*American Economic Review: Papers & Proceedings 2014, 104(5): 457–462*
*http://dx.doi.org/10.1257/aer.104.5.457*

## *THEORY OF PERSUASION*[‡]

# Costly Persuasion[†]

*By* Matthew Gentzkow and Emir Kamenica*

In many settings of economic interest, information is ex ante symmetric, but one agent designs the informational environment—i.e., controls what additional information will be generated.

A number of recent papers study such situations, with applications including Internet advertising (Rayo and Segal 2010), communication in organizations (Jehiel 2013), bank regulation (Gick and Pausch 2012; Goldstein and Leitner 2013), medical testing (Schweizer and Szech 2013), medical research (Kolotilin 2013), government control of the media (Gehlbach and Sonin 2013), entertainment (Ely, Frankel, and Kamenica 2013), and price discrimination (Bergemann, Brooks, and Morris 2013).

Identifying the optimal information structure in such settings is a difficult problem if approached by brute force. Given a state space $\Omega$, the set of all[1] information structures, or signals, is as large as $(\Delta(\Omega))^{|\Omega|}$.[2] Moreover, in many applications the objective function is not continuous in the choice of the signal.

Kamenica and Gentzkow (2011)—henceforth, KG—provide a way to simplify the problem of choosing optimal signals. They consider the following model of "Bayesian persuasion." One agent, call him Sender, wishes to persuade another agent, call her Receiver, to change her action. The two agents share a common prior. Sender chooses a *signal* (a map from the true state of the world to a distribution over some signal realization space). Receiver observes the signal realization and takes an action that affects the welfare of both players. Signals are assumed to be costless.

KG simplify Sender's problem by making two observations. First, it is possible to express Sender's payoff as a value function over the posterior belief induced by the signal realization. Second, given any distribution of posteriors whose expectation is the prior, there exists a signal that induces that distribution of posteriors. From these two observations it follows that one can derive the optimal signal from the concavification of Sender's value function.[3]

This concavification approach, however, is not generally feasible if signals are costly. In that case, Sender's payoff is not fully determined by the posterior; given the posterior, the payoff also depends on the signal (due to its cost). Since one cannot express Sender's payoff as a value function over beliefs, the concavification approach cannot be used. All of the aforementioned papers assume costless signals.

The contribution of this paper is to introduce a family of cost functions that is compatible with the concavification approach to deriving the optimal signal. A leading example is cost

[‡] *Discussants:* Navin Kartik, Columbia University; Paul Milgrom, Stanford University; Ilya Segal, Stanford University.

* Gentzkow: University of Chicago Booth School of Business, 5807 S. Woodlawn Avenue, Chicago, IL 60637, and NBER (e-mail: gentzkow@chicagobooth.edu); Kamenica: University of Chicago Booth School of Business, 5807 S. Woodlawn Avenue, Chicago, IL 60637 (e-mail: emir.kamenica@chicagobooth.edu). We are grateful to Ilya Segal and participants at the 2014 ASSA Meetings for helpful comments and suggestions. We thank the Alfred P. Sloan Foundation, as well as the Initiative on Global Markets and the Neubauer Family Foundation, both at the University of Chicago Booth School of Business, for financial support.

[†] Go to http://dx.doi.org/10.1257/aer.104.5.457 to visit the article page for additional materials and author disclosure statement(s).

[1] Brocas and Carrillo (2007) consider a much simpler version of informational design where Sender only chooses how many i.i.d. draws from a particular signal will be generated.

[2] Kamenica and Gentzkow (2011) show that it is without loss of generality to set the cardinality of the signal realization space to be the same as the cardinality of the state space. Then, the set of all signals has the same cardinality as $(\Delta(\Omega))^{|\Omega|}$.

[3] Given a function *f*, its concavification is the smallest concave function everywhere above *f*.

proportional to expected reduction in entropy (Shannon 1948). We thus expand the set of settings where the problem of designing the optimal informational environment is tractable.

