Non-Bayesian Persuasion*

Geoffroy de Clippel† Xu Zhang‡

March 4, 2020

Abstract

Following Kamenica and Gentzkow (2011), the paper studies persuasion as an information design problem. We investigate how mistakes in probabilistic inference impact optimal persuasion. The 'concavification method' is shown to extend naturally to a large class of belief updating rules, which we identify and characterize. This class comprises many non-Bayesian models discussed in the literature. We apply this new technique to gain insight in a variety of questions about, among other things, the revelation principle, the ranking of updating rules, and circumstances under which persuasion is beneficial. Our technique extends to also shed light on the question of robust persuasion.

*We thank Pedro Dal Bó, Kfir Eliaz, Ignacio Esponda, Erik Eyster, Jack Fanning, Teddy Mekonnen, Xiaosheng Mu, Bobby Pakzad-Hurson, Alex Poterack, Louis Putterman, Kareen Rozen, Roberto Serrano, Alireza Tahbaz-Salehi, Neil Thakral and seminar participants at the Brown University Theory Seminar for helpful comments.

†Department of Economics, Brown University. geoffroy_declippel@brown.edu.

‡Department of Economics, Brown University. xu_zhang@brown.edu.

1 Introduction

The past decade has seen much progress on the topic of information design. In a seminal paper, Kamenica and Gentzkow (2011), henceforth KG, studies how a rational agent (Receiver, She) can be persuaded to take a desired action by controlling her informational environment. They provide tools to determine when persuasion is profitable, and how to best persuade. They also illustrate how these techniques provide valuable insights in a variety of applications.

The purpose of the present paper is to expand the analysis to accommodate agents who make mistakes in probabilistic inference. Experimental evidence (Camerer 1998, Benjamin 2019) shows that people oftentimes systematically depart from Bayes' rule when confronted with new information. Though our analysis easily extends to other contexts as well, we find it more natural as a benchmark to keep assuming the persuader (Sender, He) is Bayesian.[1] After all, a person who exerts effort to figure out the best way to persuade is also likely to make an effort to assess probabilities accurately.

Example scenarios from KG naturally extend to our setting. Suppose a rational doctor has a patient's best interest at heart, but knows that his patient is too conservative in her updated belief upon hearing bad news. Which tests should he run to optimally acquire information and persuade the patient to undergo surgery when necessary? Consider now a prosecutor trying to maximize his conviction rate. How should he conduct his investigation when facing a judge suffering from base-rate neglect (Kahneman and Tversky 1973)?[2] A rational firm may strategize when supplying product information to prospective buyers. How can it best exploit customers who have a favorable bias towards the trademark when processing information? To what extent does it remain possible to persuade other customers who have an unfavorable updating bias? We address these questions, and others discussed below, by extending KG's main results from Bayesian persuasion to our setting. The paper thus also speaks to the robustness of their results against a richer, sometimes more realistic class of updating rules.

Receiver's action is determined by her belief. Hence the first step in understanding the limits of persuasion is to figure out how signals (or experiments) impact Receiver's belief. Under Bayesian updating, as in KG, a distribution of posteriors is achievable by some signal if, and only if, it satisfies the martingale property (the expectation of the posteriors matches the prior). This is the key observation that leads to the now-classic concavification argument to derive optimal persuasion value and strategies. A first question then is whether comparable characterization results obtain when Receiver is not Bayesian. Furthermore, Sender and Receiver will now typically have different posteriors. This raises new, interesting questions when Sender's utility is state-dependent, as her preferred action also varies with information revealed by the experiment. Thus

[1] Sender's updating rule does not even matter if his utility is state-independent.

[2] Evidence of judges' mistakes in statistical inference, including base-rate neglect, is provided in Guthrie, Rachlinski, and Wistrich (2001, 2007), Lindsey, Hertwig, and Gigerenzer (2002), Koehler (2002), Danziger, Levav, and Avnaim-Pesso (2011), and Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2017).

we must not only characterize distributions of Receiver's posteriors induced by experiments, but distributions over Sender-Receiver posterior *pairs* (see Section 2).

After obtaining such characterization results for a couple of classic non-Bayesian updating rules, we discovered a common, overarching methodology, which we present in Section 3. Let μ_0 be the common prior (our approach also applies to non-common priors, see below). As we will show, for many updating rules commonly used in the literature, there exists a *distortion function* D_{μ_0} mapping beliefs into beliefs, such that the updated belief after receiving a signal realization simply corresponds to the distortion of the accurate, Bayesian posterior ν (independently of other realizations that could arise). In other words, the Sender-Receiver posterior pair is simply $(\nu, D_{\mu_0}(\nu))$ in that case. These rules are said to *systematically distort updated beliefs*. KG's concavification argument for optimal persuasion extends to such rules with Sender's indirect utility for Bayesian posteriors modified by factoring in the distortion function.

Though straightforward once the idea of a distortion function has been recognized, this result is nonetheless powerful as it turns out that many rules commonly used in the literature do systematically distort updated beliefs in the above sense. To better gauge their prevalence and understand what they entail, Section 4 starts with a useful characterization of these rules. It then provides a variety of examples encompassed by our notion, including conservative Bayesianism (Edwards, 1968), Grether (1980)'s $\alpha - \beta$ model, Cripps' (2018) divisible updating rules, and other forms of confirmatory bias and motivated beliefs. As it turns out, our methodology also applies to situations with a Bayesian Receiver who either distorts probabilities when making decisions (a key ingredient in prospect theory), or has a different prior from Sender (a problem first studied in Alonso and Câmara (2016)). To complete the picture of the power and limits of our approach,[3] Section 4 ends with a few 'non-examples', that is, updating rules that do not systematically distort updated beliefs.

Sender faces a moral hazard problem: he has to provide the right incentives for Receiver to take the desired action. The revelation principle (see e.g. Section 6 in Myerson (1991)) applies when Receiver is Bayesian: each signal admits an outcome-equivalent incentive compatible signal where realizations are suggested actions. Section 5 starts with an example showing that the revelation principle need not hold when Receiver is subject to mistakes in probabilistic inference. Within the class of rules that systematically distort updated belief, we then essentially characterize those for which the revelation principle holds. The key property is that the distortion function maps straight lines into straight lines (as do affine functions, or more generally projective transformations). As a corollary, if optimal persuasion is achievable (which it is under some mild continuity assumption), then it can be achieved with a number of signal realizations no greater than the number of actions. This result breaks down for distortion functions that do not map straight lines into straight lines. Even then, no more than n signal realizations are needed for optimal persuasion, where n is the number of states,

[3]We should point out, however, that many conceptual contributions of the paper, in the questions we raise and the definitions we introduce to tackle them, are valid beyond the class of rules we focus on to answer them.

for rules that systematically distort updated beliefs.

KG highlights the surprising fact that effective persuasion is possible in many problems despite the fact that Receiver is rational and knows the Sender's intent to persuade her to take an action to his advantage. One may think at first that persuasion gets only easier if Receiver lacks rationality when processing information. Quickly one realizes though that departures from Bayes' rule need not equate to easier manipulation. For an extreme example, the worst for Sender is a stubborn Receiver who never updates her belief. Of course, persuasion is sometimes easier: the best for Sender is a totally gullible Receiver who adopts any suggested belief. This raises interesting, related questions. When is effective persuasion possible away from Bayesian updating? How does Sender fare as a function of the updating rule? Are some updating rules preferable to others? We tackle these questions in Sections 6 and 7.

Of course, by offering a method to compute Sender's optimal payoff, our extension of KG's concavification argument provides a tool for addressing such questions in any given persuasion problem. When ranking updating rules, however, more robust comparisons – those that hold for a large class of persuasion problems sharing a common information structure – can be more informative of the way specific belief updating rules systematically impact optimal persuasion. We carefully develop these ideas in Section 6 by introducing and comparing different definitions, and illustrate the concepts by means of examples. Then we prove that, perhaps surprisingly, no two rules that systematically distort updated beliefs (and whose distortion function is invertible) can be unambiguously compared when permitting all payoff structures. In particular, Bayesian updating is neither systematically superior, nor systematically inferior to any of these rules. Unambiguous comparisons between these rules may sometimes be possible, however, when restricting attention to interesting subclasses of problems (e.g., state-independent utility for Sender).

Effective persuasion is possible if, and only if, there is a signal that gives Sender a strictly larger payoff than with Receiver's default action (optimal for the prior). Thanks to our earlier result, this can be determined, for rules that systematically distort updated beliefs, by checking the value of a concavified function. However, given that concavifying a function can be hard, it may be worthwhile (as in KG) to provide simpler, necessary and sufficient conditions for when effective persuasion is possible. Section 7 speaks to the robustness of KG's Proposition 2 in that regard, as the result extends *verbatim* to most rules that systematically distort updated beliefs. Of course, their property 'there is information that Sender would share' must now take into account that Receiver's posterior after processing that information is a distortion of the Bayesian posterior.

Given the difficulty of concavifying general functions in the presence of multiple states, effort has been devoted to better understand the simpler case where Receiver's optimal action is measurable with respect to the state's expected value (see, e.g., KG, Kamenica and Gentzkow (2016), and Dworczak and Martini (2019)). We show in Section 8 that these results extend for updating rules associated to affine distortion functions. Indeed, the original persuasion problem can then be proved to be equivalent

to a Bayesian persuasion problem where Receiver's utility is distorted. We use this result to illustrate how mistakes in probabilistic inferences may impact optimal persuasion and consumer welfare when a firm tries to persuade a customer to buy its product (a simple model proposed by KG to illustrate their theory of Bayesian persuasion). Interestingly, the customer may actually fare better on average (in actual terms instead of perceived utility) when she suffers from an unfavorable updating bias towards the product.

Section 9 concludes by highlighting how our methodology for dealing with non-Bayesian updating extend in two directions. First, optimal persuasion value remains computable by concavification in problems where Sender views different distortion functions as possible, and attaches positive probabilities to multiple updating rules. Second, by extending the notion of distortion function to accommodate correspondences, we can capture an idea of robust persuasion. Suppose, for instance, Sender is concerned that Receiver's posterior may fall in a neighborhood of the Bayesian posterior. Fearing the worst, his indirect utility for that Bayesian posterior is now his minimal utility over Receiver's preferred actions for all beliefs in that neighborhood. Once again, Sender's optimal persuasion value remains computable by concavification, and represents a guaranteed level of profit despite the risk of Receiver's mistakes.

Related Literature

Alonso and Câmara (2016) investigate optimal persuasion for a *Bayesian* Receiver who has a different (full-support) prior than Sender. Galperti (2019) models changing worldviews, which also entail non-common priors, but with different supports. Receiver always adopts the same full-support prior (which is known to Sender and independent of the signal) after unexpected evidence and then updates based the new prior following Bayes' rule. Receivers in those papers could alternatively be interpreted as *non-Bayesian with respect to Sender's prior*, providing neat examples of rules that systematically distort updated beliefs (see Example 5). Levy, de Barreda, and Razin (2018a, 2018b) study persuasion with a Receiver who suffers from correlation neglect and prove that Sender can achieve close to his first-best in that case. We show that such updating rules do not systematically distort updated beliefs.

