Q'(k) \\
&= \underbrace{\lambda Q'(k)}_{>0}\left[-\log(k) - \log\left(\frac{1-Q(k)}{Q(k)}\right)\right] \\
&=^{sign} \log\left(\frac{Q(k)}{1-Q(k)}\right) - \log(k) \\
&=^{sign} Q(k) - P(k)
\end{aligned}$$

as desired. □

A.2. Proofs from Section 4.3

Proof of Proposition 2. **[TO BE FILLED IN]** □

B. Proofs from Section 5

B.1. Proof of Theorem 1

Theorem 1 will follow from Propositions 13 and 14 below.

To keep notation self-contained, define the function

$$\phi(x) := (\alpha + \beta \cdot x) \cdot p(x;k)$$

for fixed parameters $\alpha, \beta \in \mathbb{R}$ and $k, \lambda \in \mathbb{R}_+$. (That is, if Sender's payoff from action is $v(x) = \alpha + \beta x$ and $\eta \in \mathbb{R}$ is the Lagrange multiplier on Receiver's *activity constraint*, then we will simply redefine $\alpha := \alpha + \eta$. This is done for purposes of notational economy during the proof only.) In general, for a fixed Lagrange multiplier η, Sender's component problem is equivalent to

$$\sup_{F \leqslant^{MPS} G} \mathbb{E}_F\left[\phi(x)\right] \tag{B.1}$$

where k, λ, β are determined by model primitives, and α may incorporate a Lagrange multiplier.

§ B.1.1. *LP Duality*

We will use the following verification result, which appears as Theorem 1 it Dworczak and Martini (2018). It is the appropriate version of complementary slackness for Sender's linear program [B.1].

Theorem B.1 (Dworczak and Martini (2018))**.** *If there exists a cumulative distribution function F and a convex function $m : \mathcal{S} \to \mathbb{R}$, with $m(x) \geqslant \phi(x)$ for all $x \in \mathcal{S}$, that satisfy*

[B.2] $$\textit{supp}(F) \subseteq \{x \in \mathcal{S} : m(x) = \phi(x)\}$$

[B.3] $$\mathbb{E}_G[m(x)] = \mathbb{E}_F[m(x)]$$

[B.4] $$F \leqslant^{MPS} G$$

then F is a solution to [B.1].

To establish uniqueness, we will also use the following result, which is appears as an intermediate result in Dworczak and Martini (2018).

Corollary B.1. *Let (F, m) be a primal-dual pair satisfying the conditions of Theorem B.1, so that F is optimal. Then, if $\hat{F}$ is any other solution to Sender's problem* [B.1]*, the primal-dual pair $(\hat{F}, m)$ also satisfies the conditions of Theorem B.1.*

§ B.1.2. *Preliminary Observations*

Properties of the optimal mechanism are thus determined by concavity/convexity properties of the function ϕ. Direct computation delivers that $\phi''(x) =^{sign} \xi(x)$ where the function $\xi(\cdot)$ defined by

$$\xi(x) := \alpha + \beta(x + 2\lambda) + ke^{x/\lambda} \cdot (\beta(2\lambda - x) - \alpha)$$

Lemma B.1. *The function $\xi(\cdot)$ has the following asymptotic behavior:*

$$\lim_{x\to-\infty} \xi(x) = \lim_{x\to+\infty} \xi(x) = -\infty \qquad \textit{if } \beta > 0$$
$$\lim_{x\to-\infty} \xi(x) = \lim_{x\to+\infty} \xi(x) = +\infty \qquad \textit{if } \beta < 0$$

Proof. Assume $\beta > 0$; the proof for $\beta < 0$ is identical up to appropriate sign changes. For the first limit, we have that $\lim_{x\to-\infty} e^{x/\lambda} = \lim_{x\to-\infty} xe^{x/\lambda} = 0$, so it follows immediately that $\lim_{x\to-\infty} \xi(x) = \lim_{x\to-\infty} \beta x = -\infty$. For the second limit, we have $\xi'(x) = \beta + \frac{k}{\lambda} e^{x/\lambda} \cdot (\beta\lambda - \beta x - \alpha)$. Thus, there exists an $\overline{x}$ such that $x \geqslant \overline{x}$ implies that $\xi'(x) \leqslant -1$. It follows immediately that $\lim_{x\to\infty} \xi(x) = -\infty$. □

Lemma B.2.

(i) If $\beta > 0$, the set $\{x : \xi(x) \geqslant 0\}$ is a compact interval and $\{x : \xi(x) > 0\}$ has nonempty interior. There exist $\underline{c}_+ < \overline{c}_+$ such that $\{x : \xi(x) = 0\} = \{\underline{c}_+, \overline{c}_+\}$.

(ii) If $\beta < 0$, the set $\{x : \xi(x) \leqslant 0\}$ is a compact interval and $\{x : \xi(x) < 0\}$ has nonempty interior. There exist $\underline{c}_- < \overline{c}_-$ such that $\{x : \xi(x) = 0\} = \{\underline{c}_-, \overline{c}_-\}$.

Proof. Assume $\beta > 0$; the proof for $\beta < 0$ is identical with appropriate sign changes. By direct computation, we see that

$$\xi'(x) \geqslant 0 \qquad \iff \qquad \beta\lambda \geqslant \psi(x) := ke^{x/\lambda} \cdot (\alpha + \beta x - \lambda\beta)$$

Observe that $\psi'(x) =^{sign} \alpha + \beta x$, so that $\psi(\cdot)$ has a unique critical point $x^* = -\alpha/\beta$, and $\psi(x^*) =^{sign} -\lambda\beta < 0$. Moreover, $\psi'(x) < 0$ for $x < x^*$ and $\psi'(x) > 0$ for $x > x^*$. It follows that x^* is also a global minimum. In addition, $\lim_{x\to-\infty} \psi(x) = 0$ and $\lim_{x\to+\infty} \psi(x) = +\infty$. Thus, there exists a unique $\hat{x} > x^*$ such that $\xi'(x) > 0$ for all $x < \hat{x}$ and $\xi'(x) < 0$ for all $x > \hat{x}$. That is, $\hat{x}$ is the unique critical point of $\xi(\cdot)$ and is a global maximizer.

This directly implies that $\{x : \xi(x) \geqslant 0\}$ is an interval. By Lemma B.1, it must be bounded, and it is closed by continuity of $\xi(\cdot)$.

It remains to show that $\{x : \xi(x) > 0\}$ has nonempty interior. Because $\xi(\cdot)$ is continuous, it is sufficient to show that $\xi(\hat{x}) > 0$. To see this, note that

$$\xi(x) = \lambda\xi'(x) + \alpha + \beta x + \lambda\beta k e^{x/\lambda}$$

Thus, $\xi(\hat{x}) > \alpha + \beta\hat{x}$. Because $\hat{x}$ is a critical point, we have $\beta\lambda = \psi(\hat{x})$, which implies that $\alpha + \beta\hat{x} - \lambda\beta > 0$. Putting this together, we get $\xi(\hat{x}) > \lambda\beta > 0$. □

Corollary B.2. *Thus, when $\beta > 0$ the function $\phi(\cdot)$ is concave-convex-concave, and when $\beta < 0$ it is convex-concave-convex.*

Lemma B.3.

(i) If $\beta > 0$, $\phi(\cdot)$ is strictly decreasing on $(-\infty, \underline{c}_+]$ and strictly increasing on $[\overline{c}_+, \infty)$.
(ii) If $\beta < 0$, $\phi(\cdot)$ is strictly increasing on $(-\infty, \underline{c}_-]$ and strictly decreasing on $[\overline{c}_-, \infty)$.

Proof. Assume $\beta > 0$; the proof for $\beta < 0$ is identical up to appropriate sign changes. By direct computation,

$$\phi'(x) =^{sign} \zeta(x) := \alpha + \beta(x + \lambda) + ke^{x/\lambda}(\lambda\beta)$$

and

$$\begin{aligned}\xi(x) &= \zeta(x) + \lambda\beta - ke^{x/\lambda}(\alpha + \beta x - \lambda\beta) \\ &= \zeta(x) + \lambda\beta - \psi(x)\end{aligned}$$

where $\psi(\cdot)$ was defined in the proof of Lemma B.2. Assume $x \leqslant \underline{c}_+$. Then, it follows from the proof of Lemma B.2 that $\xi'(x) > 0$, which is equivalent to $\lambda\beta - \psi(x) > 0$. Since $\xi(x) \leqslant 0$ by the assumption that $x \leqslant \underline{c}_+$, it follows that $\zeta(x) < 0$ and hence $\phi'(x) < 0$. Similarly, if $x \geqslant \overline{c}_+$ we have $\xi(x) \geqslant 0$ and $\lambda\beta - \psi(x) < 0$, from which it follows that $\zeta(x) > 0$ and $\phi'(x) > 0$. □

§ B.1.3. *Proof when* $\beta > 0$

For any $s \in \mathcal{S}$, define $l(s) := \mathbb{E}_G[x | x \leqslant s]$ and $h(s) := \mathbb{E}_G[x | x \geqslant s]$. Say that $\phi(\cdot)$ is *superdifferentiable at y on $\mathcal{S}$* if $\phi\big|_{\mathcal{S}}(\cdot)$, the restriction of $\phi(\cdot)$ to $\mathcal{S}$, is superdifferentiable at $y \in \mathcal{S}$. That is, there exists an affine function $\partial\phi : \mathcal{S} \to \mathbb{R}$ such that $\partial\phi(x) \geqslant \phi(x)$ for all $x \in \mathcal{X}$ and $\partial\phi(y) = \phi(y)$.