## I. The Model

### A. *Costly Signals*

There is a finite state space $\Omega$ with a typical state denoted $\omega$. A *signal* $\pi$ consists of a finite *signal realization* space $S$ and a family of distributions $\{\pi(\cdot|\omega)\}_{\omega\in\Omega}$ over $S$. We denote the cost of a signal $\pi$ by $c(\pi)$.

Given a signal $\pi$ and some prior $\mu$, each signal realization $s$ leads to a posterior belief $\mu_s \in \Delta(\Omega)$. Hence, given a prior $\mu$ each signal $\pi$ induces some distribution of posteriors $\tau \in \Delta(\Delta(\Omega))$. We denote this distribution of posteriors by $\langle\pi|\mu\rangle$.

A function $H : \Delta(\Omega) \to \mathbb{R}_+$ that assigns nonnegative numbers to beliefs is a *measure of uncertainty* if it is concave (Ely, Frankel, and Kamenica 2013). This definition is motivated by Blackwell's (1953) theorem: $\mathbb{E}_{\langle\pi|\mu\rangle}H(\mu_s) \leq H(\mu)$ for all $\pi$ and $\mu$ if and only if $H$ is concave. Hence, assuming that $H$ is concave is equivalent to assuming that receiving information must on average reduce uncertainty.

Our main assumption is that the cost of a signal is proportional to the expected reduction in uncertainty relative to some fixed reference belief:

ASSUMPTION 1: *There exists an interior belief $\mu$ and a measure of uncertainty $H$ such that for all signals $\pi$:*

$$c(\pi) = \mathbb{E}_{\langle\pi|\mu\rangle}[H(\mu) - H(\mu_s)].$$

### B. *Bayesian Persuasion*

Receiver has a continuous utility function $u(a,\omega)$ that depends on her action $a \in A$ and the state of the world. Sender has a continuous utility function $v(a, \omega)$ that depends on Receiver's action and the state of the world. Sender and Receiver share an interior prior $\mu_0$. The action space $A$ is compact.

The game is as follows. Sender chooses a signal $\pi$. Receiver observes Sender's choice of the signal and a signal realization $s \in S$. She then takes her action.[4]

Receiver's payoff is $u(a, \omega)$. Sender's payoff is $v(a, \omega) - c(\pi)$.

We define the *value* of a signal to be Sender's equilibrium payoff if he chooses that signal. The *gain* from a signal is the difference between its value and Sender's equilibrium payoff if he chooses a completely uninformative signal. We say *Sender benefits from persuasion* if there is a signal with a strictly positive gain. A signal is *optimal* if no other signal has a higher value. Clearly, in equilibrium, Sender selects an optimal signal.

## II. Discussion of the Model

The model gives Sender substantial commitment power as it assumes the realization of the signal is truthfully communicated to Receiver. This makes the environment effectively nonstrategic. KG discuss at length various settings where this assumption is suitable. In the interest of space, we do not repeat that discussion here.

Instead, we focus our discussion on Assumption 1. Note that this is a substantive assumption that rules out some reasonable cost functions. For example, suppose that the state space is binary, $\Omega = \{L, R\}$, and that the cost of a signal $\pi(l|L) = \rho_L$, $\pi(r|R) = \rho_R$ is $\rho_L + \rho_R$. In this case, there does not exist a function $H(\cdot)$ and a belief $\mu$ such that $c(\pi) = \mathbb{E}_{\langle\pi|\mu\rangle}[H(\mu) - H(\mu_s)]$.