While we take the updating rule as an exogenous primitive capturing how Receiver performs probabilistic inferences, some endogenize it. Bloedel and Segal (2018), Lipnowski, Mathevet, and Wei (forthcoming), and Wei (2018) study optimal persuasion when Receiver is rationally inattentive — i.e., Receiver is Bayesian but rationally decides how much information to acquire. Such a Receiver would appear as making mistakes in probabilistic inferences if one ignored the cost she faces in acquiring information. Given her utility function, one can define an updating rule, which systematically distorts updated beliefs if the attention cost function is posterior-separable (Caplin, Dean, and Leahy 2019) as in Lipnowski et al. (forthcoming) and Wei (2018). Eliaz, Spiegler, and Thysen (2019) studies strategic interpretations where Receiver's capability of interpreting equilibrium messages is restricted by a "dictionary" chosen by Sender. By choosing different dictionaries, Sender essentially manipulates Receiver's

updating rule, which enhances his persuasiveness to the point that full persuasion is sometimes possible.

There is a contemporary effort to incorporate non-Bayesian updating in other models of communication. Lee, Lim, and Zhao (2019) investigate the implications of conservative Bayesianism (or 'prior-biased inferences' in their term) in cheap talk problems. Benjamin, Bodoh-Creed, and Rabin (2019) analyze an example where an informed persuader chooses whether or not to reveal a verifiable signal to an audience who suffers from base-rate neglect. In contrast, we study a wide range of updating rules with a communication protocol where Sender has full commitment power.

The paper fits a broader effort in the literature to accommodate features of behavioral economics in mechanism design. de Clippel (2014) studies implementation when individual choices need not be compatible with the maximization of a preference ordering. Crawford (2019), de Clippel, Saran, and Serrano (2019) and Kneeland (2019) study mechanism design with agents who need not have rational expectations. The present paper pursues this broad effort by investigating the implications of mistakes in probabilistic inferences, this time in an information design problem.

Our paper also relates to the axiomatization of non-Bayesian updating. Recent attempts in this direction include Epstein (2006), Epstein, Noor, and Sandroni (2008), Lehrer and Teper (2016), Zhao (2016), Augenblick and Rabin (2018), Cripps (2018), and Chauvin (2019). Although we do not propose any axiom, the property of systematically distorting updated beliefs can help classify different updating rules.

2 General Framework

As in KG, a state ω is drawn at random according to a full support distribution μ_0 on a finite set Ω. Sender (he) and Receiver (she) are both expected utility maximizers with continuous von Neumann-Morgenstern utility functions $v(a, \omega)$ and $u(a, \omega)$, where a is Receiver's chosen action from a compact set A. Neither player knows the state, but Sender can costlessly choose a signal π, which consists of a finite realization space S and a family of distributions $\{\pi(\cdot|\omega)\}_{\omega\in\Omega}$ over s. It will always be assumed that S does not contain any redundant signal, that is, each signal in S occurs with strictly positive probability under π. Upon observing the realization s, Sender correctly updates his belief by applying Bayes' rule:

$$\mu_s^B(\omega; \mu_0, \pi) = \frac{\pi(s|\omega)\mu_0(\omega)}{\sum_{\omega'\in\Omega} \pi(s|\omega')\mu_0(\omega')}. \tag{1}$$

In contrast to KG, Receiver may make mistakes in probabilistic inferences. Her posterior is denoted $\mu_s^R(\omega; \mu_0, \pi)$.

Given the prior μ_0 and Receiver's updating rule μ^R, signal π generates a distribution $\tau \in \Delta(\Delta(\Omega) \times \Delta(\Omega))$ over pairs of Sender-Receiver posteriors. The pair (ν, ν') occurs with probability $\sum_{s\in S(\nu,\nu')} \sum_\omega \pi(s|\omega)\mu_0(\omega)$, where $S(\nu, \nu')$ is the set of signal

realizations s such that $\nu = \mu_s^B(\,.\,;\mu_0,\pi)$ and $\nu' = \mu_s^R(\,.\,;\mu_0,\pi)$. Let $T(\mu_0,\mu^R)$ denote the set of all such distributions obtained by varying π.

Given belief ν', Receiver picks an optimal action

$$\hat{a}(\nu') \in \arg\max_{a\in A} E_{\nu'} u(a,\omega),$$

and we assume that $\hat{a}(\nu')$ maximizes Sender's expected utility whenever Receiver is indifferent between some actions at ν'. To figure out the optimal signal, Sender aims to solve the following optimization problem:

$$V(\mu_0,\mu^R) = \sup_{\tau\in T(\mu_0,\mu^R)} E_\tau \hat{v} = \sup_{\tau\in T(\mu_0,\mu^R)} \sum_{(\nu,\nu')\in\, supp(\tau)} \tau(\nu,\nu')\hat{v}(\nu,\nu'), \tag{2}$$

where

$$\hat{v}(\nu,\nu') = \sum_\omega \nu(\omega) v(\hat{a}(\nu'),\omega)$$

represents Sender's utility should the posteriors be ν for himself and ν' for the Receiver.

Remark 1. *The problem further simplifies should Sender's utility be state-independent. Indeed, $\hat{v}(\nu,\nu') = v(\hat{a}(\nu'))$ is then independent of ν, which will be denoted $\hat{v}(\nu')$. Only marginal distributions of Receiver's posteriors matter. Let $T^R(\mu_0,\mu^R)$ be the set of distributions $\tau^R \in \Delta(\Delta(\Omega))$ for which there exists $\tau \in T(\mu_0,\mu^R)$ such that $\tau^R(\nu') = \sum_{(\nu,\nu')\in\, supp(\tau)} \tau(\nu,\nu')$, for each ν' in the support of τ^R. Then*

$$V(\mu_0,\mu^R) = \sup_{\tau^R\in T^R(\mu_0,\mu^R)} E_{\tau^R}\hat{v} = \sup_{\tau^R\in T^R(\mu_0,\mu^R)} \sum_{\nu'\in\, supp(\tau^R)} \tau^R(\nu')\hat{v}(\nu'). \tag{3}$$

3 Simplifying the Problem: An Interesting Class of Updating Rules

As is clear from (2) and (3), a critical step for computing the Sender's optimal signal is to gain a better understanding of the sets $T(\mu_0,\mu^R)$ and $T^R(\mu_0,\mu^R)$. A key insight in KG is that, should Receiver be rational, a distribution τ^R of posteriors can arise if, and only if, it is *Bayes-plausible*,[4] that is,

$$\sum_{\nu\in Supp(\tau^R)} \nu\tau^R(\nu) = \mu_0.$$

Much like the revelation principle, this characterization greatly simplifies the Sender's problem as one doesn't need to worry about the multitude of possible signals, but only about the much simpler space of Bayes-plausible distributions of posteriors.

[4]See also Shmaya and Yariv (2009).

Suppose now the Receiver is prone to mistakes in probabilistic inference: $\mu^R \neq \mu^B$. Can one find a similar, simple characterization of the set $T^R(\mu_0, \mu^R)$ of distributions over posteriors? More generally, can one find a simple characterization of $T(\mu_0, \mu^R)$, which also expresses how the Receiver's posterior varies as a function the Sender's rational posterior? By 'simple', we mean characterization results that, like Bayes-plausibility, can be expressed in terms of distributions over posteriors, with no reference to signals. As we will confirm below, this would allow us to extend the tractable techniques identified in KG to solve for optimal persuasion.

Say that μ^R *systematically distorts updated beliefs* if, for all full-support prior μ_0, there exists a *distortion function* $D_{\mu_0} : \Delta(\Omega) \to \Delta(\Omega)$ such that, for all signal π and all signal realizations s, $\mu_s^R(\ .\ ; \mu_0, \pi) = D_{\mu_0}(\mu_s^B(\ .\ ; \mu_0, \pi))$. As desired, the function D_{μ_0} expresses the relationship between the Receiver's posterior and the Sender's rational one in a systematic way that is *independent of the signal*.

Remark 2. *If, in addition, the distortion functions are invertible (which, as we will see, is the case in many examples, but is not needed for our analysis), then the set $T^R(\mu_0, \mu^R)$ of distributions τ^R over Receiver's posteriors that can arise under μ^R are characterized by the following* distorted Bayes-plausibility *condition:*

$$\sum_{\mu \in Supp(\tau^R)} D_{\mu_0}^{-1}(\mu)\tau^R(\mu) = \mu_0.$$

Indeed, Receiver has a posterior μ with probability $\tau^R(\mu)$ if, and only if, the rational posterior is $D_{\mu_0}^{-1}(\mu)$. The characterization result then follows from the characterization of rational distributions over posteriors through Bayes-plausibility.

Suppose that μ^R systematically distorts updated beliefs with distortion functions $(D_{\mu_0})_{\mu_0 \in \Delta(\Omega)}$. Then, for all priors μ_0, Sender can generate a distribution τ over pairs of posteriors (ν, ν') (that is, $\tau \in T(\mu_0, \mu^R)$) if, and only if, the marginal of τ on the first component is Bayes-plausible and $\nu' = D_{\mu_0}(\nu)$ for all (ν, ν') in the support of τ. Thus

$$V(\mu_0, \mu^R) = \sup_{\rho \text{ Bayes-plausible}} \sum_{\nu \in\ supp(\rho)} \rho(\nu)\check{v}(\nu), \tag{4}$$

where

$$\check{v}(\nu) = \hat{v}(\nu, D_{\mu_0}(\nu)).^5 \tag{5}$$

Finding Sender's best signal is thus equivalent to finding the best signal under rationality provided one uses the distorted indirect utility function $\check{v}$ defined over Bayesian posteriors. The following result then follows from Kamenica and Gentzkow (2011)'s Corollary 2. For each function $f : \Delta(\Omega) \to \mathbb{R}$, let f's *concavification* (Aumann and Maschler 1995), denoted $CAV(f)$, be the smallest concave function that is everywhere weakly greater than f:

$$[CAV(f)](\mu) = \sup\{z \mid (\mu, z) \in co(f)\},$$

[5]For notational simplicity, we do not label $\check{v}$ with D_{μ_0} while keeping in mind that it depends on the distortion function and potentially the prior.

where $co(f)$ denotes the convex hull of the graph of f.

Proposition 1. *Suppose that μ^R systematically distorts updated beliefs with distortion functions $(D_{\mu_0})_{\mu_0 \in \Delta(\Omega)}$. The value of an optimal signal for the common prior μ_0 is $[CAV(\check{v})](\mu_0)$. Sender benefits from persuasion if, and only if, $[CAV(\check{v})](\mu_0) > \hat{v}(\mu_0, \mu_0)$.*

Unlike in Bayesian persuasion, to see if Sender benefits from persuasion, we need to compare $[CAV(\check{v})](\mu_0)$ to $\hat{v}(\mu_0, \mu_0) = \sum_\omega \mu_0(\omega) v(\hat{a}(\mu_0), \omega)$, rather than $\check{v}(\mu_0) = \sum_\omega \mu_0(\omega) v(\hat{a}(D_{\mu_0}(\mu_0)), \omega)$. The former is Sender's default payoff with no persuasion while the latter represents his payoff from sending an uninformative signal. The two coincide if Receiver is Bayesian, but may differ if Sender's belief can be modified by an uninformative signal ($D_{\mu_0}(\mu_0) \neq \mu_0$).

4 Characterization and Examples

The next result offers a characterization of all updating rules that systematically distort updated beliefs. We will see afterwards that multiple non-Bayesian updating rules discussed in the literature have that property.