Proposition 13. *Assume that $\beta > 0$ and that G is continuous. Assume further that:*

1. $[\underline{s}, \overline{s}] \not\subseteq [\underline{c}_+, \overline{c}_+]$.
2. *$\phi(\cdot)$ is not superdifferentiable at μ on $\mathcal{S}$.*

Then the optimal persuasion strategy is is essentially unique[69] *and has a pooling-separation-pooling form. In particular, its CDF F satisfies:*

(i) If $\underline{c}_+ \leqslant \underline{s} < \overline{c}_+ < \overline{s}$, then there exists some $b \in [\underline{s}, \overline{c}_+]$ such that

$$F(x) = \begin{cases} G(x), & \text{for } x \in [\underline{s}, b) \\ G(b), & \text{for } x \in [b, h(b)) \\ 1, & \text{for } x \in [h(b), \overline{s}] \end{cases}$$

That is, there is full separation below b and pooling at $h(b)$ above b.

(69) That is, any two optimal solutions $\hat{F}$ and F^* must satisfy $\hat{F} = F^*$ G-a.e.

(ii) If $\underline{s} < \underline{c}_+ < \overline{s} \leqslant \overline{c}_+$, then there exists some $a \in [\underline{c}_+, \overline{s}]$ such that

$$F(x) = \begin{cases} 0, & \text{for } x \in [\underline{s}, l(a)) \\ G(a), & \text{for } x \in [l(a), a) \\ G(x), & \text{for } x \in [a, \overline{s}] \end{cases}$$

That is, there is pooling at $l(a)$ below a and full separation above a.

(iii) If $[\underline{c}_+, \overline{c}_+] \subseteq (\underline{s}, \overline{s})$, then there exist some $a, b \in [\underline{c}_+, \overline{c}_+]$ with $a < b$ such that

$$F(x) = \begin{cases} 0, & \text{for } x \in [\underline{s}, l(a)) \\ G(a), & \text{for } x \in [l(a), a) \\ G(x), & \text{for } x \in [a, b) \\ G(b), & \text{for } x \in [b, h(b)) \\ 1, & \text{for } x \in [h(b), \overline{s}] \end{cases}$$

That is, there is full separation on $[a, b]$, pooling at $l(a)$ below a, and pooling at $h(b)$ above b.

Proof. To prove optimality, we will construct Lagrange multipliers and verify the conjectured solutions. We begin by establishing a series of intermediate technical results, Claims B.1 – B.4. Once these are established, it will be evident how to construct the multipliers for all relevant parameter values, and we will carry out the verification step. The proof of uniqueness is deferred to the end.

Optimality: For each $\overline{z} \geqslant \overline{c}_+$ and $\underline{z} \leqslant \underline{c}_+$, define the function $m(\cdot|\underline{z}, \overline{z}) : \mathbb{R} \to \mathbb{R}$ by

$$m(x|\underline{z}, \overline{z}) := \begin{cases} \phi(a(\underline{z})) - \phi'(\underline{z})\,(a(\underline{z}) - x)\,, & \text{for } x < a(\underline{z}) \\ \phi(x), & \text{for } x \in [a(\underline{z}), b(\overline{z})] \\ \phi(b(\overline{z})) + \phi'(\overline{z})\,(x - b(\overline{z}))\,, & \text{for } x > b(\overline{z}) \end{cases}$$

where

$$a(\underline{z}) := \max\{y \geqslant \underline{z} : \phi(\underline{z}) = \phi(y) - \phi'(\underline{z})\,(y - \underline{z})\}$$
$$b(\overline{z}) := \min\{y \leqslant \overline{z} : \phi(\overline{z}) = \phi(y) + \phi'(\overline{z})\,(\overline{z} - y)\}$$

Claim B.1. *The function $a : (-\infty, \underline{c}_+] \to [\underline{c}_+, \infty)$ is well-defined, is strictly decreasing, and satisfies $a(\underline{c}_+) = \underline{c}_+$. Similarly, the function $b : [\overline{c}_+, \infty) \to (-\infty, \overline{c}_+]$ is well-defined, strictly decreasing, and satisfies $b(\overline{c}_+) = \overline{c}_+$.*

Proof of Claim. We prove the part of the claim concerning $a(\cdot)$; the proof of the part concerning $b(\cdot)$ is symmetric. Let $\underline{z} \leqslant \underline{c}_+$ be given. We claim that the equation

$$\phi(y) = h(y) := \phi(\underline{z}) + \phi'(\underline{z})\,(y - \underline{z})$$

has at least one, and most two, solution(s) on $[\underline{z}, \infty)$. Clearly $\underline{z}$ is always a solution. If $\underline{z} = \underline{c}_+$, then it is the only one. This follows from the fact that $\{\underline{c}_+\} = \arg\min_{y \in \mathbb{R}} \phi'(y)$, which is a consequence of part (i) of Lemma B.2 and part (i) of Lemma B.3, so that $\phi(y) > h(y)$ for all $y > \underline{z}$. If $\underline{z} < \underline{c}_+$, then there is exactly one other solution. Because $\phi(\cdot)$ is locally strictly concave at $\underline{z}$, there exists $\varepsilon > 0$ such that $\phi(y) < h(y)$ for $y \in (\underline{z}, \underline{c}_+ + \varepsilon)$. Moreover, $\phi(\cdot)$ is bounded below on $\mathbb{R}$ and $h(\cdot)$ is unbounded below on $[\underline{z}, \infty)$. Both functions are continuous, so we have $\phi(y^*) = h(y^*)$ for some $y^* > \underline{c}_+$ by the Intermediate Value Theorem. Moreover, it must be that $h'(y^*) < \phi'(y^*)$, and it follows that any such intersection must be unique.

The above argument establishes that $a(\cdot)$ is well-defined and that $a(\underline{c}_+) = \underline{c}_+$. To see that it is strictly decreasing, let $\underline{z}' < \underline{z} \leqslant \underline{c}_+$. Given y, define $j : \mathbb{R} \to \mathbb{R}$ by

$$j(x, y) := \phi(y) - \phi(x) - \phi'(x)(y - x)$$

Note that $j(\underline{z}, a(\underline{z})) = 0$ by construction and

$$\frac{\partial j(x, a(\underline{z}))}{\partial x} = -\phi''(x)(a(\underline{z}) - x) > 0 \qquad \text{for } x \leqslant \underline{z}$$

Thus, $j(\underline{z}', a(\underline{z})) < 0$. Moreover, there exists $b > 0$ such that

$$\frac{\partial j(\underline{z}, y)}{\partial y} = \phi'(y) - \phi'(\underline{z}') > b \qquad \text{for } y \geqslant a(\underline{z})$$

It follows that there exists some $y^* > a(\underline{z})$ such that $j(\underline{z}', y^*) = 0$. Thus $a(\underline{z}') > a(\underline{z})$, which proves that $a(\cdot)$ is strictly decreasing. □

Define $\mathfrak{Z} := \left\{(\underline{z}, \overline{z}) \in (-\infty, \underline{c}_+] \times [\overline{c}_+, \infty) : a(\underline{z}) \leqslant b(\overline{z})\right\}$. The following claim is an immediate corollary of the definition of $m(\cdot|\underline{z}, \overline{z})$ and Claim B.1.

Claim B.2. *For $(\underline{z}, \overline{z}) \in \mathfrak{Z}$, the function $m(\cdot|\underline{z}, \overline{z})$ is convex and lies pointwise above $\phi(\cdot)$. In particular,*

$$M(\underline{z}, \overline{z}) := \{x : m(x|\underline{z}, \overline{z}) = \phi(x)\} = \{\underline{z}, \overline{z}\} \cup [a(\underline{z}), b(\overline{z})]$$

Define $\overline{x} := \sup\{z : b(z) \geqslant \underline{s}\}$, $\underline{x} := \inf\{z : \overline{s} \geqslant a(z)\}$, $\overline{y} := \min\{\overline{s}, \overline{x}\}$, and $\underline{y} := \max\{\underline{s}, \underline{x}\}$.

Claim B.3.

(i) If $\overline{c}_+ < \overline{s}$, then there exists a unique $z^* \in (\overline{c}_+, \overline{y})$ such that $z^* = h((b(z^*))$.
(ii) If $\underline{c}_+ > \underline{s}$, there exists a unique $z_* \in (\underline{y}, \underline{c}_+)$ such that $z_* = l(a(z_*))$.

Proof of Claim. Consider part (i). By hypothesis, we have $\underline{s} < \overline{c}_+ < \overline{s}$. Then the monotonicity of $b(\cdot)$ established in Claim B.1 implies that $\overline{x} > \overline{c}_+$. Thus $(\overline{c}_+, \overline{y}) \neq \emptyset$. Define the function

$$\begin{aligned} \overline{\Delta} : [\overline{c}_+, \overline{y}] &\to \mathbb{R} \\ z &\mapsto z - h(b(z)) \end{aligned}$$

where, recall, $h(x) := \mathbb{E}_G[s|s \geqslant x]$. It is immediate from Claim B.1 and the full support assumption on G that $\overline{\Delta}(\overline{c}_+) = \overline{c}_+ - h(\overline{c}_+) < 0$. We claim that $\overline{\Delta}(\overline{y}) > 0$. If $\overline{s} = \overline{y}$, then $\overline{\Delta}(\overline{y}) = \overline{s} - h(b(\overline{s})) > 0$. If $\overline{x} = \overline{y}$, then $\overline{\Delta}(\overline{y}) = \overline{x} - h(\underline{s}) = \overline{x} - \mu$. Suppose this is not strictly positive, i.e., $\mu \geqslant \overline{x}$. Then $b(\mu) \leqslant \underline{s}$ and, by construction of the function $b(\cdot)$, $\phi(\cdot)$ is superdifferentiable on $\mathcal{S}$ at μ. This violates the hypotheses of the lemma, so it follows that $\overline{x} < \mu$. Thus, we have $\Delta(\overline{c}_+) < 0$ and $\Delta(\overline{y}) > 0$. As $b(\cdot)$ is strictly decreasing by Claim B.1, it is easy to see that $\Delta(\cdot)$ is strictly increasing. Moreover, it is continuous by the assumption that G is continuous. Existence of the desired z^* follows from the Intermediate Value Theorem and uniqueness follows from strict monotonicity.