One natural measure of uncertainty that can serve as a basis for a cost function that satisfies Assumption 1 is entropy: $H(\mu) \equiv -\sum_\omega \mu(\omega) \ln(\mu(\omega))$ (Shannon 1948). In fact, the economics literature on limited attention typically assumes that the cost of processing information is related to expected reduction in entropy. Sims (2003) develops a model where a decision maker faces information-processing limitations that impose a constraint on the expected reduction in entropy. Dessein, Galeotti, and Santos (2013) study organizational design when communication is constrained by a budget

[4] Gentzkow and Kamenica (2012) show that, when Receiver has a unique optimal action at each belief, this game has the same set of equilibrium outcomes as the game where Sender privately observes the signal realization and then sends a verifiable message about the signal realization to Receiver.

of entropy reduction. Martin (2012) considers a model where buyers choose how much information to obtain about the quality of a firm's product and assumes that the cost of information is proportional to the reduction in entropy. Yang (2013) studies coordination games with such costs of information acquisition. Caplin and Dean (2013) derive behavioral implications of entropy-based costs of information-processing and contrast those implications with behavior of subjects in a lab experiment.

While entropy is a natural measure of uncertainty that satisfies a rich set of appealing properties (Cover and Thomas 2006), Assumption 1 also admits many other measures. For instance, residual variance $H(\mu) = \sum_{\omega} \mu(\omega)\,(1-\mu(\omega))$ is an alternative, intuitive measure of uncertainty.

Given any measure of uncertainty $H(\cdot)$ and any affine function $f(\cdot)$, $H' = H + f$ is another measure of uncertainty. Moreover, $\mathbb{E}_{\langle \pi \mid \mu \rangle}[H(\mu) - H(\mu_s)] = \mathbb{E}_{\langle \pi \mid \mu \rangle}[H'(\mu) - H'(\mu_s)]$ for any $\mu$ and $\pi$. Hence, it is helpful to normalize $H(\cdot)$ by setting $H(\mu_s) = 0$ for all degenerate $\mu_s$.[5] With this normalization, there is a unique measure of uncertainty implied by a given cost function.

Finally, note that the belief $\mu$ in the statement of Assumption 1 is not assumed to be $\mu_0$, the prior held by Sender and Receiver. Making the stronger assumption that there exists a measure of uncertainty $H$ such that $c(\pi) = \mathbb{E}_{\langle \pi \mid \mu_0 \rangle}[H(\mu_0) - H(\mu_s)]$ would make our analysis easier, but this stronger assumption would be incompatible with many interpretations of signal costs. In particular, the stronger assumption implies that the cost of a particular signal depends on the prior, i.e., on what previous information was observed. Even the answer to the question of whether one signal or another is more costly could depend on the prior. Thus, if $c(\pi)$ represents some fixed cost of resources required to conduct an experiment that generates $\pi$ (e.g., a drug trial), the stronger assumption is inappropriate. Accordingly, we stipulate a fixed reference belief relative to which the reduction in uncertainty is measured.

## III. Main Result

As KG show, when signals are costless there is a simple way of deriving the optimal signal. Their approach builds on two observations.

First, Sender's payoff is fully determined by the posterior induced by the signal realization. Let $\hat{v}(\mu_s) = \mathbb{E}_{\mu_s}[v(a^*(\mu_s), \omega)]$ where $a^*(\mu_s)$ denotes some selection[6] from $\text{argmax}_{a \in A} \mathbb{E}_{\mu_s} u(a, \omega)$. If the posterior belief is $\mu_s$, Sender's payoff is $\hat{v}(\mu_s)$.

The second observation is that for any $\tau$ such that $\mathbb{E}_{\tau}[\mu_s] = \mu_0$, there exists a $\pi$ such that $\tau = \langle \pi \mid \mu_0 \rangle$. Hence, we can express Sender's problem as

$$(1) \qquad \max_{\tau \text{ s.t. } \mathbb{E}_{\tau}[\mu_s] = \mu_0} [\mathbb{E}_{\tau} \hat{v}(\mu_s)].$$

This problem has a simple geometric interpretation. Let $V$ denote the concavification of $\hat{v}$—the smallest concave function that is everywhere weakly greater than $\hat{v}$. From the formulation of Sender's problem as equation (1), we can see that the value of an optimal signal is $V(\mu_0)$ and that Sender benefits from persuasion if and only if $V(\mu_0) > \hat{v}(\mu_0)$.