Proposition 2. *The updating rule μ^R systematically distorts updated beliefs if, and only if, given any full-support prior μ_0, $\mu_s^R(\,.\,; \mu_0, \pi) = \mu_{\hat{s}}^R(\,.\,; \mu_0, \hat{\pi})$ for all signal-realization pairs (π, s) and $(\hat{\pi}, \hat{s})$ such that the likelihood ratio $\frac{\hat{\pi}(\hat{s}|\omega)}{\pi(s|\omega)}$ is constant as a function of ω.*[6]

In that case, the distortion function D_{μ_0} is uniquely defined as follows:

$$D_{\mu_0}(\mu) = \mu_{\hat{s}}^R(\,.\,; \mu_0, \hat{\pi}_\mu),$$

for all $\mu \in \Delta(\Omega)$, where $\hat{\pi}_\mu$ is any signal giving the realization $\hat{s}$ with probability $\hat{\pi}_\mu(\hat{s}|\omega) = \frac{\mu(\omega)}{\mu_0(\omega)} \min_{\omega'} \frac{\mu_0(\omega')}{\mu(\omega')}$, for all $\omega \in \Omega$.

Let's pause a moment and have a second look at the necessary and sufficient condition identified in Proposition 2. A first requirement is that the Receiver's updated belief after a signal realization should be independent of the label used to describe that realization and the set of other realizations that could have occurred. All that matters is the likelihood of getting that signal realization as a function of the different states. More substantially, the updated belief should remain unchanged when rescaling those probabilities by a common factor, a property of homogeneity of degree zero.

It is easy to provide some further intuition for Proposition 2. Let's restrict attention, for simplicity, to signals π such that $\pi(s|\omega) > 0$ for all s and all ω. It follows from (1) that

$$\frac{\mu_s^B(\omega; \mu_0, \pi)}{\mu_s^B(\omega'; \mu_0, \pi)} = \frac{\pi(s|\omega)}{\pi(s|\omega')} \frac{\mu_0(\omega)}{\mu_0(\omega')},$$

[6] With the convention that $\frac{0}{0}$ can be set to any desired value, that is, no restriction is imposed for states ω such that $\pi(\hat{s}|\omega) = \pi(s|\omega) = 0$.

for all s, ω, ω'. Hence, given μ_0, Bayes' rule defines a bijection (independent of π and s) between posterior μ_s^B and the set of all likelihood ratios $\{\frac{\pi(s|\omega')}{\pi(s|\omega)} \mid \omega, \omega' \in \Omega\}$. If μ^R also defines a bijection between this set and the posteriors it generates, then we have a bijection (independent of π and s) between Receiver's and Sender's posteriors. In that case, if $\frac{\pi(s|\omega')}{\pi(s|\omega)} = \frac{\hat{\pi}(\hat{s}|\omega')}{\hat{\pi}(\hat{s}|\omega)}$ for all ω, ω', then Receiver's belief given the realization s for the signal π must coincide with her belief given the realization $\hat{s}$ for the signal $\hat{\pi}$. In the appendix, we provide a formal proof of Proposition 2, which in addition does not rely on distortion functions to be bijections, and accommodates signal realizations that have zero probability under some state.

4.1 Examples

Many non-Bayesian models in the literature systematically distort updated beliefs. Importantly, for each of the examples below, we indicate the associated distortion functions.

Example 1 (Affine Distortion)**.** *Under this class of updating rules, Receiver's posterior falls in between a given belief $\nu^* \in \Delta(\Omega)$ and the correct, Bayesian posterior:*

$$\mu_s^R(\ .\ ; \mu_0, \pi) = \chi\nu^* + (1-\chi)\mu_s^B(\ .\ ; \mu_0, \pi)$$

where $0 \leq \chi \leq 1$ is a constant parameter. This rule matches the simple, tractable functional form Gabaix (2019) suggests to unify various aspects of behavioral economics. A larger χ means moving further away from Bayesian updating, and the tendency to update towards ν^* becomes stronger. Clearly, the updating rule systematically distorts updated beliefs, with

$$D_{\mu_0}^{\chi,\nu^*}(\nu) = \chi\nu^* + (1-\chi)\nu$$

which is affine in ν.

Depending on the nature of ν^*, this updating rule captures different biases. One may think of ν^* as an 'ideal' belief, but other interpretations can also be interesting. When ν^* is the uniform distribution over Ω, it means Receiver tends to smooth out posterior beliefs. We can also allow ν^* to vary with μ_0. In the presence of two states, for instance, putting full weight under ν^* on the more likely state under μ_0 captures the idea of 'confirmatory bias' (Rabin and Schrag, 1999). When $\nu^* = \mu_0$, it generates posteriors that are closer to the prior than the Bayesian ones and is called 'conservative Bayesianism' (Edwards, 1968), with the distortion function

$$D_{\mu_0}^{CB\chi}(\nu) = \chi\mu_0 + (1-\chi)\nu.$$

Example 2 (Motivated Updating)**.** *Consider a motivated belief updating model where Receiver suffers a psychological loss when her posterior is away from a reference belief ν^* and can adjust her posterior relative to the Bayesian posterior ν with some cost.*[7] *Such*

[7]Beliefs also impact Receiver's utility in Lipnowski and Mathevet (2018), but belief-based utilities do not determine posteriors in their framework. Instead, they study optimal information disclosuse when Sender is a benevolent expert who pursues a *Bayesian* Receiver's best interest.

a trade-off is summarized by a belief-based utility, $\mathcal{U}(\mu, \nu, \nu^*)$, which Receiver wants to maximize, then the associated distortion function is

$$D_{\mu_0}^{MU}(\nu) = \arg\max_{\mu \in \Gamma(\nu)} \mathcal{U}(\mu, \nu, \nu^*)$$

where $\Gamma(\nu) \subset \Delta(\Omega)$ is the set of Receiver's possible posteriors given the Bayesian one. We assume the above constrained maximization has a unique solution at each $\nu \in \Delta(\Omega)$.

For a concrete example, consider a model of motivated conservative Bayesian updating (Hagmann and Loewenstein 2017) where $\nu^* = \mu_0$. Receiver updates her beliefs in a conservative Bayesian fashion defined in Example 1, but chooses χ to maximize $\mathcal{U}(D_{\mu_0}^{CB\chi}(\nu), \nu, \mu_0)$. Therefore

$$D_{\mu_0}^{MCB}(\nu) = \chi^* \mu_0 + (1 - \chi^*)\nu$$

where $\chi^* = \arg\max_{\chi \in [0,1]} \mathcal{U}(D_{\mu_0}^{CB\chi}(\nu), \nu, \mu_0)$.

The updating rules above naturally satisfy the property of systematically distorting updated beliefs since they are defined by their distortion functions. Below are some subtler examples.

Example 3 (Grether's $\alpha - \beta$ Model)**.** *The two-parameter updating rule below is the most common specification of non-Bayesian updating in the literature (Grether 1980, Benjamin, Rabin, and Raymond 2016, Augenblick and Rabin 2018, Benjamin 2019, Benjamin et al. 2019):*

$$\mu_s^R(\omega; \mu_0, \pi) = \frac{\pi(s|\omega)^\beta \mu_0(\omega)^\alpha}{\sum_{\omega' \in \Omega} \pi(s|\omega')^\beta \mu_0(\omega')^\alpha}$$

where $\alpha, \beta > 0$. With different parameter values, it compromises four common biases in belief updating: base-rate neglect for $0 < \alpha < 1$, overweighting prior for $\alpha > 1$, underinference for $0 < \beta < 1$ and overinference for $\beta > 1$. To see it satisfies the condition in Proposition 2, note that for any signal-realization pairs (π, s) and $(\hat{\pi}, \hat{s})$ described in Proposition 2, there exists a constant λ such that $\pi(s|\omega) = \lambda \hat{\pi}(\hat{s}|\omega)$, so

$$\mu_s^R(\omega; \mu_0, \pi) = \frac{(\lambda \hat{\pi}(\hat{s}|\omega))^\beta \mu_0(\omega)^\alpha}{\sum_{\omega' \in \Omega} (\lambda \hat{\pi}(\hat{s}|\omega'))^\beta \mu_0(\omega')^\alpha} = \frac{(\hat{\pi}(\hat{s}|\omega))^\beta \mu_0(\omega)^\alpha}{\sum_{\omega' \in \Omega} (\hat{\pi}(\hat{s}|\omega'))^\beta \mu_0(\omega')^\alpha} = \mu_s^R(\omega; \mu_0, \hat{\pi}).$$

Therefore, it systematically distorts updated beliefs, with:

$$D_{\mu_0}^{\alpha,\beta}(\nu) = \frac{\nu^\beta \mu_0^{\alpha-\beta}}{\sum_{\omega' \in \Omega} \nu(\omega')^\beta \mu_0(\omega')^{\alpha-\beta}}.$$ [8]

[8] For vectors $s, t \in \mathbb{R}^N$, st denotes the component-wise product, i.e., $(st)_i = s_i t_i$. Similarly, $(\frac{s}{t})_i = \frac{s_i}{t_i}$ and $(s^\alpha)_i = s_i^\alpha$.

Example 4 (Divisible Updating)**.** *Cripps (2018) axiomatically characterizes the belief updating processes that are independent of the grouping of multiple signals, i.e., divisible updating rules. Any divisible updating rule is characterized by a homeomorphism* $F : \Delta(\Omega) \to \Delta(\Omega)$ *such that*

$$\mu_s^R(\ .\ ; \mu_0, \pi) = F^{-1}\left(\mu_s^B(\ .\ ; F(\mu_0), \pi)\right).$$

For any signal-realization pairs (π, s) *and* $(\hat{\pi}, \hat{s})$ *described in Proposition 2,* $\mu_s^B(\ .\ ; F(\mu_0), \pi) = \mu_s^B(\ .\ ; F(\mu_0), \hat{\pi})$*, so by Proposition 2, it systematically distorts updated beliefs, with*

$$D_{\mu_0}^{DU}(\nu) = F^{-1}\left(\frac{\nu\frac{F(\mu_0)}{\mu_0}}{\sum_{\omega'\in\Omega}\nu(\omega')\frac{F(\mu_0)(\omega')}{\mu_0(\omega')}}\right).$$

Note that $D_{\mu_0}^{DU}(\mu_0) = \mu_0$*.*

4.2 Broader Examples

Examples below involve a Bayesian Receiver, but depart from KG in other dimensions. These cases can also be addressed using the techniques established in this paper. The first example shows how previous results by Alonso and Câmara (2016) and Galperti (2019) relates to our general approach.