The proof for part (ii) is entirely symmetric if we replace $\overline{\Delta}(\cdot)$ with the function

$$\begin{aligned} \underline{\Delta} : [\underline{x}, \underline{c}_+] &\to \mathbb{R} \\ z &\mapsto z - l(a(z)) \end{aligned}$$

where, recall, $l(x) := \mathbb{E}_G[s|s \leqslant x]$. The details are omitted. □

Claim B.4. *Define the functions*

$$\begin{aligned}\gamma : \mathcal{S} &\to \mathbb{R} \\ s &\mapsto \phi(h(s)) - \phi'(h(s))(h(s) - s)\end{aligned}$$

and

$$\begin{aligned}\delta : \mathcal{S} &\to \mathbb{R} \\ s &\mapsto \phi(l(s)) + \phi'(l(s))(s - l(s))\end{aligned}$$

If (z_*, z^*), defined in Claim B.3, satisfies $(z_*, z^*) \notin \mathfrak{Z}$, then there exists a unique $y^* \in [\underline{c}_+, \overline{c}_+]$ such that $\gamma(y^*) = \delta(y^*) > \phi(y^*)$. If $(z_*, z^*) \in \mathfrak{Z}$, then no such point exists.

Proof of Claim. We first show that there is at most one point of intersection. Observe that $\gamma(\cdot)$ is strictly increasing on $h^{-1}([\overline{c}_+, \overline{s}]) := \{s : h(s) \in [\overline{c}_+, \overline{s}]\}$. To see this, define the (smooth) function $\hat{\gamma}(x, y) := \phi(x) - \phi'(x)(x - y)$ so that $\gamma(s) = \hat{\gamma}(h(s), s)$. Let $s', s \in h^{-1}([\overline{c}_+, \overline{s}])$ with $s' > s$. By Taylor's Theorem,

$$\begin{aligned}\hat{\gamma}(h(s'), s') - \hat{\gamma}(h(s), s) &= \frac{\partial \hat{\gamma}(h(s), s)}{\partial x}(h(s') - h(s)) + \frac{\partial \hat{\gamma}(h(s), s)}{\partial y}(s' - s) + o(\varepsilon) \\ &= -\phi''(h(s))(h(s) - s)(h(s') - h(s)) + \phi'(h(s))(s' - s) + o(\varepsilon)\end{aligned}$$

where $\varepsilon := ||(h(s'), s) - (h(s), s)||$, the usual Euclidean norm on $\mathbb{R}^2$. The first term is non-negative as $h(\cdot)$ is weakly increasing and $h(s) \geqslant \overline{c}_+$. Moreover, we have $\phi'(h(s)) > 0$ by Lemma B.3 and $s' > s$ by assumption, so the second term is strictly positive. These terms are of order ε, so for sufficiently small $\varepsilon > 0$ (i.e., $|s' - s|$ sufficiently small, as $h(\cdot)$ is continuous) we have $\gamma(s') > \gamma(s)$. It follows that $\gamma(\cdot)$ is strictly increasing on this domain, as claimed. Using an analogous argument, we can show that $\delta(\cdot)$ is strictly decreasing on $l^{-1}([\underline{s}, \underline{c}_+]) := \{s : l(s) \in [\underline{s}, \underline{c}_+]\}$. It follows that the functions $\gamma(\cdot)$ and $\delta(\cdot)$ have at most one point of intersection.

We now establish (non)existence in the relevant cases.

Suppose $(z_*, z^*) \notin \mathfrak{Z}$. Thus, $\gamma(b(z^*)) < \delta(b(z^*))$ and $\gamma(a(z_*)) > \delta(a(z_*))$. By the Intermediate Value Theorem and strict monotonicity, there exists a unique $y^* \in (b(z^*), a(z_*))$ such that $\delta(y^*) = \gamma(y^*)$. As both functions are strictly monotone, $\phi(b(z^*)) = \gamma(b(z^*))$ and $\phi(a(z)*)) = \delta(a(z_*))$ by construction, and $\phi(\cdot)$ is convex on $(b(z^*), a(z_*))$, it follows that $\delta(y^*) = \gamma(y^*) > \phi(y^*)$. This gives the desired existence.

Suppose $(z_*, z^*) \in \mathfrak{Z}$. By the strict monotonicity of $\gamma(\cdot)$ and $\delta(\cdot)$ established above and properties established in Claims B.1 and B.4, we have $\gamma(s) < \phi(s)$ for all $s \in h^{-1}([\overline{c}_+, z^*)) \neq \emptyset$. Similarly, $\delta(s) < \phi(s)$ for all $s \in l^{-1}([z_*, \underline{c}_+]) \neq \emptyset$. Clearly $\gamma(\cdot)$ and $\delta(\cdot)$ cannot intersect at any $y \notin [a(z_*), b(z^*)]$ in this case, so it follows that any intersection occurs at some $y \in [a(z_*), b(z^*)]$ such that $\phi(y) \leqslant \delta(y) = \gamma(y)$ (with a strict inequality whenever $z_* < z^*$), establishing nonexistence of y^* in this case. □

We are now in a position to construct the Lagrange multipliers. For notational purposes, and with a slight abuse of notation, extend the functions $a(\cdot)$ and $b(\cdot)$ as follows. Add point $\underline{\omega}$ to the domain of $a(\cdot)$ so that $a(\underline{\omega}) = -\infty$ and, similarly, add point $\overline{\omega}$ to the domain of $b(\cdot)$ so that $b(\overline{\omega}) = +\infty$. There are several cases to consider, though the construction is similar in each instance:

Case 1. Let $\underline{c}_+ \leqslant \underline{s} < \overline{c}_+ < \overline{s}$, as in part (i) of the proposition. By Claim B.3, there exists an appropriate z^*. Define the function $m^*(\cdot) := m(\cdot|\underline{\omega}, z^*)$ and the CDF

$$F^*(x) = \begin{cases} G(x), & \text{for } x \in [\underline{s}, b(z^*)) \\ G(b(z^*)), & \text{for } x \in [b(z^*), z^*) \\ 1, & \text{for } x \in [z^*, \overline{s}] \end{cases}$$

To show that (F^*, m^*) forms an optimal primal-dual pair for Sender's problem [B.1], we must verify that the conditions of Theorem B.1 are satisfied. The pair $(\underline{\omega}, z^*) \in \mathfrak{X}$ by construction, so by Claim B.2, $m^*(\cdot)$ is convex and lies pointwise above $\phi(\cdot)$. By construction, $\text{supp}(F^*) \subseteq M(\underline{\omega}, z^*) = [\underline{s}, b(z^*)] \cup \{z^*\}$. Also by construction, $m^*(\cdot)$ is affine on the interval $[b(z^*), \overline{s}]$ where F^* is pooling, so it follows that $\mathbb{E}_G[m^*(x)] = \mathbb{E}_{F^*}[m^*(x)]$. Finally, it is obvious that $F^* \leqslant^{MPS} G$. Thus, by Theorem B.1, F^* is optimal.

Case 2. Let $\underline{s} < \underline{c}_+ < \overline{s} \leqslant \overline{c}_+$, as in part (ii) of the proposition. The proof is essentially identical to Case 1 above, with $m^*(\cdot) := m(\cdot | z_*, \overline{\omega})$ and

$$F^*(x) = \begin{cases} 0, & \text{for } x \in [\underline{s}, z_*) \\ G(z_*), & \text{for } x \in [z_*, a(z_*)) \\ G(x), & \text{for } x \in [a(z_*), \overline{s}] \end{cases}$$

and, as such, the details are omitted.

Case 3. Let $[\underline{c}_+, \overline{c}_+] \subseteq (\underline{s}, \overline{s})$, as in part (iii) of the proposition. Assume moreover that $(z_*, z^*) \in \mathfrak{X}$. Define the function $m^*(\cdot) := m(\cdot | z_*, z^*)$ and the CDF

$$F^*(x) = \begin{cases} 0, & \text{for } x \in [\underline{s}, z_*) \\ G(a(z_*)), & \text{for } x \in [z_*, a(z_*)) \\ G(x), & \text{for } x \in [a(z_*), b(z^*)) \\ G(b(z^*), & \text{for } x \in [b(z^*), z^*) \\ 1, & \text{for } x \in [z^*, \overline{s}] \end{cases}$$

We claim that (F^*, m^*) is an optimal primal-dual pair. By construction, $m^*(\cdot)$ is affine on the intervals $[\underline{s}, a(z_*)]$ and $[b(z^*), \overline{s}]$ on which F^* is pooling. Thus, $\mathbb{E}_G[m^*(x)] = \mathbb{E}_{F^*}[m^*(x)]$. Moreover, $\text{supp}(F) \subseteq M(z_*, z^*)$ by construction. The remaining conditions of Theorem B.1 are verified as in Case 1 above.