We wish to extend this approach to the case where signals are costly. The key obstacle to doing so is the fact that the first observation above—Sender's payoff is fully determined by the posterior—does not necessarily hold for an arbitrary cost function $c(\pi)$. Sender's payoff at a posterior may depend on the signal that induced this belief. The key import of Assumption 1 is that it allows us to represent Sender's payoff from signal $\pi$ as $\mathbb{E}_{\langle \pi \mid \mu_0 \rangle}[\hat{v}_c(\mu_s)]$ where $\hat{v}_c$ denotes value of $\mu_s$ suitably adjusted for the cost of inducing this belief.

At first glance, it may not be obvious that Assumption 1 will suffice for the existence of such a representation. Since the reference belief $\mu$ may be different from $\mu_0$, all that Assumption 1 implies directly is that the payoff from $\pi$ is $\mathbb{E}_{\langle \pi \mid \mu_0 \rangle}[\hat{v}(\mu_s)] - \mathbb{E}_{\langle \pi \mid \mu \rangle}[H(\mu) - H(\mu_s)]$.

[5] Note that both entropy and residual variance satisfy this normalization.

[6] In general, Receiver might have multiple optimal actions at a given belief. Each optimal map from Receiver's belief to her action defines a separate $\hat{v}(\cdot)$ and a separate maximization problem for Sender. Some of these maximization problems may not have a solution and thus Receiver's actions that lead to those cannot be part of an equilibrium. An equilibrium always exists, however, because if Receiver chooses a Sender-optimal action at each belief, $\hat{v}(\cdot)$ is guaranteed to be upper semicontinuous.

Thus, one remaining step in our argument is to show that for any $\mu$ and $H$, there exists a function $\hat{v}_c$ such that $\mathbb{E}_{\langle\pi|\mu_0\rangle}[\hat{v}(\mu_s)] - \mathbb{E}_{\langle\pi|\mu\rangle}[H(\mu) - H(\mu_s)] = \mathbb{E}_{\langle\pi|\mu_0\rangle}[\hat{v}_c(\mu_s)]$ for all $\pi$.

To complete this final step, we draw on an insight from Alonso and Câmara's (2013) extension of KG to a setting with heterogeneous priors. In particular, suppose there are two individuals $a$ and $b$, with interior priors $\mu_0^a$ and $\mu_0^b$, respectively. Suppose we know both individuals observed the same signal realization from the same signal, but we do not know what the signal was or what the signal realization was. Can we determine $b$'s posterior from $a$'s posterior? The answer is yes. In particular, if $a$'s posterior is $\mu_s^a$, $b$'s posterior must be

$$\mu_s^b(\omega) \;=\; \mu_s^a(\omega)\,\frac{\mu_0^b(\omega)/\mu_0^a(\omega)}{\sum_{\omega'}\mu_s^a(\omega')\,\mu_0^b(\omega')/\mu_0^a(\omega')}.$$

Accordingly, we can define a function $m(\cdot\,|\,\mu_0^a;\,\mu_0^b)$ such that if an agent with the prior $\mu_0^a$ has the posterior $\mu_s^a$, then an agent with the prior $\mu_0^b$ has the posterior $m(\mu_s^a\,|\,\mu_0^a;\,\mu_0^b)$.

Given any cost function $c(\pi)$ that satisfies Assumption 1, let $\hat{v}_c(\mu_s) = \hat{v}(\mu_s) - [H(\mu) - H(m(\mu_s\,|\,\mu_0,\,\mu))]$. We then have that $\mathbb{E}_{\langle\pi|\mu_0\rangle}[\hat{v}_c(\mu_s)] = \mathbb{E}_{\langle\pi|\mu_0\rangle}[\hat{v}(\mu_s)] - c(\pi)$ for all $\pi$. Let $V_c$ be the concavification of $\hat{v}_c$. We then have our main result:

PROPOSITION 1: *Suppose the cost function satisfies Assumption A1. The value of an optimal signal is $V_c(\mu_0)$. Sender benefits from persuasion if and only if $V_c(\mu_0) > \hat{v}_c(\mu_0)$.*

The main implication of Proposition 1 is that one can derive the optimal signal by drawing the value function $\hat{v}_c(\cdot)$ and its concave closure $V_c(\cdot)$ and then "reading off" the optimal $\tau$ from the picture.