Example 5 (Bayesian Updating with a Different Prior)**.** *A Bayesian persuasion problem where Sender and Receiver have different full-support priors,* μ_0 *and* μ_0^R*, is equivalent to a common-prior non-Bayesian persuasion problem where Receiver's updating rule is*

$$\mu_s^R(\ .\ ; \mu_0, \pi) = \mu_s^B(\ .\ ; \mu_0^R, \pi).$$

For any signal-realization pairs (π, s) *and* $(\hat{\pi}, \hat{s})$ *described in Proposition 2,* $\mu_s^B(\ .\ ; \mu_0^R, \pi) = \mu_{\hat{s}}^B(\ .\ ; \mu_0^R, \hat{\pi})$*, so by Proposition 2, it systematically distorts updated beliefs, with*

$$D_{\mu_0}^{NCP}(\nu) = \frac{\nu\frac{\mu_0^R}{\mu_0}}{\sum_{\omega'\in\Omega}\nu(\omega')\frac{\mu_0^R(\omega')}{\mu_0(\omega')}}.$$

This is exactly Equation (6) in Alonso and Câmara (2016, Proposition 1). Note that $D_{\mu_0}^{NCP}(\mu_0) = \mu_0^R \neq \mu_0$*.*

A distortion function may exist even if Sender and Receiver have non-common priors with different supports. Galperti (2019) assumes that when an unexpected realization happens, Receiver first changes her prior to a full-support one, which is fixed and known to Sender, and then applies Bayesian updating. Since Receiver's prior only depends on whether the evidence is expected (not π *itself), her posterior is pinned down by the Bayesian one. The corresponding distortion function only differs from the above in that Receiver's prior is either the original* μ_0^R *or the full-support one, depending on whether the Bayesian posterior disproves* μ_0^R*, see Galperti (2019, Proposition 1).*

Example 6 (Probability Weighting)**.** *When making choices, people oftentimes attribute excessive weight to events with low probabilities and insufficient weight to events with high probability (a feature accommodated in prospect theory, for instance). While perhaps updating beliefs accurately, Receiver may not use the Bayesian posterior correctly when deciding on which action to take, but instead use* $W(\mu_s^B(\ .\ ;\mu_0,\pi))$ *for some probability weighting function* $W:\Delta(\Omega)\to\Delta(\Omega)$*. Of course, this is an example of rule that systematically distorts updated beliefs, where the distortion function is simply the probability weighting function itself.*

That being said, we don't know much about the interplay between probability weighting and belief updating. In another scenario, Receiver might first subconsciouly distort the prior, using $W(\mu_0)$ *instead of* μ_0*, then apply Bayesian updating, and finally distort the posterior once again, using the posterior* $W(\mu_s^B(\ .\ ;W(\mu_0),\pi))$*. It is straightforward to check, as for Examples 4 and 5, that the condition in Proposition 2 is satisfied.*

4.3 Non-Examples

To better understand the realm of our approach, it is also informative to see examples of rules that do not share this feature of systematically distorting updated beliefs. Many of these examples will be used to illustrate ideas in the rest of the paper. We start with a couple of simple, more technical examples.

Example 7. *[No Learning Without Full Disclosure] Consider a Receiver who does not learn unless every realization of the signal reveals a state with certainty:*

$$\mu_s^R(\omega;\mu_0,\pi)=\begin{cases}1, & \text{if } \pi(s|\omega)>0 \text{ and for all } s',\exists!\omega' \text{ s.t. } \pi(s'|\omega')>0\\ 0, & \text{if } \pi(s|\omega)=0 \text{ and for all } s',\exists!\omega' \text{ s.t. } \pi(s'|\omega')>0\\ \mu_0(\omega), & \text{otherwise.}\end{cases}$$

This updating rule satisfies the martingale property but does not systematically distort updated beliefs since the mapping between the Bayesian posterior and the non-Bayesian one depends on the signal.

Example 8 (Normalized Exponential Transformation)**.** *Given a function* $f:[0,1]\to\mathbb{R}_+$ *such that* $f(x)>0$ *if* $x>0$*, think of the general updating rule*

$$\mu_s^f(\omega;\mu_0,\pi)=\frac{f(\pi(s|\omega)\mu_0(\omega))}{\sum_{\omega'\in\Omega}f(\pi(s|\omega')\mu_0(\omega'))}.$$

Bayes' rule corresponds to the special case where f *is the identity. For many other functions, however,* μ^f *is not homogenous of degree zero in* $\pi(s|\cdot)$ *and thus does not systematically distort updated beliefs, by Proposition 2.*

Signal realizations are oftentimes multi-dimensional. A doctor, for instance, may run multiple tests to guide his patient; a prosector's investigation may follow multiple

lines of inquiry to support his case; a drug company may present multiple evidence to get its new drug approved. In these cases, a signal realization s is more precisely described as a vector $(s_1, \ldots, s_K)$, that is, Sender uses a signal where S has a product structure: $S = S_1 \times \cdots \times S_K$. As for rational updating, that structure is inconsequential for rules studied thus far. Indeed, all that matters is how states correlate with signal realizations s, and the nature of those realizations does not matter. By contrast, the next two examples illustrate mistakes in probabilistic inferences that can arise with multi-signals.

Example 9 (Information Aggregation Mistakes)**.** *Similar to the non-Bayesian social learning literature,*[9] *see DeGroot (1974), Jadbabaie et al. (2012), and Molavi et al. (2018), Receiver treats each aspect of the signal realization in isolation (e.g., Receiver reads sections of a report one at a time, independently drawing inferences from each), and averages the K induced Bayesian posteriors:*

$$\mu_s^{AVG}(\ .\ ; \mu_0, \pi) = \sum_{k=1}^{K} \frac{1}{K} \mu_s^B(\ .\ ; \mu_0, \pi_k),$$

where π_k is the marginal of π on dimension k, for $k = 1, \ldots, K$. We can apply Proposition 2 to see that μ^{AVG} does not systematically distort updated beliefs. Suppose there are two equally-likely states, $\Omega = \{A, B\}$, and a signal π^γ delivering realizations in $\{a, b\} \times \{a', b'\}$ according to the conditional distributions given in Table 1, where γ is a parameter between $1/4$ and $1/2$. For systematic distortion, it must be that the probability of A conditional on receiving the realization (a, a') is independent of γ (e.g., equal to $1/2$ in case of Bayesian updating). By contrast, the updated belief under μ^{AVG} is equal to $\frac{1+4\gamma}{1+8\gamma}$, which does vary with γ.

	a'	b'
a	γ	1/4
b	1/4	1/2-γ

(a) Likelihoods under state A

	a'	b'
a	γ	0
b	0	1-γ

(b) Likelihoods under state B

Table 1. Signal π^γ

Example 10 (Correlation Neglect)**.** *Alternatively, Receiver is said to suffer from 'correlation neglect' (Levy et al. 2018a, 2018b) if she processes all K signals as a whole but applies Bayesian updating to the wrong joint distribution, treating each component of the joint signal as an independent signal:*

[9] In social learning, the multiple sources of information come from different people one is connected to in a network.

$$\mu_s^{CN}(\ .\ ;\mu_0,\pi) = \mu_s^B(\ .\ ;\mu_0,\prod_{k=1}^{K}\pi_k).$$

Again, Proposition 2 shows that this rule does not systematically distort updated beliefs: going back to the example scenario of Table 1, the updated belief for A conditional on (a,a') – $\frac{1+8\gamma+16\gamma^2}{1+8\gamma+32\gamma^2}$ – *also varies with* γ*. Clearly, correlation neglect can also arise under* μ^{AVG}*, but the two updating rules are quite different. For a stark example, when* $(s_1,...,s_K)$ *are fully correlated,* μ^{CN} *is equivalent to an* $\alpha-\beta$ *rule where* $\alpha=1$ *and* $\beta=K$*. By contrast,* μ^{AVG} *agrees with Bayesian updating in this case, recognizing that information can be gleaned only from a single component of the realizations, as others are just copies of it. Also, notice how distributions over posteriors satisfy the martingale property under* μ^{AVG} *but not always under* μ^{CN}*.*

Unlike Examples above, some complicated information processing models cannot be reduced to an updating rule $\mu_s^R(\omega;\mu_0,\pi)$ that, given μ_0 and π, maps each realization s to a posterior belief and are thus beyond our framework. For example, Rabin and Schrag (1999) model 'confirmatory bias' with a binary signal as probabilistically mistaking a disconfirming realization for a confirming one, so μ^R maps a confirming signal to its Bayesian posterior but a disconfirming realization to a distribution over two posteriors.[10] With this model, τ^R induced by a signal has the same support as the Bayesian one but different likelihoods of posteriors, which renders it hard to adapt the concavification method. Another example is Bloedel and Segal (2018), where a rationally inattentive Receiver's attention cost is proportional to the mutual information between Senders signals and her perceptions. Due to the optimal attention strategy, Receiver's updated belief depends on her incentive u and the entire signal structure. Hence even if we extend our definition to allow μ^R to vary with u, it still does not systematically distort updated beliefs for each given u.

5 Revelation Principle

Sender faces a moral hazard problem: he wants Receiver to take some action, but Receiver is free to choose what she desires. KG establishes a version of the revelation principle (Myerson 1991, Section 6) for Bayesian persuasion. Specifically, with a Bayesian Receiver, any value v^* achievable with some signal π can be achieved with a straightforward signal π' that produces a "recommended action" always followed by Receiver, that is, $S'\subset A$ and $\pi'(a|\omega)=\sum_{s\in S^a}\pi(s|\omega)$, where $S^a=\{s|\hat{a}(\mu_s^R)=a\}$ for each $a\in A$. With $\mu^R=\mu^B$, since a was an optimal response to each $s\in S^a$, it must also be an optimal response to the realization a from π', so the distribution of Receiver's actions conditional on the state under π' is the same as under π. However, this may not

[10] In contrast, the specification we give in Example 1 models 'confirmatory bias' with a general signal and is applicable under our non-Bayesian persuasion framework.

be true for a non-Bayesian updating rule μ^R even if it systematically distorts updated beliefs.

Example 11. *Consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$, a uniform prior μ_0, and $A = \{a_1, a_2\}$. Receiver applies the following updating rule:*

$$\mu_s^R(\omega; \mu_0, \pi) = \frac{\pi(s|\omega)^2 \mu_0(\omega)}{\sum_{\omega' \in \Omega} \pi(s|\omega')^2 \mu_0(\omega')}$$

which is a special case of the $\alpha - \beta$ model where $\alpha = 1$ and $\beta = 2$. With signal π and Receiver's utility function u shown in Table 2, simple algebra confirms that $S^{a_1} = \{s_1, s_2\}$ while Receiver strictly prefers a_2 when facing the straightforward signal realization a_1, so the revelation principle fails.

ω \ s	s_1	s_2	s_3
ω_1	1	0	0
ω_2	0	1	0
ω_3	0.49	0.49	0.02

(a) Signal π

ω \ a	a_1	a_2
ω_1	0.3	0
ω_2	0.3	0
ω_3	0	1

(b) Receiver's utility u

Table 2. Example Where the Revelation Principle Fails

The reason why the revelation principle holds with a Bayesian Receiver is that any Bayesian posterior induced by a recommendation a is a convex combination of the Bayesian posteriors induced by the original signal realizations in S^a. Since a is Receiver's optimal choice for each of these original posteriors, it will remain optimal for the new posterior obtained through convex combination. Similarly, if μ^R systematically distorts updated beliefs and the distortion function D_{μ_0} always maps a convex combination of Bayesian posteriors to a convex combination of the distorted posteriors, the revelation principle will remain valid. The formal proof is in the appendix.

Proposition 3. *Given Ω, if the distortion function D_{μ_0} satisfies that for any ν_1 and $\nu_2 \in \Delta(\Omega)$ and any $\lambda \in [0, 1]$, there exists $\gamma \in [0, 1]$, such that*

$$D_{\mu_0}(\lambda \nu_1 + (1 - \lambda)\nu_2) = \gamma D_{\mu_0}(\nu_1) + (1 - \gamma) D_{\mu_0}(\nu_2),$$

then the revelation principle holds.