Case 4. Let $[\underline{c}_+, \overline{c}_+] \subseteq (\underline{s}, \overline{s})$, as in part (iii) of the proposition. Assume moreover that $(z_*, z^*) \notin \mathfrak{X}$. Define the function

$$m^*(x) := \begin{cases} \delta(x), & \text{for } x \in [\underline{s}, y^*) \\ \gamma(x), & \text{for } x \in [y^*, \overline{s}] \end{cases}$$

where y^* is the point defined in Claim B.4, and define the CDF

$$F^*(x) = \begin{cases} 0, & \text{for } x \in [\underline{s}, l(y^*)) \\ G(y^*), & \text{for } x \in [l(y^*), h(y^*)) \\ 1, & \text{for } x \in [h(y^*), \overline{s}] \end{cases}$$

By construction, $m^*(\cdot)$ is affine and decreasing on $[\overline{s}, y^*)$, affine and increasing on $[y^*, \overline{s}]$, and has a kink at y^*. By Claim B.4, it is continuous and lies pointwise above $\phi(\cdot)$. Convexity follows from the piecewise affine structure and continuity. It is affine precisely on the intervals on which F^* pools, and so satisfies $\mathbb{E}_G[m^*(x)] = \mathbb{E}_{F^*}[m^*(x)]$. Moreover, F^* is supported on $\{l(y^*), h(y^*)\}$, and by construction $\{x : m^*(x) = \phi(x)\} = \{l(y^*), h(y^*)\}$. It is clear that $F^* \leqslant^{MPS} G$. Thus the conditions of Theorem B.1 are met, and it follows that F^* is optimal.

Uniqueness: It remains to establish uniqueness of the solutions. Consider any one of the Cases 1-4 above, and let (F^*, m^*) be the constructed primal-dual pair. Corollary B.1 implies that, for any other optimal solution $\hat{F} \neq F^*$, the primal-dual pair $(\hat{F}, m^*)$ satisfies the conditions of Theorem B.1. Thus $\text{supp}(\hat{F}) \subseteq \{x : m^*(x) = \phi(x)\}$. In Case 4, this immediately implies that $\hat{F} = F$ G-a.e., as $|\{x : m^*(x) = \phi(x)\}| = 2$ and $\mu \notin \{x : m^*(x) = \phi(x)\}$ by the non-superdifferentiability assumption. In Cases 1-3, it is immediate that $\hat{F} = F^*$ G-a.e. on the separating region (interval on which $F^* = G$), as $m^*(\cdot)$ is strictly convex there by construction and any (G-non-null) pooling

would violate the condition that $\mathbb{E}_{\hat{F}}[m^*(x)] = \mathbb{E}_G[m^*(x)]$. It follows, by the same argument used for Case 4, that $\hat{F} = F^*$ G-a.e. on $[\underline{s}, \overline{s}]$.

□

§ B.1.4. *Proof when* $\beta = 0$

Proposition 14. *Assume that $\beta = 0$, $\alpha > 0$, and that G is continuous. Assume further that $\phi(\cdot)$ is not superdifferentiable at μ on $\mathcal{S}$. Then the optimal persuasion strategy F is essentially unique and has a separating-pooling form. In particular, its CDF F satisfies*

$$F(x) = \begin{cases} G(x), & \text{for } x \in [\underline{s}, b) \\ G(b), & \text{for } x \in [b, h(b)) \\ 1, & \text{for } x \in [h(b), \overline{s}] \end{cases}$$

for some $b < c_0$. That is, there is full separation below b and pooling at $h(b)$ above b.

Proof. When $\beta = 0$ and $\alpha > 0$, it is clear that there exists some $c_0 \in \mathbb{R}$ such that $\phi(\cdot)$ is strictly convex on $(-\infty, c_0)$ and strictly concave on (c_0, ∞). The proof thus follows almost verbatim the proof of part (i) of Proposition 13. The reader is referred to that proof, and in particular the construction of an appropriate primal-dual pair in Case B.1.3, for details.

□

B.2. Proof of Theorem 2

Claim B.5. *G is trivial if and only if exactly one of the following is true:*

1. $\mathbb{E}_F\left[e^{x/\lambda}\right] \leqslant 1$ *for all* $F \leqslant^{MPS} G$.
2. $\mathbb{E}_F\left[e^{-x/\lambda}\right] \leqslant 1$ *for all* $F \leqslant^{MPS} G$.

Proof. The "if" direction is obvious from the definition. For the "only if" direction, suppose that G is trivial. By Jensen's inequality, either (i) $\mathbb{E}_G\left[e^{s/\lambda}\right] \leqslant 1$ or (ii) $\mathbb{E}_G\left[e^{-s/\lambda}\right] \leqslant 1$, but not both. To see this, note that the function $h_0(y) := 1/y$ is strictly convex and G has non-singleton support by assumption, so that, e.g., $\mathbb{E}_G\left[e^{s/\lambda}\right] \leqslant 1$ implies that $1 \leqslant h_0\left(\mathbb{E}_G\left[e^{s/\lambda}\right]\right) < \mathbb{E}_G\left[h_0\left(e^{s/\lambda}\right)\right] = \mathbb{E}_G\left[e^{-s/\lambda}\right]$. In case (i), note that the function $h_1(y) := e^{y/\lambda}$ is strictly convex and so $\mathbb{E}_G[h_1(s)] = \sup_{F \leqslant^{MPS} G} \mathbb{E}_F[h_1(x)]$. This implies part 1. In case (ii), similarly, the function $h_2(y) := e^{-y/\lambda}$ is strictly convex and so $\mathbb{E}_G[h_2(s)] = \sup_{F \leqslant^{MPS} G} \mathbb{E}_F[h_2(x)]$. This implies part 2. □

The statement in the theorem regarding trivial priors follows from Claim B.5 and part 4 of Proposition 1. In particular, by part 4(a) of Proposition 1, part 1 of Claim B.5 is equivalent to $K(G) = \{0\}$; by part 4(b) of Proposition 1, part 2 of Claim B.5 is equivalent to $K(G) = \{\infty\}$. We now turn to parts 1-3 of the theorem.

Claim B.6. *For any prior G:*

1. *If $\mu < 0$, then* $\inf_{F \leqslant^{MPS} G} \mathbb{E}_F\left[e^{x/\lambda}\right] \leqslant 1$ *and* $\inf_{F \leqslant^{MPS} G} \mathbb{E}_F\left[e^{-x/\lambda}\right] > 1$.
2. *If $\mu > 0$, then* $\inf_{F \leqslant^{MPS} G} \mathbb{E}_F\left[e^{x/\lambda}\right] > 1$ *and* $\inf_{F \leqslant^{MPS} G} \mathbb{E}_F\left[e^{-x/\lambda}\right] \leqslant 1$.

Proof. Let the functions $h_1(\cdot)$ and $h_2(\cdot)$ be defined as in the proof of Claim B.5. Both are strictly convex functions, so $\inf_{F \leqslant^{MPS} G} \mathbb{E}_F [h_i(x)] = h_i(\mu)$ for $i = 1, 2$. The claim follows from observing that $h_1(\mu) \geqslant 1$ if and only if $\mu \geqslant 0$ and $h_2(\mu) \geqslant 1$ if and only if $\mu \leqslant 0$. □

B.3. Proofs from Section 5.2

[OUTDATED AND INCOMPLETE]

Proof of Proposition 4. **[TO BE COMPLETED]** Formally, [RP] may be written as

$$\sup_{F \leqslant^{MPS} G,\ k \in \mathbb{R}_+} k \tag{B.5}$$

$$\text{s.t.} \quad k \in K(F) \tag{B.6}$$

where the constraint correspondence

$$K(F) := \{k \in \mathbb{R}_+ : \mathbb{E}_F [p(x; k)] - P(k) \geqslant 0\}$$

consists of those activity parameters k such that Receiver is willing to act with *at least* probability $P(k)$. Due to Assumption 2, it is clear that the inequality constraint defining $K(F)$ will bind at the optimum. This gives us the following: □

Proof of Proposition 5. We begin by proving parts 2-4, and return to part 1 at the end.

Part 2 follows from Proposition 3 and Berge's Theorem.

For part 3, observe that Receiver's expected payoff (both gross and net of attention costs) must always be non-negative in the optimal mechanism. This is by revealed preference: Receiver can always choose to pay no attention and set $k = 0$ (i.e., always choose inaction), which yields payoff zero. Recall that, under full attention, Receiver gets expected payoff zero conditional on any signal realization, and hence also in expectation (see, e.g., Kolotilin (2015).) By the above observations and upper semi-continuity of Receiver's value function (which follows from Berge's Theorem), Receiver's expected utility must converge to zero as $\lambda \to 0$. As $\lambda \to \infty$, it is clearly optimal to pay vanishingly little attention, so as to economize on attention costs. Formally, Receiver's utility in the optimal mechanism is bounded above by his payoff under full information. By Lemma 2 of Matějka and McKay (2015), the full information payoff is

$$\max_{k \in \overline{\mathbb{R}}_+} \mathbb{E}_G \left[\lambda \log \left(P(k) e^{x/\lambda} + 1 - P(k) \right) \right]$$

In turn, this full-information payoff is bounded above by

$$U(\lambda) := \mathbb{E}_G \left[\lambda \log \left(\max \left\{ e^{x/\lambda}, 1 \right\} \right) \right]$$

For each $\lambda > 0$, let $\tilde{G}_\lambda(\cdot)$ be the CDF defined by $\tilde{G}_\lambda(y) = G(\lambda y)$. Because $\mathcal{S}$, the support of G, is compact and satisfies $0 \in \text{int}(\mathcal{S})$, it is clear that $\tilde{G}_\lambda \to^{w^*} \delta_{\{0\}}$ as $\lambda \to 0$, where $\delta_{\{x\}}$ is the Dirac measure on the set $\{x\}$. Thus, it follows that

$$U(\lambda) = \mathbb{E}_{\tilde{G}_\lambda} [\log (\max \{e^y, 1\})] \to 0$$

Part 3 of the proposition follows from the fact that **[NEED TO FILL IN.]**

Towards the proof of Part 4, define the function

$$\begin{aligned} V : \mathbb{R}_{++} \times \mathbb{R}_{+} &\to \mathbb{R}_{+} \\ (\lambda, k) &\mapsto \mathbb{E}_{F^*_\lambda}\left[p(x; k, \lambda)\right] \end{aligned}$$