For example, if $\hat{v}_c(\cdot)$ has the shape as in Figure 1, the optimal $\tau$ induces $\mu_L$ and $\mu_R$. Given the optimal $\tau$, the optimal $\pi$ is determined by the following equation:

$$(2)\qquad \pi(s\,|\,\omega) \;=\; \frac{\mu_s(\omega)\,\tau(\mu_s)}{\mu_0(\omega)},$$

which implies that $\langle\pi\,|\,\mu_0\rangle = \tau$.

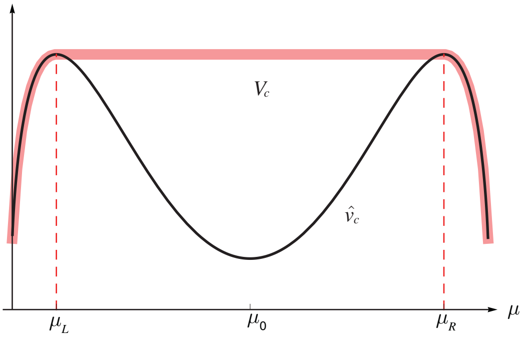

FIGURE 1. GEOMETRIC DERIVATION OF THE OPTIMAL SIGNAL

## IV. Example

To illustrate the main result above, we consider an extension of the motivating example in KG. A prosecutor (Sender) is trying to convince a judge that a defendant is guilty. The judge (Receiver) chooses whether to *acquit* or *convict* the defendant. There are two states of the world: the defendant is either *guilty* or *innocent*. The judge gets utility 1 for choosing the just action (convict when guilty and acquit when innocent) and utility 0 for choosing the unjust action. The prosecutor gets utility 1 if the judge convicts and utility 0 if the judge acquits (minus the signal cost), regardless of the state. The prosecutor and the judge share a prior belief $\Pr(guilty) = 0.3$.

The prosecutor conducts an investigation and is required by law to report its full outcome. One can think of the investigation as a choice of how to structure the arguments, whom to subpoena, what forensic tests to conduct, etc. Formally, an investigation is a signal $\pi$ that specifies distributions $\pi(\cdot\,|\,guilty)$ and $\pi(\cdot\,|\,innocent)$ on signal realizations $\{i, g\}$. The cost of investigation $\pi$ is $k\mathbb{E}_{\langle\pi|\mu_*\rangle}[H(\mu_*) - H(\mu_s)]$ where $H$ denotes entropy $(H(\mu) \equiv -\sum_\omega \mu(\omega)\,\ln(\mu(\omega)))$, $\mu_*$ denotes the uniform belief $(\mu_*(guilty) = \mu_*(innocent) = \frac{1}{2})$, and $k \geq 0$ is a cost parameter.

What is the prosecutor's optimal investigation? If he conducts no investigation ($\pi$ is perfectly uninformative), his payoff is zero because the judge acquits under her prior (and the cost of a completely uninformative signal is zero). A very informative investigation might be overly costly but is suboptimal even when signals are costless. In fact, as KG show, when $k = 0$, the optimal investigation is partially informative with

$$\pi(i \,|\, innocent) = \tfrac{4}{7} \qquad \pi(i \,|\, guilty) = 0$$

$$\pi(g \,|\, innocent) = \tfrac{3}{7} \qquad \pi(g \,|\, guilty) = 1,$$

which yields a payoff of 0.6 to the prosecutor. Panel A of Figure 2 shows the value function and its concavification under costless signals.