Any affine distortion function certainly satisfies the condition in Proposition 3 with $\gamma = \lambda$, so the revelation principle holds for the variety of cases discussed in Example 1. More generally, any projective transformation satisfies it, with γ being in general a

function of ν_1, ν_2 and λ. For example, the revelation principle holds for the updating rule discussed in Example 5: $D_{\mu_0}^{NCP}$ is a projective transformation with

$$\gamma = \frac{\lambda < \nu_1, \frac{\mu_0^R(\omega')}{\mu_0(\omega')} >}{\lambda < \nu_1, \frac{\mu_0^R(\omega')}{\mu_0(\omega')} > + (1-\lambda) < \nu_2, \frac{\mu_0^R(\omega')}{\mu_0(\omega')} >}.$$

Similarly, the revelation principle holds for an $\alpha - \beta$ rule where $\beta = 1$ and for a divisible rule where F is a projective transformation (so $D_{\mu_0}^D$ is a composition of two projective transformations). Also note that the condition in Proposition 3 holds often when there are only two states, as it suffices that the distortion function is monotonic.

The following proposition shows that a slightly weaker version of the condition identified in Proposition 3 is necessary: If the images of three collinear beliefs under D_{μ_0} are non-collinear, then the revelation principle fails in some persuasion problem involving D_{μ_0}.

Proposition 4. *Given $|\Omega| \geq 3$, if there exist two beliefs ν_1, $\nu_2 \in \Delta(\Omega)$ and $0 < \lambda < 1$ such that $D_{\mu_0}(\lambda\nu_1 + (1-\lambda)\nu_2)$ is not collinear with $D_{\mu_0}(\nu_1)$ and $D_{\mu_0}(\nu_2)$, then there exists an action space A, a Receiver's utility function u and a signal π such that the revelation principle does not hold.*

With Proposition 4, it is easy to see that the revelation principle does not hold for any $\alpha - \beta$ rule where $\beta \neq 1$ if there are more than two states. Combining the above two propositions and the fundamental theorem of projective geometry leads to the following corollary:

Corollary 1. *Given $|\Omega| \geq 3$, the revelation principle holds with an one-to-one distortion function D_{μ_0} if, and only if, D_{μ_0} is a projective transformation.*

When the revelation principle holds in a persuasion game, any value v^* achievable with some signal can be achieved with a straightforward signal. Thus optimal persuasion (if well-defined) can be achieved in such cases with at most $|A|$ signal realizations. Many more signal realizations may be required for optimal persuasion when the revelation principle fails.[11] For a straightforward example, consider a Receiver who does not learn unless there is full disclosure (Example 7). When $E_{\mu_0} v(\hat{a}(\mu_0), \omega) < E_{\mu_0} v(\hat{a}(\omega), \omega)$ (which is generically true if $u = v$), Sender will choose full disclosure, so optimal persuasion requires $|\Omega|$ signal realizations, which may be much larger than $|A|$.

Another result from KG tells us that optimal Bayesian persuasion can be achieved with a signal that has at most $|\Omega|$ realizations. This follows from an application of Caratheodory's theorem, and remains true independently of the properties of the indirect utility functions. Thus this result extends to any rule that systematically distorts updated beliefs.

[11] There are also situations with an optimal signal involving less than $|A|$ realizations while the revelation principle fails.

Proposition 5. *Whenever an optimal signal exists,*[12] *optimal persuasion can be achieved using a signal with at most* $|\Omega|$ *realizations.*

6 Which Updating Rule is Preferable?

By now, we understand better how each persuasion problem generates a value for Sender. Receivers can then be ranked based on the value they generate given the updating rules they use. In this section, we intend to better understand the role of non-Bayesian updating rules on optimal persuasion by uncovering more robust comparisons, that is, comparisons that hold for a large class of persuasion problems sharing a common information structure.

Adopting Bayes' rule to update beliefs is the rational, correct thing to do. Without thinking much about it, one could conjecture that Sender can always profitably nudge a receiver who suffers from mistakes in statistical inferences. Very quickly, one realizes that this view is naive. For instance, a close-minded person who never updates beliefs is the worst possible Receiver. Of course, other updating rules may be preferable to Bayesian updating for Sender. A totally gullible person, who adopts the belief stated in a signal realization without paying attention to the probability distribution generating it, is best for Sender. How do more realistic updating rules compare, to Bayesian updating and with each others?

Fix the set Ω of states of the world, and the common prior μ_0. Suppose Ann uses the updating rule μ^R, while Beth uses the updating rule $\hat{\mu}^R$. Sender *unambiguously prefers* Ann over Beth (or μ^R over $\hat{\mu}^R$, denoted $\mu^R \succeq \hat{\mu}^R$) if, for all action set A, all utility functions (u, v), and all signal $\hat{\pi}$, there exists a signal π such that his expected utility when using π in the persuasion problem $(\Omega, \mu_0, A, (u, v), \mu^R)$, is larger or equal than his expected utility when using $\hat{\pi}$ in the modified persuasion problem with $\hat{\mu}^R$ replacing μ^R. The comparison is strict ($\mu^R \succ \hat{\mu}^R$) if, in addition, there exists $\varepsilon > 0$, (u, v), A, and a signal π such that his expected utility when using π given μ^R in the persuasion problem $(\Omega, \mu_0, A, (u, v), \mu^R)$, is larger or equal than ε plus his expected utility when using $\hat{\pi}$ in the modified persuasion problem with $\hat{\mu}^R$ replacing μ^R.

What limits Sender in his information design problem is the set of distributions over Sender-Receiver posterior pairs he can generate. Say that Ann is *easier to persuade* than Beth if $T(\mu_0, \hat{\mu}^R) \subseteq T(\mu_0, \mu^R)$. The comparison is strict if the inclusion is strict, that is, strictly more distributions over posterior pairs are achievable when facing Ann.

Section 2 suggests a close link between unambiguous preference comparisons and being easier to persuade. A couple of subtleties arise though. Obviously, Sender unambiguously prefers Ann over Beth if Ann is easier to persuade. However, the converse need not hold and being strictly easier to persuade need not imply a strict preference. This is because *not all distributions over posterior pairs are critical for optimal persuasion.*

[12] As is the case, for instance, if the distortion function is continuous.

Example 12. *Suppose Receiver gets overwhelmed by, and stops paying attention to, signals with too many realizations. Formally, $\mu_s^R(\omega;\mu_0,\pi) = \mu_s^B(\omega;\mu_o,\pi)$, if π has at most K signal realizations, while $\mu_s^R(\omega;\mu_o,\pi) = \mu_0(\omega)$ for other signals π. Suppose for the sake of this example that there happens to be fewer states than K: $|\Omega| \leq K$. Clearly, persuasion is strictly easier with μ^B than μ^R, since distributions over posterior pairs with more than K elements in the support are achievable when Receiver pays attention to realizations of all signals. Remember that, given any signal and Bayesian updating, Sender can achieve the same expected value using a signal with at most $|\Omega|$ signal realizations. Hence, Sender does not strictly prefer μ^B over μ^R.*[13] *It is easy then to also construct a variant $\hat{\mu}^R$ of μ^R that Sender unambiguously strictly prefers over μ^B, while $T(\mu_0,\hat{\mu}^R)$ is not a subset of $T(\mu_0,\mu^B)$.*

Say that the posterior $\nu' \neq \mu_0$ is *feasible for Ann* (or given μ^R) if ν' belongs to the support of some element of $T^R(\mu_0,\mu^R)$, that is, if it arises with strictly positive probability for some signal, when Receiver updates her beliefs according to μ^R. Otherwise, it is said to be *unfeasible* for Ann. As we will argue in the proof of the following proposition, if a posterior is feasible given μ^R, but not given $\hat{\mu}^R$, then there exists a pair of utility functions such that Sender gets a strictly higher value of persuasion with μ^R instead of $\hat{\mu}^R$. We are now ready to provide a fuller understanding of how unambiguous preference comparisons relate to the notion of being easier to persuade.

Proposition 6. *(a) Sender unambiguously prefers Ann over Beth if she is easier to persuade;*

(b) If Sender unambiguously prefers Ann over Beth, and at least one posterior is feasible for Ann but not for Beth, then Sender unambiguously strictly prefers Ann over Beth;

(c) If one posterior is feasible for Ann, but not for Beth, and another is feasible for Beth, but not for Ann, then there is no unambiguous comparison between Ann and Beth.

Example 9-10 (Continued)**.** *Notice that a Receiver suffering from correlation neglect is easier to persuade than its Bayesian counterpart, since any distribution of Bayesian posteriors can be achieved by unidimensional signals. Hence, $\mu^{CN} \succeq \mu^B$ by Proposition 6(a). Similarly, $\mu^{AVG} \succeq \mu^B$. In fact, Levy et al. (2018b, Theorem 1) prove that Sender can approach his first-best payoff under μ^{CN} by using signals with sufficiemtly many components. Hence Sender unambiguously prefers μ^{CN} to essentially all alternative updating rules. However, no such universal dominance holds for μ^{AVG} because the set of distributions over posteriors in this case remains rather limited even if one allows for any number of signal dimensions. Indeed, they must always satisfy the martingale*

[13] Given Proposition 5, this example extends when replacing μ^B by any rule that systematically distorts updated beliefs.

property. [14]

Unambiguous preference comparisons are demanding, as they must hold for *all* persuasion problems sharing the same original information structure (Ω, μ_0). If no such unambiguous comparison holds, a fuller understanding arises by focusing on smaller classes of persuasion problems. We will consider the case where Sender's utility is state independent. The definitions of $\succ$ and $\succeq$ can be adapted at once to reflect this additional restriction on Sender's utility. The resulting relations allow to compare more updating rules, but remain incomplete. Formally, Sender *unambiguously prefers* Ann over Beth *whatever his state-independent utility* (or μ^R over $\hat{\mu}^R$, denoted $\mu^R \succeq^* \hat{\mu}^R$) if, for all (A, u, v) such that v is state independent, and all signal $\hat{\pi}$, there exists a signal π such that his expected utility when using π in the persuasion problem $(\Omega, \mu_0, A, (u, v), \mu^R)$, is larger or equal than his expected utility when using $\hat{\pi}$ in the modified persuasion problem with $\hat{\mu}^R$ replacing μ^R. The definition of $\succ$ is adapted similarly.

Part (a) of Proposition 6 becomes a bit stronger, as Sender unambiguously prefers Ann over Beth, whatever his state-independent utility, as soon as $T^R(\mu_0, \hat{\mu}^R) \subseteq T^R(\mu_0, \mu^R)$ (comparing only feasible posteriors for Ann and Beth, instead of having an inclusion in terms of Sender-Receiver posterior pairs). The astute reader will have noticed that parts (b) and (c) were proved with a state-independent utility function for Sender, and hence remain valid.

Proposition 6^* *If $T^R(\mu_0, \hat{\mu}^R) \subseteq T^R(\mu_0, \mu^R)$, then $\mu^R \succeq^* \hat{\mu}^R$. Furthermore, parts (b) and (c) of Proposition 6 continue to apply with $\succ^*$ instead of $\succ$.*

The next example illustrates Proposition 6^*.

Example 13 (Conservative Bayesian)**.** *It is easy to check that $T^R(\mu_0, \mu^{CB\chi})$ is strictly decreasing in χ, that is, more distributions over Receiver's posteriors are feasible the closer she is from updating rationally. By Proposition 6^*, we conclude that Sender unambiguously (strictly) prefers smaller degrees of conservatism (that is, being closer to rationality), whatever his state-independent utility.*

Does this comparison extend to the whole class of Sender's preferences? It may come as a surprise that the answer is negative: there are intuitive persuasion problems (with state-dependent utility, of course) where Sender prefers conservative Bayesian updating over rationality.