Thus, $\tilde{V}(\lambda) := V(\lambda, k^*_\lambda)$ is Sender's value when the attention cost is $\lambda > 0$. Take $k = k^*_\lambda$ as given. By the Envelope Theorem (see, e.g., Corollary 4 of Milgrom and Segal (2002)), we have

$$V'(\lambda, k^*_\lambda) = \mathbb{E}_{F^*_\lambda}\left[\frac{\partial p(x; k^*_\lambda, \lambda)}{\partial \lambda}\right]$$

and, by direct computation,

$$\frac{\partial p(x; k, \lambda)}{\partial \lambda} = -\frac{x}{\lambda^2} \cdot p(x; k)$$

so that

$$V'(\lambda, k^*_\lambda) = -\frac{1}{\lambda^2} \cdot \underbrace{\mathbb{E}_{F^*_\lambda}\left[x \cdot p(x; k^*_\lambda, \lambda)\right]}_{=:\, U^R(\lambda)}$$

But note that $U^R(\lambda)$ simply defines Receiver's expected material (i.e., gross of attention costs) payoff under the optimal mechanism. By revealed preference, we have $U^R(\cdot) \geqslant 0$, with a strict inequality whenever Receiver pays strictly positive attention cost. By Assumption 3, strictly positive attention is equivalent to $k^*_\lambda > 0$. It follows that$V(\lambda', k^*_\lambda) < P(k^*_\lambda)$ for any $\lambda' > \lambda$, so that k^*_λ is not an implementable activity level when attention costs are λ'. It follows that $k^*_{\lambda'} < k^*_\lambda$ and by Proposition 4, that $V(\lambda') < V(\lambda)$. This completes the proof of part 4.

Finally, we prove part 1 of the proposition. Recall the form of the optimal persuasion strategy described in Proposition 3. Let $h(\lambda) \geqslant b(\lambda)$ denote, as functions of λ, the location of the high atom (conditional mean of the pooling region) and the boundary between the separation region and pooling region under the optimal signal structure F^*_λ. By definition, $h(\lambda) = \mathbb{E}_G\left[s | s \geqslant b(\lambda)\right]$ so that $h(\lambda)$ is (locally) increasing iff $b(\lambda)$ is (locally) increasing. Thus, it is easy to see that part 1 of the proposition is equivalent to $h(\cdot)$ being an increasing function. We show that this is true in two steps. Define $\tilde{h}(\lambda, k)$ as the high atom induced by the solution to

$$\max_{F \leqslant^{MPS} G} \mathbb{E}_F\left[p(x; k, \lambda)\right]$$

which takes the activity parameter k as given (though it needn't be implementable). Clearly, $h(\lambda) = \tilde{h}(\lambda, k^*_\lambda)$. The first step is to show that $\tilde{h}(\cdot, k^*_\lambda)$ is locally strictly increasing at λ. The second step is to show that $\tilde{h}(\lambda, \cdot)$ is locally strictly increasing in k^*_λ. **[NEED TO FILL IN.]**

Claim B.7. *For any $\lambda' > \lambda > 0$ and $k \in \mathbb{R}_{++}$, we have $h(\lambda', k) > h(\lambda, k)$.*

Proof of Claim B.7. Define $z := \lambda'/\lambda > 1$ and consider the change of variables from x to $y := x/z$. For any CDF H, let $\tilde{H}_z(\cdot)$ be the CDF defined by $\tilde{H}_z(t) = H(zt)$. Thus, if $X \sim F$, then $Y := X/z \sim \tilde{F}_z$. It is easy to see that $F \leqslant^{MPS} G$ if and only if $\tilde{F}_z \leqslant^{MPS} \tilde{G}_z$ and, moreover,

$$\mathbb{E}_F\left[p(x; k, \lambda')\right] = \frac{1}{z} \cdot \mathbb{E}_{\tilde{F}_z}\left[p(y; k, \lambda)\right]$$

Thus, the problems

[B.7] $$\max_{F \leqslant^{MPS} G} \mathbb{E}_F\left[p(x; k, \lambda')\right]$$

[B.8] $$\max_{\tilde{F}_z \leqslant^{MPS} \tilde{G}_z} \mathbb{E}_{\tilde{F}_z}\left[p(y; k, \lambda)\right]$$

are equivalent in the sense that $\tilde{h}_z(\lambda, k)$, the high atom induced by the solution to problem [B.8], satisfies $z \cdot \tilde{h}_z(\lambda, k) = h(\lambda', k)$.

□

Claim B.8. *For any $\lambda > 0$ and $k, k' \in \mathbb{R}_{++}$ such that $k' < k$, we have $h(\lambda, k') > h(\lambda, k)$.*

Proof of Claim B.8. □

B.4. Proofs from Section 5.3

Proof of Proposition 6. Fix $k = 1$ and consider the *unconstrained problem*

[B.9] $$\max_{F \leqslant^{MPS} G} \mathbb{E}_F \left[\tau(x; 1)\right]$$

where $\tau(x; 1) := x \cdot p(x; 1) - x/2$. Since $\mathbb{E}_F \left[x/2\right] = 0$ for all $F \leqslant^{MPS} G$, this is equivalent to Sender's component problem at $k = 1$, absent the activity constraint.

Claim B.9. *The function $\tau(\cdot; 1)$ satisfies:*

1. *Symmetry: $\tau(-x; 1) = \tau(x; 1)$ for all $x \in \mathbb{R}$;*
2. *It is non-negative, strictly increasing on $\mathbb{R}_+$, and strictly decreasing on $\mathbb{R}_-$;*
3. *There exists $y > 0$ such that $\tau(\cdot; 1)$ is strictly convex on $(-y, y)$ and strictly concave on $(-\infty, -y]$ and $[y, \infty)$;*
4. *$\tau(x; 1) \leqslant |x|/2$ for all $x \in \mathbb{R}$, with strict inequality for all $x \neq 0$;*
5. *$\tau'(x; 1) > 1/2$ for all $x \geqslant y$ and, symmetrically, $\tau'(x) < -1/2$ for all $x \leqslant -y$.*

Proof of Claim B.9. Parts 1 and 2 are easy to verify by direct calculation. Part 3 follows from Part 1 together with the special case of part (i) of Lemma B.2 in which $\alpha = 0$ and $\beta = 1$. For part 4, assume without loss that $x \geqslant 0$. Since $p(x; 1) \in (0, 1)$, we have $\tau(x; 1) \leqslant x - x/2 = x/2$ with strict inequality when $x > 0$. For part 5, it is again without loss to assume $x > 0$. We compute that

[B.10] $$\tau''(x; 1) =^{sign} \xi(x) := x + 2\lambda + e^{x/\lambda}(2\lambda - x)$$

[B.11] $$\frac{d}{dx}\left(\frac{x}{2} - \tau(x; 1)\right) =^{sign} \zeta(x) = \lambda + e^{x/\lambda}(\lambda - x)$$

Thus, $\zeta(x) = \xi(x) - x - 2\lambda e^{x/\lambda} < \xi(x)$. Hence, for all $x \geqslant y$ we have $\zeta(x) < 0$, which by construction is equivalent to $\tau'(x; 1) > 1/2$. □

By part 2 of Theorem 1, we know a solution to [B.9] exists and that it must be monotone partitional with pooling at the bottom and top, and separation in the middle. By the uniqueness part of Theorem 1, the solution is essentially unique. Thus, by Claim B.11 there exists a symmetric solution F^* of this monotone partitional form. Denote the separation interval of F^* by $(-b, b)$.

Claim B.10. *The separation region of F^* is nonempty, i.e., $b > 0$.*

Proof of Claim B.10. If the y defined in Claim B.9 satisfies $y \geqslant \overline{s}$, the $\tau(\cdot;1)$ is strictly convex on $\mathcal{S}$ and so full revelation is uniquely optimal. Thus, $b = \overline{s} > 0$. So assume that $y < \overline{s}$. Then we are in either Case 3 or Case 4 from the proof of Theorem 1 (see Appendix B.1.3), and it remains to show that we are in Case 3. That this is true follows from Claim B.9. In particular, the multiplier $m(\cdot)$ that certifies optimality of F^* must be tangent to $\tau(\cdot;1)$ at the points $\{-h, h\}$ for some $h \geqslant y$. Because $\tau(h;1) < h/2$ and $\tau'(h;1) > 1/2$, it must be that $m^*(z) = 0$ for some $z > 0$ and, since $\tau(\cdot;1) \geqslant 0$, it must be that $z \leqslant b$. (This argument is illustrated in the right-hand panel of Figure 6 in the main text.) □