Panel B of the figure depicts $\hat{v}_c(\cdot)$ and its concavification when $k = 2$. In this case, the optimal investigation induces beliefs $\mu_i = 0.15$ and $\mu_g = \frac{1}{2}$. Since $\mathbb{E}_\tau[\mu_s] = 0.3$, we know that $\tau(\mu_i) = 0.57$ and $\tau(\mu_g) = 0.43$. Thus, applying equation (2), we derive the optimal signal as

$$\pi(i \,|\, innocent) = 0.69 \quad \pi(i \,|\, guilty) = 0.28$$

$$\pi(g \,|\, innocent) = 0.31 \quad \pi(g \,|\, guilty) = 0.72.$$

Since $\tau(\mu_g) = 0.43$, the prosecutor induces conviction in 43 percent of the cases. Note that the costs reduce the likelihood of conviction because a definitive proof of *innocence* has become prohibitively costly (which in turn increases the probability that innocence is indicated by the signal realization.)

If investigations are very costly, e.g., $k = 10$, the optimal choice is a completely uninformative investigation which yields a payoff of zero to the prosecutor. Panel C of Figure 2 depicts the value function when $k = 10$. The concavification coincides with the value function at the prior, so the prosecutor cannot benefit from conducting an investigation.

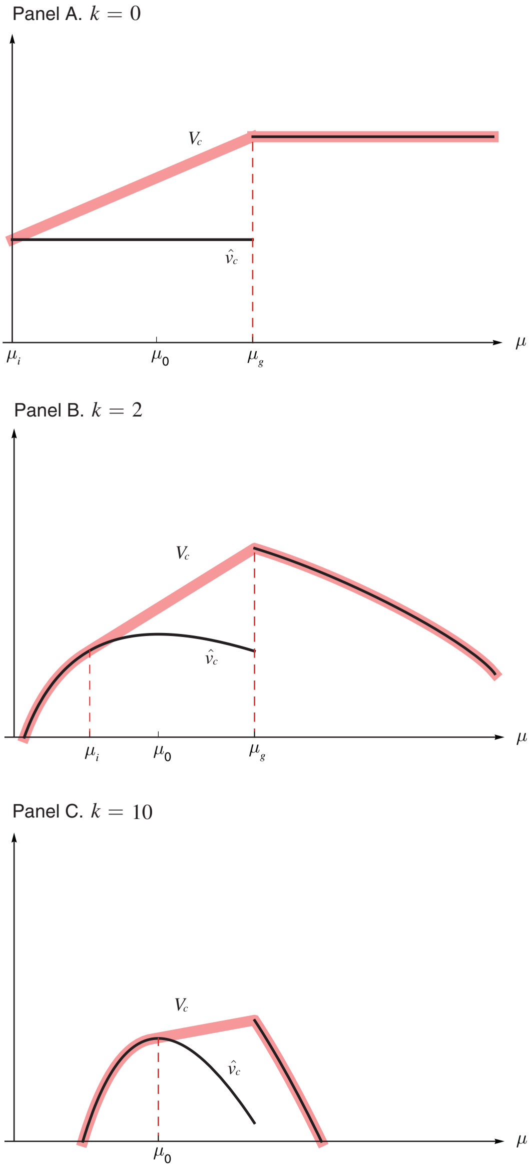

Figure 2. Optimal Investigations

## V. Comparative Statics

This example above illustrates how the concavification approach can be used to solve for the equilibrium even when signals are costly. It also illustrates some implications of the magnitude of signal costs. First, the optimal signal under $k = 0$ is Blackwell more informative than the optimal signal when $k = 2$, which is in turn Blackwell more informative than the uninformative signal that arises when $k = 10$. In fact, it is easy to see that in this example, a lower $k$ always leads to a Blackwell more informative signal. Second, since the optimal signal never induces a belief that would make Receiver strictly prefer a nondefault action, Receiver's payoff is unaffected by $k$. Finally, Sender's payoff decreases as the cost of signals increases.