Example 13 (Continued)**.** *A manager must decide whether to assign an employee*

[14] By following the arguments developed in the rest of this section, the reader can easily check that Sender is then indifferent between μ^B and μ^{AVG} when his utility is state-independent, but that he may benefit from the discrepancy between μ^B and μ^{AVG} otherwise (implying that $\mu^{AVG} \succ \mu^B$). We will see, for instance, that Bayesian and conservative Bayesian are not unambiguously comparable (see Example 13 and Proposition 7 below). Yet $\mu^{AVG} \succeq \mu^{CB\frac{1}{2}}$ (simply adding to any signal a second component that is uninformative).

to a new venture, and if so, whether to assign Abe or Bob to it. Requirements in terms of effort and qualification, as well as levels of profit, are uncertain. For the sake of this example, we simply consider two equally-likely states, ω_1 and ω_2. The manager's and Abe's payoffs are provided in Table 3 (as will be clear shortly, Bob's payoffs are irrelevant). Abe gets a zero payoff if he is not picked, and the manager gets a zero payoff if the new venture is not pursued. The manager's payoff from the venture is positive in ω_1 and negative in ω_2, whatever the selected employee, but losses and gains are amplified when picking Bob. When selected, Abe's payoffs are perfectly aligned with those of his manager. Abe is in charge of gathering preliminary information about the state to help his manager decide what to do.

ω \ a	Abe	Bob	No one
ω_1	1	2	0
ω_2	-2	-6	0

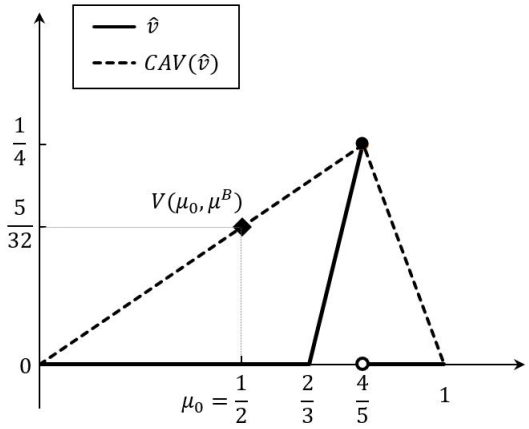
(a) Manager (Receiver)

ω \ a	Abe	Bob	No one
ω_1	1	0	0
ω_2	-2	0	0

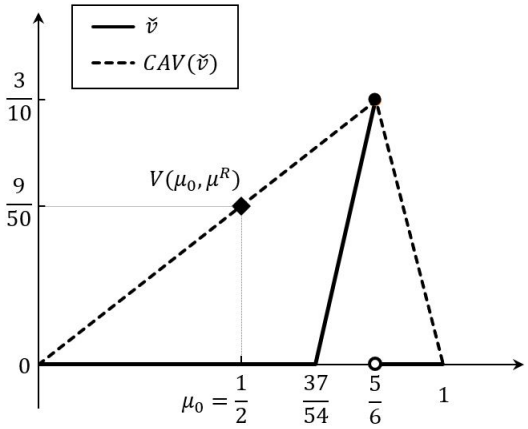
(b) Abe (Sender)

Table 3. Players' Payoffs

Consider first the case of a rational manager. If hiring Bob was not an option, then incentives would be perfectly aligned, and optimal persuasion would be fully informative. However, this strategy is clearly sub-optimal in the presence of Bob, as the manager would then either pick him, or no one. Instead, Abe picks the signal that generates the posterior 0 *with probability* 3/8*, and the posterior* 4/5 *(the threshold above which the manager will pick Bob) with probability* 5/8*. Abe's expected utility is* 5/8 *times* $(4/5) - 2(1/5)$*, or* 1/4*. Figure 1(a) depicts the situation.*



(a) Rational Manager



(b) Conservative Bayesian Manager

Figure 1. Optimal Persuasion

Suppose instead the manager updates her beliefs conservatively, say with $\chi = 1/10$. Now Abe can use a signal that will more accurately reveal to his manager the state ω_1, without jeopardizing his chance of being selected over Bob. Optimal persuasion generates the Bayesian posterior 0 with probability 2/5, and the Bayesian posterior 5/6 (so that the manager's incorrect updated belief is again precisely 4/5) with probability 3/5. Abe's expected utility is 3/5 times $(5/6) - 2(1/6)$, or 3/10. Figure 1(b) depicts the situation. Thus Abe ends up strictly better off with a conservative Bayesian manager ($\chi = 1/10$) than with a rational manager ($\chi = 0$).

The previous example shows how $T^R(\mu_0, \hat{\mu}^R) \subseteq T^R(\mu_0, \mu^R)$ does not imply that Sender unambiguously prefers μ^R over $\hat{\mu}^R$. Indeed what matters for such comparisons are the distributions over Sender-Receiver *posterior pairs* generated by μ^R and $\hat{\mu}^R$. With this in mind, we come to the perhaps surprising result that most rules that systematically distort updated beliefs are incomparable when Sender's preference is unrestricted. Though the conclusion holds even more generally, we focus on the large class of *invertible* distortion functions (which encompasses all the examples in Section 4.1).

Proposition 7. *Let μ^R and $\hat{\mu}^R$ be two distinct rules that systematically distort updated beliefs. If the associated distortion functions are one-to-one, then neither $\mu^R \succeq \hat{\mu}^R$, nor $\hat{\mu}^R \succeq \mu^R$.*

Since Bayes rule also has an invertible distortion function (identity function), Proposition 7 implies that in general whether Sender prefers a non-Bayesian Receiver who systematically distorts updated beliefs over a Bayesian one depends on the class of persuasion problems considered.

7 When Does Sender Benefit from Persuasion?

KG's analysis highlights the following surprising fact. By designing the right experiment, a sender can oftentimes nudge a receiver's decision to her advantage, even though the receiver is rational and aware of the sender's intent to persuade.

A priori, one may have conjectured that profitable persuasion becomes only more prevalent when Receiver does not update beliefs rationally. Yet, we understand by now that mistakes in probabilistic inferences need not make persuasion easier. Thanks to the previous section, we can say this: if $\mu^R \succeq \hat{\mu}^R$ and Sender benefits from persuasion given $\hat{\mu}^R$, then so does he given μ^R. A similar result holds when replacing $\succeq$ by $\succeq^*$, and restricting attention to state-independent utility for Sender. Given that $\succeq$ and $\succeq^*$ are incomplete, more effort is required to understand circumstances where persuasion is profitable.

In fact, we already have a characterization result: persuasion is profitable if and only if $[CAV(\check{v})](\mu_0) > \hat{v}(\mu_0, \mu_0)$ (see Proposition 1). However, while insightful whenever $\check{v}$ can be graphed to construct its concavification, checking this inequality can be

challenging when there are more than three states. A similar issue arises in KG. To address it, they propose a simpler condition that characterizes profitable persuasion for almost all prior μ_0 (using the fact that A is finite). We now show that this result extends to updating rules that systematically distort updated beliefs, provided that the distortion function is *regular*, that is, such that D_{μ_0} is continuous and $D_{\mu_0}(\mu_0) = \mu_0$.

As in KG, we say "Receiver's *preference is discrete* at belief μ if Receiver's expected utility from her preferred action $\hat{a}(\mu)$ is bounded away from her expected utility from any other action, i.e., if there is an $\epsilon > 0$ such that $\forall a \neq \hat{a}(\mu)$, $E_\mu u(\hat{a}(\mu), \omega) > E_\mu u(a, \omega) + \epsilon$." This is copied verbatim from KG, as it corresponds to a joint restriction on Receiver's belief and utility, which has nothing to do with how Receiver updates her beliefs. Clearly, Receiver's preference is discrete at almost all belief μ (since A is finite, a non-discrete preference requires indifference between at least two distinct actions).

We say *there is information Sender would share* (given the prior μ_0 and the distortion function D_{μ_0}) if there is a belief ν such that $\check{v}(\nu) > \hat{v}(\nu, \mu_0)$. In words, if Sender had in his possession private information in the form of a signal realization that led him to believe ν, then he'd prefer sharing that information with Receiver (leading him to believe $D_{\mu_0}(\nu)$) rather than having him act based on the prior. This extends KG's property to reflect the fact that Receiver's posterior is distorted. We can now prove the following.

Proposition 8. *Fix any updating rule that systematically distorts updated beliefs, with a regular distortion function. The two following properties hold.*

(a) If there is no information Sender would share at μ_0, then Sender does not benefit from persuasion.

(b) The converse hold if Receiver's preference is discrete at the prior (which is generically true when A is finite): if there is information Sender would share, then Sender benefits from persuasion.

We see that KG's Proposition 2 is quite robust. While regularity is not needed for part (a), we show in the Online Appendix (by means of counter-examples) that part (b) does not extend to irregular distortion functions (and, a fortiori, more general updating rules).

8 Receiver's Action Varies with Expected State

When the state space is large, the standard concavification method has limited applicability. For this reason, KG (and subsequent papers including Gentzkow and Kamenica (2016) and Dworczak and Martini (2019)) extend their analysis to persuasion problems where Sender's preference is state-independent and Receiver's optimal action varies only with the expected state. In other words, there exists $\tilde{v} : \mathbb{R} \to \mathbb{R}$ such that $v(\hat{a}(\nu')) = \tilde{v}(E_{\nu'}(\omega))$, for all Receiver's belief ν'.

In this section, we establish that such techniques further extend to accommodate updating rules that systematically distort updated beliefs with affine distortion func-

tions.[15] Indeed, we show next that Sender's optimal signals can be found by analyzing a modified problem where Receiver is rational, but her *utility* is distorted.

Proposition 9. *Suppose μ^R systematically distorts updated beliefs, with an affine distortion function D_{μ_0}. Then a signal is optimal in the original persuasion problem $(\Omega, \mu_0, A, (u, v), \mu^R)$ if, and only if, it is optimal in the Bayesian persuasion problem $(\Omega, \mu_0, A, (\tilde{u}, v), \mu^B)$, where $\tilde{u}(a, \omega)$ is the expected utility of action a under the distortion of the Dirac probability measure δ_ω: $\tilde{u}(a, \omega) = E_{D_{\mu_0}(\delta_\omega)} u(a, \cdot)$.*

Since Sender's utility remains unchanged when transforming the original problem $(\Omega, \mu_0, A, (u, v), \mu^R)$ into its fictitious variant $(\Omega, \mu_0, A, (\tilde{u}, v), \mu^B)$, Sender's expected utility from the optimal signals coincides in both problems. The two problems are thus identical for Sender, but for a small difference. The above result captures the best Sender can achieve by using some signal, but of course he also has the option to use no signal at all. This may be different from using an uninformative signal, as $D_{\mu_0}(\mu_0)$ need not equal μ_0. For instance, being irrationally biased towards a cause, Receiver may interpret uninformative evidence about its value as positive news. To account for this possibility, we must keep in mind that Sender benefits from persuasion if, and only if, his expected utility from the optimal signal identified in Proposition 9 is strictly larger than $\hat{v}(\mu_0, \mu_0)$, Sender's expected utility in the absence of persuasion. To summarize, Sender's maximal payoff in $(\Omega, \mu_0, A, (u, v), \mu^R)$ is the maximum of $\hat{v}(\mu_0, \mu_0)$ and his maximal payoff in $(\Omega, \mu_0, A, (\tilde{u}, v), \mu^B)$.