Claim B.11. *If* [B.9] *has a solution, it has a symmetric solution F^*.*

Proof of Claim B.11. For any $F \leqslant^{MPS} G$, define its *reflection* by $F^r(x) := 1 - F(-x)$. We claim that symmetry of G implies that $F \leqslant^{MPS} G$ if and only if $F^r \leqslant^{MPS} G$. First, it's easy to see that $\mathbb{E}_F[x] = \mathbb{E}_{F^r}[x]$. Second, for every $s \in \mathcal{S}$ we have

$$\begin{aligned}
\int_{-\overline{s}}^{s} [F(x) - G(x)]\, dx &= \int_{-\overline{s}}^{s} [(1 - F^r(-x)) - (1 - G(-x))]\, dx && \text{by symmetry of } G \\
&= \int_{-\overline{s}}^{s} [G(-x) - F^r(-x)]\, dx \\
&= \int_{-\overline{s}}^{s} [F^r(x) - G(x)]\, dx && \text{by change of variable}
\end{aligned}$$

which establishes the claim that the feasible set is symmetric under reflection. By part 1 of Claim B.9, $\tau(x;1)$ is symmetric about $x = 0$, from which it follows that if F is optimal then F^r is also optimal. Define

$$F^*(x) := \frac{F(x) + F^r(x)}{2}$$

It is symmetric by construction, attains the optimal value because the objective in [B.9] is linear, and feasible because the feasible set is convex. □

Claim B.12. *If F is symmetric, then $k = 1$ solves $\mathbb{E}_F[p(x;k)] = P(k)$.*

Proof of Claim B.12. It is easy to verify that

$$1 - p(x;k) = \frac{ke^{-x/\lambda}}{k^2 + ke^{-x/\lambda}} =: h(x;k)$$

and, in particular, $h(x;1) = p(-x;1)$. Thus,

$$\begin{aligned}
\mathbb{E}_F[p(x;1)] &= \int_0^{\overline{s}} p(x;1) dF(x) + \int_{-\overline{s}}^{0} p(x;1) dF(x) \\
&= \int_0^{\overline{s}} p(x;1) dF(x) + \int_{-\overline{s}}^{0} [1 - p(-x;1)]\, dF(x) && \text{by } h(x;1) = p(-x;1) \\
&= \frac{1}{2} + \int_0^{\overline{s}} p(x;1) dF(x) - \int_0^{\overline{s}} p(x;1) dF(x) && \text{by symmetry of } F \\
&= \frac{1}{2}
\end{aligned}$$

The claim follows from $P(1) = 1/2$. □

We have thus constructed a symmetric solution to the unconstrained problem [B.9] with the properties claimed in the proposition. Moreover, Claim B.12 establishes that the activity constraint [A] is not binding at this solution. As the value of the unconstrained problem is weakly greater than that of the component problem [CP] at $k = 1$, the constructed F^* also solves the latter problem. The proposition follows from these observations together with Lemma **??**, as $k = 1$ is interior. □

Proof of Proposition 8. For fixed k, it follows from Lemma 2 of Matějka and McKay (2015) that the optimal F is given by the solution to

$$\max_{F \leqslant^{MPS} G} \mathbb{E}_F \left[\lambda \log \left(P(k) e^{x/\lambda} + 1 - P(k) \right) \right]$$

It is easy to verify that the integrand is convex in x, and strictly convex whenever $k \in \mathbb{R}_{++}$, which proves the lemma.

[TO BE FILLED IN]

□

Proof of Proposition 9. We begin by proving parts 2 and 3 of the proposition. Recalling [B.10] from the proof of Claim B.9, define the function $y : \mathbb{R}_{++} \to \mathbb{R}_{++}$ by

$$y(\lambda) := \sup \left\{ x : \xi(x, \lambda) := x + 2\lambda + e^{x/\lambda} (2\lambda - x) \geqslant 0 \right\}$$

Thus, $(-y(\lambda), y(\lambda))$ is the interval over which $\tau(x; 1)$ is (strictly) convex.

Claim B.13. *The function $y(\cdot)$ is continuous, strictly increasing, unbounded above, and satisfies $\lim_{\lambda \to 0} y(\lambda) = 0$.*

Proof of Claim B.13. Continuity is obvious from Berge's Theorem. We have $\xi(0, \lambda) = 4\lambda$, which implies $y(\lambda) \geqslant 0$ with a strict inequality for $\lambda > 0$. Moreover, the function $\xi(\cdot)$ has strict increasing differences in (x, λ) on $\mathbb{R}_{++} \times \mathbb{R}_{++}$:

$$\frac{\partial^2 \xi}{\partial x \partial \lambda} = \frac{x^2 e^{x/\lambda}}{\lambda^3} > 0$$

It follows that $y(\cdot)$ is strictly increasing. To see that it is unbounded above, notice that $\xi(2\lambda, \lambda) = 4\lambda$, from which it follows that $y(\lambda) > 4\lambda$. Finally, for any fixed $x > 0$ we have $\lim_{\lambda \to 0} \xi(x, \lambda) = -\infty$, so it must be that $\lim_{\lambda \to 0} y(\lambda) = 0$. □

Now, parts 2 and 3 of the proposition follow from Claim B.13. Because $y(\cdot)$ is unbounded above and strictly increasing, there exists $\overline{\lambda} > 0$ such that for all $\lambda \geqslant \overline{\lambda}$, $y(\lambda) \geqslant \overline{s}$ so that $\tau(\cdot; 1)$ is strictly convex on $\mathcal{S}$, in which case full disclosure is uniquely optimal. This gives part 2. Moreover, by continuity and the fact that $y(0) = 0$, as $\lambda \to 0$ the separation region, a subset of $(-y(\lambda), y(\lambda))$, must converge to $\{0\}$. This implies part 2.

The following fact, the proof of which follows from direct computation, is useful in the sequel.

Claim B.14. *The function $\tau(x;1)$ satisfies:*

$$\frac{\partial \tau}{\partial \lambda} = -\frac{x^2 e^{x/\lambda}}{\lambda^2 \left(e^{x/\lambda}+1\right)^2} \leqslant 0 \qquad (< 0 \text{ for } x \neq 0) \tag{B.12}$$

$$\frac{\partial^2 \tau}{\partial x \partial \lambda} =^{sign} -x \cdot \xi(x,\lambda) \tag{B.13}$$

Thus, $\partial^2\tau/\partial x\partial\lambda > 0$ if $|x| > y(\lambda)$ and < 0 if $|x| < y(\lambda)$.

We turn now to part 1 of the proposition. Let $h(\lambda) \geqslant b(\lambda) > 0$ denote, as functions of λ, the location of the high atom (conditional mean of the upper pooling region) and the boundary between the pooling region and upper pooling region under the optimal signal structure F^*_λ. By definition, $h(\lambda) = \mathbb{E}_G[s|s \geqslant b(\lambda)]$ so that $h(\lambda)$ is (locally) increasing iff $b(\lambda)$ is (locally) increasing. Thus, it is easy to see that part 1 is implied by the following claim:

Claim B.15. *The function $h(\cdot)$ is non-decreasing, and locally strictly increasing whenever $h(\lambda) < \overline{s}$.*

Proof of Claim B.15. If $h(\lambda) = \overline{s}$, so that $F^*_\lambda = G$ is fully separating, then it follows from Claim B.13 that $F^*_{\lambda'} = G$ for all $\lambda' \geqslant \lambda$. Thus, suppose that $h(\lambda) < \overline{s}$. From the proof of Theorem 1, we see that $(h(\lambda), b(\lambda))$ is given by the unique (strictly positive) solution to the following system:

$$h(\lambda) = \mathbb{E}_G[s|s \geqslant b(\lambda)] \tag{B.14}$$

$$\tau(h(\lambda);1,\lambda) = \tau(b(\lambda);1,\lambda) + (h(\lambda)-b(\lambda)) \cdot \tau'(h(\lambda);1,\lambda) \tag{B.15}$$

Moreover,

$$\tau(h(\lambda);1,\lambda) \begin{cases} < \tau(x;1,\lambda) + (h(\lambda)-x)\cdot\tau'(h(\lambda);1,\lambda) & \text{for } x \in [0,b(\lambda)) \\ > \tau(x;1,\lambda) + (h(\lambda)-x)\cdot\tau'(h(\lambda);1,\lambda) & \text{for } x \in (b(\lambda),y(\lambda)) \end{cases} \tag{B.16}$$

Let $\lambda' := \lambda + \varepsilon$, with $\varepsilon > 0$ sufficiently small so that $h(\lambda) > y(\lambda')$. Define the function $\zeta : (y(\lambda'),\infty) \to (0, y(\lambda'))$ by

$$\zeta(x) := \inf\{y \in \mathbb{R}_+ : \tau(x;1,\lambda') - (x-y)\cdot\tau'(x;1,\lambda') \geqslant \tau(y;1,\lambda')\}$$

It is easy to see from, e.g., Claims B.9 and B.10, that $\zeta(\cdot)$ is well-defined. In particular, for any $x > y(\lambda')$, $\tau(x) \in (0, y(\lambda))$ is the unique point at which the line tangent to $\tau(\cdot;1,\lambda')$ at x intersects (and in particular crosses from below) the curve $\tau(\cdot;1,\lambda')$. We claim that $\zeta(h(\lambda)) > b(\lambda)$. By [B.13] in Claim B.14 and the assumption that $h(\lambda) > y(\lambda')$, we have $\tau'(h(\lambda);1,\lambda')$, and by [B.12] we have $\tau(h(\lambda);1,\lambda) > \tau(h(\lambda);1,\lambda')$. Thus, from [B.15] and [B.16] we see that, for any $x \in (0, b(\lambda)]$,

$$\begin{aligned} &\tau(h(\lambda);1,\lambda') - (h(\lambda)-x)\,\tau'(h(\lambda);1,\lambda') \\ <&\tau(h(\lambda);1,\lambda) - (h(\lambda)-x)\,\tau'(h(\lambda);1,\lambda) \\ \leqslant&\tau(x;1,\lambda) \end{aligned}$$

where the final inequality is an equality if and only if $x = b(\lambda)$. This establishes that $\zeta(h(\lambda)) > b(\lambda)$, as claimed. Towards contradiction, suppose that $h(\lambda') \leqslant h(\lambda)$. If $h(\lambda') = h(\lambda)$, then since $\zeta(h(\lambda)) > b(\lambda)$, we have $h(\lambda') < \mathbb{E}_G[s|s \geqslant \zeta(h(\lambda'))]$. Thus, it is impossible to jointly satisfy the MPS constraint and tangency condition as in [B.14] and [B.15]. Contradiction. Moreover, it is easy to see that $\zeta(\cdot)$ is a strictly decreasing function. (This follows from the same argument used in the proof of Claim B.1.) Thus, if $h(\lambda') < h(\lambda)$ we again have $h(\lambda') < \mathbb{E}_G[s|s \geqslant \zeta(h(\lambda'))]$, a contradiction. The claim follows. □

Finally, part 4 of the proposition follows from Claims B.16 and B.18 below.