In this final section, we consider the extent to which these comparative statics hold in general. Specifically, consider the general version of the model (with an arbitrary state space, action space, preferences, and prior) and suppose that $c(\pi) = k\mathbb{E}_{\langle\pi|\mu\rangle}[H(\mu) - H(\mu_s)]$ for some reference belief $\mu$ and some measure of uncertainty $H$. How do outcomes vary with the cost parameter $k$?

It is easy to see that as $k$ increases, Sender's payoff must decrease. In fact, even if we ignore the signal-cost component of Sender's payoff, the expected value of $v(a, \omega)$ is weakly lower when $k$ is higher.

It is less clear how Receiver's payoff varies with $k$. Receiver's payoff can clearly decrease when signals become more expensive. For example, if $u = v$, Sender and Receiver's payoffs are perfectly aligned, so higher $k$ must reduce Receiver's payoff. It is also possible, however, for Receiver's payoff to strictly increase when signals become more costly.

The last observation implies that, unlike in the example above, lower costs do not generally induce Blackwell more informative signals. That said, the concavity of $H$ implies that if there is a uniquely optimal signal $\pi_l$ when the cost parameter is $k_l$ and a distinct uniquely optimal signal $\pi_h$ when the cost parameter is $k_h > k_l$, it cannot be the case that $\pi_h$ is Blackwell more informative than $\pi_l$.[7]

## REFERENCES

**Alonso, Ricardo, and Odilon Câmara.** 2013. "Persuading Skeptics and Reaffirming Believers." Unpublished.

**Bergemann, Dirk, Benjamin Brooks, and Stephen Morris.** 2013. "The Limits of Price Discrimination." Unpublished.

**Blackwell, David.** 1953. "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics* 24 (2): 265–72.

**Brocas, Isabelle, and Juan D. Carrillo.** 2007. "Influence Through Ignorance." *RAND Journal of Economics* 38 (4): 931–47.

**Caplin, Andrew, and Mark Dean.** 2013. "Rational Inattention, Entropy, and Choice: The Posterior-based Approach." Unpublished.

**Cover, Thomas M., and Joy A. Thomas.** 2006. *Elements of Information Theory*. 2nd ed. Hoboken: Wiley-Interscience.

**Dessein, Wouter, Andrea Galeotti, and Tano Santos.** 2013. "Rational Inattention and Organizational Focus." Unpublished.

**Elliott, Matthew, Benjamin Golub, and Andrei Kirilenko.** 2012. "How Better Information Can Garble Experts' Advice." Unpublished.

**Ely, Jeffrey, Alexander Frankel, and Emir Kamenica.** 2013. "Suspense and Surprise." Unpublished.

**Gehlbach, Scott, and Konstantin Sonin.** 2013. "Government Control of the Media." Unpublished.

**Gentzkow, Matthew, and Emir Kamenica.** 2012. "Disclosure of Endogenous Information." Unpublished.

**Gick, Wolfgang, and Thilo Pausch.** 2012. "Persuasion by Stress Testing: Optimal Disclosure of Supervisory Information in the Banking Sector." Unpublished.

**Goldstein, Itay, and Yaron Leitner.** 2013. "Stress Tests and Information Disclosure." Unpublished.

**Jehiel, Philippe.** 2013. "On Transparency in Organizations." Unpublished.

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

**Kolotilin, Anton.** 2013. "Experimental Design to Persuade." Unpublished.

**Martin, Daniel.** 2012. "Strategic Pricing with Rational Inattention to Quality." Unpublished.

**Rayo, Luis, and Ilya Segal.** 2010. "Optimal Information Disclosure." *Journal of Political Economy* 118 (5): 949–87.

**Schweizer, Nikolaus, and Nora Szech.** 2013. "Optimal Revelation of Life-changing Information." Unpublished.

**Shannon, Claude E.** 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

**Sims, Christopher A.** 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.

**Yang, Ming.** 2013. "Coordination with Flexible Information Acquisition." Unpublished.

[7] In contrast, Elliott, Golub, and Kirilenko (2012) construct a model with multiple senders where senders' access to more informative signals can reduce the amount of information revealed in equilibrium.