As illustration, recall the distortion function in Example 1, $D_{\mu_0}^{\chi,\nu^*}(\nu) = \chi\nu^* + (1-\chi)\nu$, is affine. Though somewhat limited, this class of updating rules is nonetheless rich enough to gain insight into how mistakes in probabilistic inferences can impact outcomes in applications. Let's revisit for instance KG's Example B in Section V.

Example 14 (Supplying Product Information)**.** *A firm (Sender) faces a single, risk neutral consumer (Receiver) who decides whether to buy one unit of the firm's product. The state $\omega \in [0, 1]$ measures the match quality between the consumer's preference and the product, and represents her consumption utility. Her outside option utility is $\underline{u} \in [0, 1]$, should she decide not to purchase. Therefore, she buys the product if, and only if, $E_{\nu'}[\omega] \geq \underline{u}$, where ν' is her belief regarding ω. The firm and the consumer share a common prior μ_0 about ω which has full support and no atoms on $[0, 1]$. For the problem to be interesting, we assume that $E_{\mu_0}[\omega] < \underline{u}$ (otherwise, the firm sells without persuading). The firm can choose a signal $\pi : [0, 1] \to \Delta(S)$ to reveal some information about the match quality ω (e.g., a trial version of the product or an advertisement with certain details).*

Let's apply Proposition 9 to a consumer with $D_{\mu_0}^{\chi,\nu^*}$. With 0 denoting 'not buying' and 1 denoting 'buying', Receiver's modified utility is given by:

$$\tilde{u}(0, \omega) = \chi E_{\nu^*}[\underline{u}] + (1 - \chi)\underline{u} = \underline{u}$$

$$\tilde{u}(1, \omega) = \chi E_{\nu^*}[\omega'] + (1 - \chi)\omega$$

[15] That is, $D_{\mu_0}(\lambda\nu_1 + (1-\lambda)\nu_2) = \lambda D_{\mu_0}(\nu_1) + (1-\lambda) D_{\mu_0}(\nu_2)$, for all $\lambda \in [0, 1]$ and all beliefs ν_1, ν_2.

for all $\omega \in [0,1]$. *Persuasion in the original problem is equivalent to Bayesian persuasion with Receiver's modified utility* $\tilde{u}$. *A strategically irrelevant re-parametrization of* $\tilde{u}$ *brings us back to a Bayesian version of the original problem where only the outside option's utility has been modified:*

$$\tilde{\tilde{u}}(0,\omega) = \frac{\underline{u} - \chi E_{\nu^*}[\omega']}{1-\chi} \quad \textit{and} \quad \tilde{\tilde{u}}(1,\omega) = \omega,$$

for all $\omega \in [0,1]$.

Using the terminology introduced for Bayesian persuasion in KG's Section IV, it is easy to check that preferences are more aligned when the outside option's value goes down. Thus Sender's optimal profit goes up when $E_{\nu^*}[\omega]$ *increases. Let's now look at the impact of changes in* χ. *Say that the updating bias is unfavorable to the product if the Bayesian posterior is pulled in a direction that makes Receiver less likely to buy, that is, if* $E_{\nu^*}[\omega] < \underline{u}$. *In that case, Sender's optimal profit goes down as* χ *increases, and vice versa when the updating bias is favorable to the product. Two special cases are worth noting. First, the consumer's updating bias can be so unfavorable to the product that, contrary to the rational case, she cannot be persuaded to buy in any way. Indeed, the modified outside option value is larger than* 1 *when* $E_{\nu^*}[\omega] < \underline{u}$ *and* $1 \geq \chi > \frac{1-\underline{u}}{1-E_{\nu^*}[\omega]}$. *Second, the consumer's updating bias can be so favorable to the product that the firm succeeds at selling by using an uninformative signal. This happens when* $E_{\mu_0}[\omega]$ *is larger or equal to the modified outside option value, that is, when* $E_{\nu^*}[\omega] > \underline{u}$ *and* $1 \geq \chi \geq \frac{u - E_{\mu_0}[\omega]}{E_{\nu^*}[\omega] - E_{\mu_0}[\omega]}$.

We now turn our attention to consumer's welfare. Sobel (2013) makes the following observation:

> *"Systematic evidence of behavioral biases will motivate different ways in which opportunistic Senders can relax the Bayesian plausibility restriction and take advantage of biased Receivers. It is not necessary that a cognitive bias will make the Receiver worse off. It might be interesting to investigate circumstances in which behavioral biases are not costly. When biases are not costly, they would presumably survive evolutionary arguments designed to eliminate non-optimizing decision rules."*

We investigate this idea in our example by computing how well the consumer fares on average (in actual terms, not in terms of perceived utility). For this, we must know what the optimal persuasion strategy is, which is doable thanks to Proposition 9 and Corollary 2 of Dworczak and Martini (2019). It is optimal to reveal whether ω *is below or above* u^* *where* $E[\omega|\omega \geq u^*] = \frac{\underline{u} - \chi E_{\nu^*}[\omega]}{1-\chi}$ *(remember that the revelation principle holds, and hence two signal realizations are sufficient). With such a signal, the consumer's actual average payoff is* $E_{\mu_0}[\underline{u} 1_{[\omega < u^*]} + \omega 1_{[\omega \geq u^*]}] = \underline{u} + \int_{u^*}^{1} (\omega - \underline{u}) d\mu_0$, *which is maximized at* $u^* = \underline{u}$ *(coinciding with the first-best), increases below that threshold and decreases above it. If the consumer's updating bias is favorable to the product, then* u^* *decreases with* χ, *and the highest payoff is reached when* $\chi = 0$ *(Bayesian updating). However, when*

the consumer's updating bias is unfavorable to the product, her actual average payoff increases with χ up to the first best when $\chi = \frac{E[\omega|\omega \geq \underline{u}] - \underline{u}}{E[\omega|\omega \geq \underline{u}] - E_{\tilde{\nu}}[\omega]}$. Receiver over consumes under optimal persuasion when she is rational ($\chi = 0$) and an updating bias against the product turns out to be beneficial, as it forces Sender to recommend buying only for states above a larger threshold, which better aligns the consumer's preference.

9 Extensions

9.1 Belief over Updating Rules

Suppose Sender is unsure about Receiver's updating rule, and instead maximizes his expected payoff given a probabilistic belief λ regarding μ^R. Each realization s of a signal π now generates a distribution over posterior pairs: $(\mu_s^B(\ .\ ; \mu_0, \pi), \mu_s^R(\ .\ ; \mu_0, \pi))$ arises with probability $\lambda(\mu^R) \sum_\omega \pi(s|\omega)\mu_0(\omega)$. As π varies, we obtain a set $T(\mu_0, \lambda)$ that generalizes the definition of $T(\mu_0, \mu^R)$. One can then adapt (2) by averaging payoffs over all possible updating rules in the support of λ, which resembles the multiple-receivers, public-information case in KG's Section VI.B.

When focusing on rules that systematically distort updated beliefs, λ can be thought of as a distribution over distortion functions. For instance, Sender may use Grether (1980) to model Receiver's inferences, along with a probabilistic belief regarding the specific values of the parameters α and β. Interestingly, optimal persuasion value can still be found by concavification in such cases, simply by adjusting Sender's indirect utility for Bayesian posteriors. Indeed, Sender's optimization problem becomes:

$$V(\mu_0, \lambda) = \sup_{\rho \text{ Bayes-plausible}} \sum_{\nu \in \, supp(\rho)} \rho(\nu) v_\lambda(\nu)$$

where

$$v_\lambda(\nu) = \sum_{D_{\mu_0} \in \, supp(\lambda)} \lambda(D_{\mu_0}) \check{v}(\nu | D_{\mu_0})$$

is the expectation of the value function defined in (5).[16] Proposition 1 extends, and optimal persuasion requires at most $|\Omega|$ signal realizations (Proposition 5).

9.2 Robust Persuasion

Suppose Sender models Receiver's probabilistic inferences using Bayesian updating as a benchmark, but fears that she may make some limited mistakes. For instance, by not making careful computations, Receiver's perceived posteriors may be in the ballpark of, but not always equal to, the correct posterior. To fix ideas, suppose Sender is confident that Receiver's posterior falls within distance ε of the Bayesian posterior ν, but is

[16] Recall that $\check{v}$ defined in (5) depends on the distortion function D_{μ_0}. Here we make such dependence explicit since we are varying D_{μ_0}.

concerned though that any posterior within the ball $B(\nu, \varepsilon)$ is possible. Sender may then be interested in finding a signal that *guarantees* him a good profit *whatever* the mistakes Receiver may make within those bounds. Hence Sender's indirect utility for a posterior ν is the infimum of his indirect utility for posteriors in its neighborhood, and the ε-robust optimal persuasion value can be found by concavification of this function. Indeed, Sender solves the following optimization problem:

$$V(\mu_0, \varepsilon) = \sup_{\rho \text{ Bayes-plausible}} \sum_{\nu \in \, supp(\rho)} \rho(\nu) v_\varepsilon(\nu)$$

where

$$v_\varepsilon(\nu) = \inf_{\nu' \in B(\nu, \varepsilon)} \hat{v}(\nu, \nu').$$

With v_ε substituting $\check{v}$, our key simplifying results still apply. Effectively, $B(\cdot, \varepsilon)$ should be viewed as a distortion *correspondence*.[17] By selecting the worst posterior for Sender in each set, to reflect the desire for a guaranteed payoff, it reduces to a distortion function as in the rest of the paper.

Robustness has been a recent topic of interest in mechanism design, aiming to find institutions guaranteeing good outcomes for a large class of model misspecifications. The above can be seen as an analogous exercise in simple information design problems. Because the information designer (Sender) faces a single agent (Receiver), discontinuity issues arising in mechanism design (see Oury and Tercieux (2012)) are not a problem here: v_ε converges to v as ε goes to 0 as soon as $\hat{v}$ is continuous.

Robust persuasion has intuitive implications. Consider KG's prosecutor-judge example: the prosecutor wants a guilty verdict, but the judge chooses to convict only when her belief of the defendant being guilty surpasses some threshold τ^* larger than the prior (that is, conviction occurs only with persuasive evidence). Under optimal Bayesian persuasion, signal realizations (evidence) either make the judge certain of innocence, or bring her belief exactly at τ^*. Indeed, it maximizes the conviction rate by pooling the largest possible fraction of innocent defendants with the guilty while keeping the judge willing to convict. This strategy, however, is very risky for the prosecutor, should he fear the judge's probabilistic inferences might not always be perfectly accurate. For ε-robustness, the prosecutor will reduce a bit the conviction rate by conservatively triggering a larger Bayesian posterior of $\tau^* + \varepsilon$. Indeed, v_ε remains a step function as in the Bayesian case, but with a threshold at $\tau^* + \varepsilon$ instead of τ^*.

References

Alonso, R. and O. Câmara (2016). Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory 165*, 672–706.

[17] Clearly, our reasoning extends to cases where the notion of neighborhood is more complex than a ball, possibly varying with ν, and where the reference posterior ν around which the neighborhood is defined, may itself be a distortion of the Bayesian posterior.

Augenblick, N. and M. Rabin (2018). Belief movement, uncertainty reduction, and rational updating. Working Paper.