Claim B.16. *Sender's expected utility under the optimal strategy is strictly decreasing in λ.*

Proof of Claim B.16. Let $\lambda' > \lambda$, and let $F^*_{\lambda'}$ and F^*_{λ} be the corresponding optimal signal structures. By Proposition 6, the supports of both signal structures are strict supersets of $\{0\}$. It follows from [B.12] in Claim B.14 that $\mathbb{E}_{F^*_{\lambda'}}[\tau(x;1,\lambda')] < \mathbb{E}_{F^*_{\lambda'}}[\tau(x;1,\lambda)]$ and, by revealed preference, $\mathbb{E}_{F^*_{\lambda'}}[\tau(x;1,\lambda)] \leqslant \mathbb{E}_{F^*_{\lambda}}[\tau(x;1,\lambda)]$. Combining these inequalities yields the claim. □

Claim B.17. *Let $k \in \overline{\mathbb{R}}_+$ be given. For any $t \in \mathbb{R}_{++}$, let (A_t, X) be a $\{0,1\} \times \mathbb{R}$-valued random vector with the joint distribution induced by $X \sim F$ and $(A_t|X=x) \sim p(x;k,t)$. Then, for any $\lambda, \lambda' \in \mathbb{R}_{++}$,*

$$\lambda I(A_\lambda; X) = (\lambda - \lambda')H(A_\lambda) + \lambda' I(A_{\lambda'}; X) \tag{B.18}$$

Proof. We have $H(X|A_t) = \mathbb{E}_F[h(p(x;k,t))]$, where $h(x) := -x\log(x) - (1-x)\log(1-x)$ is the binary entropy function. Define the constant $z := \lambda'/\lambda$ and the random variable $Y := zX$, which has CDF $\tilde{F}(y) := F(y/z)$. Then,

$$\begin{aligned} H(X|A_\lambda) &= \int h(z\cdot x; k, \lambda')dF(x) \\ &= z \cdot \int h(p(y;k,\lambda'))d\tilde{F}(y) \\ &= z \cdot H(A_{\lambda'}|Y) \end{aligned}$$

where the second line is a standard change of variable. Thus, by Lemma A.1

$$\begin{aligned} \lambda I(A_\lambda; X) &= \lambda H(A_\lambda) - \lambda H(A_\lambda|X) \\ &= \lambda H(A_\lambda) - \lambda \cdot z \cdot H(A_{\lambda'}|Y) \\ &= (\lambda - \lambda')H(A_\lambda) + \lambda' I(A_{\lambda'}; Y) \end{aligned}$$

where the third equality uses $H(A_\lambda) = h(P(k)) = H(A_{\lambda'})$. Finally, because the mapping $x \mapsto y := zx$ is bijective, we have $I(A_{\lambda'}; Y) = I(A_{\lambda'}; X)$ by the data processing inequality (Theorem 2.8.1 of Cover and Thomas (2006)). Combining with the above display yields the claim. □

Claim B.18. *Receiver's expected utility under the optimal mechanism is strictly decreasing in λ.*

Proof of Claim B.18. Let $U^S(\lambda)$ and $U^R(\lambda)$ denote, respectively, Sender and Receiver's expected utilities under the optimal mechanism as a function of λ. Because preferences are aligned, $U^R(\lambda) = U^S(\lambda) - \lambda I(A_\lambda; X_\lambda)$ where $X_\lambda \sim F^*_\lambda$ and $(A_\lambda|X_\lambda = x) \sim p(x;1,\lambda)$. For any $\lambda' > \lambda > 0$, we have

$$\begin{aligned} U^R(\lambda) &= U^S(\lambda) - (\lambda - \lambda')H(P(1)) - \lambda' I(A_{\lambda'}; X_\lambda) \\ &> U^S(\lambda) - \lambda' I(A_{\lambda'}; X_\lambda) \\ &> U^S(\lambda') - \lambda' I(A_{\lambda'}; X_\lambda) \\ &\geqslant U^R(\lambda') \end{aligned}$$

where the first line follows from Claim B.17, the second line follows from $\lambda' > \lambda$ and $H(P(1)) = \log(2) > 0$, and the third line follows from Claim B.16. The fourth line is a consequence of the inequality $I(A_{\lambda'}; X_{\lambda'}) \geqslant I(A_{\lambda'}; X_\lambda)$. To see this, note that by Proposition 6 and part 1 of Proposition 9, the partition of $[\underline{s}, \overline{s}]$ generated by $X_{\lambda'}$ is strictly finer than that generated by X_λ. Thus, there exists some measurable function $f(\cdot)$ such that $X_\lambda = f(X_{\lambda'})$ and therefore $A_{\lambda'} \to X_{\lambda'} \to X_\lambda$ is a Markov chain. The desired inequality then follows from the data processing inequality (Theorem 2.8.1 of Cover and Thomas (2006)). □

C. Solution to the Binary-State Example

Proof of Proposition 7. □

[TO BE ADDED]

D. Details for Section 6

[OUTDATED AND INCOMPLETE]

The model is exactly as in Section 3.1, save for one change in terminology and a few new pieces of notation. (As was the case there, we formulate the model for completely general attention cost functions.) To reflect Sender's lack of commitment, we refer to his chosen information structure as a *communication strategy* and refer to realizations as *messages*. Given an attention strategy $(\mu, \mathscr{M})$, an *action strategy* for Receiver, denoted α, is a Markov kernel from $\mathscr{M}$ to $\mathscr{A}$. That is, $\alpha : m \mapsto \alpha(m, \cdot) \in \Delta(\mathscr{A})$ maps perceptions into distributions over actions. A *strategy* for Receiver is a pair consisting of an action strategy and attention strategy. Finally, given an attention strategy $(\mu, \mathscr{M})$, a *belief system* for Receiver, denoted ν, is a Markov kernel from $\mathscr{M}$ to $\mathcal{S}$. That is, $\nu : m \mapsto \nu(m, \cdot) \in \Delta(\mathcal{S})$ maps perceptions into posterior beliefs about the state.

Definition 8. A *perfect Bayesian equilibrium (PBE)* consists of (i) a communication strategy $(\pi, \mathfrak{X})$ for Sender and (ii) a strategy $(\alpha, \mu, \mathscr{M})$ and a belief system ν for Receiver such that:

1. ν is obtained from $(\pi, \mathfrak{X})$ and $(\mu, \mathscr{M})$ via Bayesian updating, whenever possible;
2. $(\mu, \mathscr{M})$ solves Receiver's problem [RP] given $(\pi, \mathfrak{X})$;
3. $\text{supp}(\alpha(m, \cdot)) \subseteq \arg\max_{a \in \mathscr{A}} \int_{\mathcal{S}} u(a, s) \nu(m, ds)$, for every $m \in \mathscr{M}$;
4. $\text{supp}(\pi(s, \cdot)) \subseteq \arg\max_{x \in \mathfrak{X}} \int_{\mathscr{M} \times \mathscr{A}} v(a, s) \alpha(da, dm) \mu(x, dm)$, for every $s \in \mathcal{S}$.

This is essentially the standard definition of equilibrium in the cheap talk literature. The only slightly non-standard piece is part 1, which gives the appropriate consistency condition for Receiver's belief system. In a typical cheap talk model, Receiver's beliefs are contingent on Sender's messages, and so must be consistent with Sender's strategy and Bayes' rule on the equilibrium path. Here, Receiver doesn't directly "see" Sender's messages, so the consistency condition is with respect to both Sender's communication strategy and Receiver's attention strategy. Parts 2-3 correspond to Receiver optimality, and part 4 corresponds to Sender optimality.

Under Assumption 1 (imposing the RI cost function), we may substantially simplify the description of equilibrium.[70] As was the case under Sender commitment, every communication strategy $(\pi, \mathfrak{X})$ induces a

(70) The formal arguments follow those from Section 4 nearly verbatim, so we simply summarize them here.

distribution of posterior means, denoted F_π, that summarizes all of the information- and payoff-relevant content of Sender's messages. Indeed, we may identify Sender's message space with $\text{CH}(\mathcal{S})$, drop references to $(\pi, \mathfrak{X})$ altogether, and simply identify Sender's strategy with the induced (random) posterior mean $X \sim F := F_\pi$. Given F, the solution to Receiver's RI problem is exactly as described in Proposition 1. Namely, Receiver's strategy will involve direct recommendations (i.e., $\mathscr{M} = \mathscr{A}$ and $\alpha(m, \cdot) = \delta_{\{m\}}(\cdot)$) and be described by a shifted Logit stochastic choice rule parametrized by k. Indeed, since she now takes k as given when choosing which messages to transmit, Sender's optimization problem is much simpler than in Sections 4 and 5. The one subtlety is that, now, it is *not* necessarily without loss to assume that the stochastic choice rule takes the shifted Logit shape outside the support of F, i.e., the "π-almost everywhere" qualifier in part 2 of Proposition 1 cannot be immediately dispensed with. Such an assumption would amount to an equilibrium refinement on Receiver's off-path actions.[71] However, it turns out that our equilibrium characterization does not require any assumptions about such off-path behavior.

[INCOMPLETE.]

Assume that G has strictly positive density $g(\cdot) > 0$ on $[\underline{s}, \overline{s}]$, where $\underline{s} < 0 < \overline{s}$. Preferences and actions are as described in Section 3.1.

[MOVE DETAILS OF EQ DEFN HERE.]

D.1. Full Attention

Assume that Receiver is fully attentive, i.e., $\lambda = 0$. Assume, moreover, that $\beta > 0$ to rule out the case where babbling is the only equilibrium.

Claim D.1. *All equilibria are monotone partitional. Every equilibrium is outcome-equivalent to one with at most two cells.*

Proof sketch. Standard argument, both agents' utilities have strictly increasing differences in (a, s). Reduction to two-cell partitions follows from usual pooling argument. □

Suppose there is an equilibrium of the above form with two cells. Let the cells be denoted $[\underline{s}, c)$ and $[c, \overline{s}]$, where c stands for *cutoff*. Denote the corresponding messages by m_l and m_h, respectively. Assume that Receiver acts with probability one on the high cell, and is inactive with probability one on the low cell.

In equilibrium, Sender must be indifferent between sending m_l and m_h:

$$0 = \alpha + \beta c \qquad \iff \qquad c = -\frac{\alpha}{\beta} \tag{D.1}$$

(Note that $c = 0$ when prefs are aligned, and $c \downarrow -\infty$ as bias gets larger.) Since $c < 0$, inaction with probability one on the low cell is IC for Receiver. Action is optimal on the high cell iff

$$\mathbb{E}_G[s | s \geqslant c] \geqslant 0$$

which is equivalent to

$$\int_c^{\overline{s}} sG(ds) \geqslant 0$$

If $f(\cdot)$ is $U[\underline{s}, \overline{s}]$, then this is equivalent to $|\overline{s}| \geqslant |c|$. Thus, binary partitions can be supported for small divergence in preferences. As divergence becomes large, only babbling can be sustained.

(71) We remind the reader that this was indeed without loss under Sender commitment. Matějka and McKay (2012) implicitly assume such a refinement, while Ravid (2018) takes a more standard game-theoretic approach based on trembling-hand perfection.

D.2. Limited Attention

For a given Sender strategy, let $\hat{u}(m) := \mathbb{E}[s|m]$ be the expected state given message m.

Let $k \in \mathbb{R}_{++}$ be fixed; assume it is implemented in equilibrium. Receiver's stochastic choice rule is

$$p(x;k) = \frac{k \exp(\hat{u}(x)/\lambda)}{1 + k \exp(\hat{u}(x)/\lambda)}$$

This is strictly increasing in x (endowing the message space with the order consistent with increasing $\hat{u}(x)$).

Claim D.2. *All equilibria are monotone partitional, with* at most *two cells.*

Proof sketch. Similar to before. But now, Receiver's best response is either constant (if $k \in \{0, \infty\}$) or *strictly* increasing (if $k \in (0, \infty)$) in the expected state. Thus, if $\alpha + \beta \cdot s > 0$ and $x(s)$ is not the highest message, Sender has strict incentive to deviate upward. Etc .... □

Sender's indifference at a cutoff point c is

$$p(\mathbb{E}[s|s < c], k) \cdot (\alpha + \beta \cdot c) = p(\mathbb{E}[s|s \geqslant c], k) \cdot (\alpha + \beta \cdot c)$$

which implies that

$$0 = (\alpha + \beta \cdot c) \qquad \iff \qquad c = -\frac{\alpha}{\beta}$$

as before. Thus the equilibrium partitions must be identical.

Note that, in this case, we do not have the same IC conditions for Receiver. Instead, we need to verify whether $k \in (0, \infty)$ can be implemented. (This assumption is what led us to the cutoff value for Sender.) If only $k \in \{0, \infty\}$ can be implemented, then we are effectively in a babbling equilibrium.

Given that the signal structure is binary, we can solve in closed form for the k that's implemented in equilibrium. As a function of c, the unconditional probability of action is given by

$$\overline{p}(c) := \frac{(1 - F(c)) \exp\left(\frac{\mathbb{E}[s|s \geqslant c]}{\lambda}\right) + F(c) \exp\left(\frac{\mathbb{E}[s|s < c]}{\lambda}\right) - 1}{\left(\exp\left(\frac{\mathbb{E}[s|s \geqslant c]}{\lambda}\right) - 1\right)\left(1 - \exp\left(\frac{\mathbb{E}[s|s < c]}{\lambda}\right)\right)}$$

Then the two-cell partition can be implemented if and only if $\overline{p}(c) \in (0, 1)$. (This is equivalent to $k(c) := \frac{\overline{p}(c)}{1 - \overline{p}(c)}$ being interior.)

It is clear that $\mathbb{E}[s|s \geqslant c] > 0$ is a necessary condition for $\overline{p}(c) > 0$. Thus, implementing a binary partition is strictly harder with RI receiver.

E. Details for Section 7.3

[OUTDATED AND INCOMPLETE]

For fixed price q and in our notation, Yang (2017) considers the problem:

[E.1] $$\max_{k, x(\cdot)} \mathbb{E}_G[(q - \delta x(s)) \cdot p(x(s) - q; k)]$$

[E.2] subject to $x(s) \in [0, s] \qquad \forall s \in \mathbb{R}_+$

[E.3] and $$\mathbb{E}_G[p(x(s) - q; k)] = \frac{k}{k+1}$$

We may write $q - \delta x(s) = q(1-\delta) - \delta t(s)$ where $t(s) := x(s) - q$. Under this change of variables, the problem becomes:

$$\max_{k, t(\cdot)} \mathbb{E}_G \left[(q(1-\delta) - \delta t(s)) \cdot p(t(s); k) \right] \tag{E.4}$$

$$\text{subject to} \quad t(s) \in [-q, s-q] \qquad \forall s \in \mathbb{R}_+ \tag{E.5}$$

$$\text{and} \quad \mathbb{E}_G \left[p(t(s); k) \right] = \frac{k}{k+1} \tag{E.6}$$

Define the random variable $S_q := S - q$, which has distribution $G_q(x) := G(x - q)$ and is supported on $[-q, \infty)$. Program [E.4] – [E.6] can be written in relaxed form as:

$$\max_{k, F \leqslant^{FOSD} G_q} \mathbb{E}_F \left[(q(1-\delta) - \delta x) \cdot p(x; k) \right] \tag{E.7}$$

$$\text{subject to} \quad \text{supp}(F) \subseteq [-q, \infty) \tag{E.8}$$

$$\text{and} \quad \mathbb{E}_F \left[p(x; k) \right] = \frac{k}{k+1} \tag{E.9}$$

To see that this is a legitimate relaxation, note that the random variables $t(S)$ and S_q live on the same probability space and $t(S) \leqslant S_q$ almost surely, which implies the FOSD ordering of their CDFs. Of course, the formulation of program [E.7] – [E.9] allow for additional randomization independent of the realized state s, which is not allowed in the original problem. In turn, program [E.7] – [E.9] is a special case of the following problem (in particular, the special case in which $\alpha > 0$ and $\beta < 0$):

$$\max_{k, F \leqslant^{FOSD} G} \mathbb{E}_F \left[(\alpha + \beta x) \cdot p(x; k) \right] \tag{E.10}$$

$$\text{subject to} \quad \text{supp}(F) \subseteq CH\left(\text{supp}(G)\right) \tag{E.11}$$

$$\text{and} \quad \mathbb{E}_F \left[p(x; k) \right] = \frac{k}{k+1} \tag{E.12}$$

Ignoring the support constraint [E.11], this is identical to our Sender's component problem [CP] with an FOSD constraint taking place of the MPS constraint. (The FOSD constraint corresponds to surplus division, while the MPS constraint corresponds to garbling. The support constraint [E.11] is implied by the MPS constraint in our persuasion model.) Moreover, there is a simple analogue of the verification result we have relied on.

Proposition 15 (LP Duality)**.** *Consider the optimization problem*

$$\sup_{F \leqslant^{FOSD} G} \mathbb{E} \left[\phi(x) \right] \tag{E.13}$$

subject to $supp(F) \subseteq CH\left(supp(G)\right) = [\underline{x}, \overline{x}]$. *Suppose there exists a CDF* F *and a non-decreasing function* $p : [\underline{x}, \overline{x}] \to \mathbb{R}$ *such that* $p(x) \geqslant \phi(x)$ *for all* x *that satisfy*

$$supp(F) \subseteq \{x \in [\underline{x}, \overline{x}] : p(x) = \phi(x)\} \tag{E.14}$$

$$\mathbb{E}_F \left[p(x) \right] = \mathbb{E}_G \left[p(x) \right] \tag{E.15}$$

$$G \text{ first-order stochastically dominates } F \tag{E.16}$$

Then F *solves* [E.13].

Proof. Same as proof of Proposition 1 in Dworczak-Martini. □

Theorem E.1. *Suppose* $\phi(x) = \alpha + \beta \cdot x$ *for* $\alpha > 0$ *and* $\beta < 0$*. Then* ***every*** *solution to program* [E.10] — [E.12] *is "debt." That is, if* F *solves* [E.10]*, it is of the form*

$$F(x) = \begin{cases} G(x), & \text{if } x < \hat{x} \\ 1, & \text{if } x \geqslant \hat{x} \end{cases}$$

for some $\hat{x} \in [\underline{x}, \overline{x}]$.