Aumann, R. J., M. Maschler, and R. E. Stearns (1995). *Repeated games with incomplete information*. MIT press.

Benjamin, D., A. Bodoh-Creed, and M. Rabin (2019). Base-rate neglect: Foundations and implications. working paper.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), *Handbook of Behavioral Economics - Foundations and Applications 2*, Chapter 2, pp. 69–186. North-Holland.

Benjamin, D. J., M. Rabin, and C. Raymond (2016). A model of nonbelief in the law of large numbers. *Journal of the European Economic Association 14*(2), 515–544.

Bloedel, A. W. and I. R. Segal (2018). Persuasion with rational inattention. Available at SSRN 3164033.

Camerer, C. (1998). Bounded rationality in individual decision making. *Experimental economics 1*(2), 163–183.

Caplin, A., M. Dean, and J. Leahy (2019). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. working paper.

Chauvin, K. P. (2019). Euclidean properties of Bayesian updating. Working paper.

Crawford, V. P. (2019). Efficient mechanisms for level-k bilateral trading. Working Paper.

Danziger, S., J. Levav, and L. Avnaim-Pesso (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences 108*(17), 6889–6892.

de Clippel, G. (2014). Behavioral implementation. *American Economic Review 104*(10), 2975–3002.

de Clippel, G., R. Saran, and R. Serrano (2019). Level-k mechanism design. *Review of economic studies 86*(3), 1207–1227.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), 118–121.

Dworczak, P. and G. Martini (2019). The simple economics of optimal persuasion. *Journal of Political Economy 127*(5), 000–000.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment*, pp. 17–52. New York: Wiley.

Eliaz, K., R. Spiegler, and H. C. Thysen (2019). Strategic interpretations. CEPR Discussion Paper No. DP13441.

Epstein, L. G. (2006). An axiomatic model of non-Bayesian updating. *The Review of Economic Studies 73*(2), 413–436.

Epstein, L. G., J. Noor, and A. Sandroni (2008). Non-Bayesian updating: a theoretical framework. *Theoretical Economics 3*(2), 193–229.

Gabaix, X. (2019). Behavioral inattention. In B. D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), *Handbook of Behavioral Economics - Foundations and Applications 2*, Chapter 4, pp. 261–343. North-Holland.

Galperti, S. (2019). Persuasion: The art of changing worldviews. *American Economic Review 109*(3), 996–1031.

Gentzkow, M. and E. Kamenica (2016). A rothschild-stiglitz approach to Bayesian persuasion. *American Economic Review 106*(5), 597–601.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics 95*(3), 537–557.

Guthrie, C., J. J. Rachlinski, and A. J. Wistrich (2001). Inside the judicial mind. *Cornell L. Rev. 86*, 777.

Guthrie, C., J. J. Rachlinski, and A. J. Wistrich (2007). Blinking on the bench: How judges decide cases. *Cornell L. Rev. 93*, 1.

Hagmann, D. and G. Loewenstein (2017). Persuasion with motivated beliefs. mimeo.

Jadbabaie, A., P. Molavi, A. Sandroni, and A. Tahbaz-Salehi (2012). Non-Bayesian social learning. *Games and Economic Behavior 76*(1), 210–225.

Kamenica, E. and M. Gentzkow (2009). Bayesian persuasion. NBER Working Paper No. 15540.

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

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017). Human decisions and machine predictions. *The quarterly journal of economics 133*(1), 237–293.

Koehler, J. J. (2002). When do courts think base rate statistics are relevant? *Jurimetrics*, 373–402.

Leadbetter, R., S. Cambanis, and V. Pipiras (2014). *A Basic Course in Measure and Probability: Theory for Applications*. Cambridge university press.

Lee, Y.-J., W. Lim, and C. Zhao (2019). Cheap talk with non-Bayesian updating. Working paper.

Lehrer, E. and R. Teper (2016). Who is a Bayesian. Working paper.

Levy, G., I. M. de Barreda, and R. Razin (2018a). Persuasion with correlation neglect. Unpublished manuscript, London School Econ.

Levy, G., I. Moreno de Barreda, and R. Razin (2018b). Persuasion with correlation neglect: media power via correlation of news content. CEPR Discussion Paper No. DP12640.

Lindsey, S., R. Hertwig, and G. Gigerenzer (2002). Communicating statistical dna evidence. *Jurimetrics 43*, 147.

Lipnowski, E. and L. Mathevet (2018). Disclosure to a psychological audience. *American Economic Journal: Microeconomics 10*(4), 67–93.

Lipnowski, E., L. Mathevet, and D. Wei (Forthcoming). Attention management. *American Economic Review: Insights*.

Molavi, P., A. Tahbaz-Salehi, and A. Jadbabaie (2018). A theory of non-Bayesian social learning. *Econometrica 86*(2), 445–490.

Myerson, R. B. (1991). *Game theory: Analysis of Conflict*. Harvard University Press.

Oury, M. and O. Tercieux (2012). Continuous implementation. *Econometrica 80*(4), 1605–1637.

Rabin, M. and J. L. Schrag (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics 114*(1), 37–82.

Shmaya, E. and L. Yariv (2009). Foundations for Bayesian updating. Unpublished paper, CalTech.[989].

Sobel, J. (2013). Giving and receiving advice. *Advances in economics and econometrics 1*, 305–341.

Tversky, A. and D. Kahneman (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology 5*(2), 207–232.

Wei, D. (2018). Persuasion under costly learning. Available at SSRN 3188302.

Zhao, C. (2016). Pseudo-Bayesian updating. Working Paper.

Appendix

Proof of Proposition 2. For sufficiency, we start by fixing a full-support prior μ_0 and let D_{μ_0} be the function defined in the statement. Consider now any signal π and any signal realization s. We must prove that $\mu_s^R(\,.\,;\mu_0,\pi) = D_{\mu_0}(\mu_s^B(\,.\,;\mu_0,\pi))$. To do this, we apply the assumption for sufficiency to show that $\mu_s^R(\,.\,;\mu_0,\pi) = \mu_s^R(\,.\,;\mu_0,\hat{\pi}_\nu)$ where $\nu = \mu_s^B(\,.\,;\mu_0,\pi)$. Notice that, by Bayes rule, $\pi(s|\omega) = 0$ if and only if $\nu(\omega) = 0$, so by definition of $\hat{\pi}_\nu$, $\hat{\pi}_\nu(\hat{s}|\omega) = 0$ if and only if $\pi(s|\omega) = 0$. Hence it remains to check that $\frac{\hat{\pi}(\hat{s}|\omega)}{\pi(s|\omega)}$ is constant over the set of ω's such that both $\pi(s|\omega) > 0$ and $\hat{\pi}(\hat{s}|\omega) > 0$. Notice that

$$\frac{\hat{\pi}_\nu(\hat{s}|\omega)}{\pi(s|\omega)} = \frac{\nu(\omega)}{\mu_0(\omega)\pi(s|\omega)} \min_{\omega'} \frac{\mu_0(\omega')}{\nu(\omega')} = \frac{1}{\sum_{\omega''\in\Omega} \pi(s|\omega'')\mu_0(\omega'')} \min_{\omega'} \frac{\mu_0(\omega')}{\nu(\omega')},$$

which is indeed independent of ω.

For necessity, suppose that μ^R systematically distorts updated beliefs with distortion functions $(\hat{D}_{\mu_0})_{\mu_0\in\Delta(\Omega)}$. Consider now some full-support prior μ_0 and two signal-realization pairs (π, s) and $(\hat{\pi}, \hat{s})$ such that the likelihood ratio $\frac{\hat{\pi}(\hat{s}|\omega)}{\pi(s|\omega)}$ is constant over the set of ω's for which $\pi(s|\omega) > 0$ and $\hat{\pi}(\hat{s}|\omega) = 0$ whenever $\pi(s|\omega) = 0$. We have to prove that $\mu_s^R(\,.\,;\mu_0,\pi) = \mu_{\hat{s}}^R(\,.\,;\mu_0,\hat{\pi})$, or equivalently that $\hat{D}_{\mu_0}(\mu_s^B(\,.\,;\mu_0,\pi)) = \hat{D}_{\mu_0}(\mu_{\hat{s}}^B(\,.\,;\mu_0,\hat{\pi}))$. To establish this last equality, we simply check that $\mu_s^B(\,.\,;\mu_0,\pi) = \mu_{\hat{s}}^B(\,.\,;\mu_0,\hat{\pi})$. For the constant ratio condition to hold, it must be that, for each ω, $\pi(s|\omega) > 0$ if and only if $\hat{\pi}(\hat{s}|\omega) > 0$. If $\pi(s|\omega) = \hat{\pi}(\hat{s}|\omega) = 0$, then both $\mu_s^B(\omega;\mu_0,\pi)$ and $\mu_{\hat{s}}^B(\omega;\mu_0,\hat{\pi})$ equal 0. If both $\pi(s|\omega)$ and $\hat{\pi}(\hat{s}|\omega)$ are strictly positive, then

$$\mu_s^B(\omega;\mu_0,\pi) = \frac{\pi(s|\omega)\mu_0(\omega)}{\sum_{\omega'\in\Omega}\pi(s|\omega')\mu_0(\omega')} = \frac{\hat{\pi}(\hat{s}|\omega)\mu_0(\omega)}{\sum_{\omega'\in\Omega}\hat{\pi}(\hat{s}|\omega')\mu_0(\omega')} = \mu_{\hat{s}}^B(\omega;\mu_0,\hat{\pi}),$$

as desired. Since the necessary condition has been established, we know now from the first part of the proof that the distortion functions D_{μ_0} defined in the statement can be used instead, and hence $\hat{D}_{\mu_0} = D_{\mu_0}$. □

Proof of Proposition 3. Given an arbitrary compact action set A, an arbitrary utility function $u(a,\omega)$ for Receiver, and an arbitrary signal on Ω with a realization set S. let $S^a = \{s|\hat{a}(\nu_s') = a\}$ for each $a \in A$, where ν_s' is Receiver's posterior after observing realization s. Define a straightforward signal with $S' = A$ and $\pi'(a|\omega) = \sum_{s\in S^a}\pi(s|\omega)$. We want to show that a is also an optimal response to the realization a from π'. For any $a \in S^a$

$$\begin{aligned}\mu_a^B(\omega;\mu_0,\pi') &= \frac{\sum_{s\in S^a}\mu_0(\omega)\pi(s|\omega)}{\sum_{\omega'}\sum_{s'\in S^a}\mu_0(\omega')\pi(s'|\omega')} \\ &= \sum_{s\in S^a}\frac{\mu_0(\omega)\pi(s|\omega)}{\sum_{\omega''}\mu_0(\omega'')\pi(s|\omega'')}\frac{\sum_{\omega''}\mu_0(\omega'')\pi(s|\omega'')}{\sum_{\omega'}\sum_{s'\in S^a}\mu_0(\omega')\pi(s'|\omega')} \\ &= \sum_{s\in S^a}\lambda_s\mu_s^B(\omega;\mu_0,\pi)\end{aligned}$$

where $\lambda_s = \frac{\sum_{\omega''}\mu_0(\omega'')\pi(s|\omega'')}{\sum_{\omega'}\sum_{s'\in S^a}\mu_0(\omega')\pi(s'|\omega')} = \frac{\tau_s}{\sum_{s'\in S^a}\tau_{s'}}$.

By assumption, there exists $\{\gamma_s\}_{s\in S^a}$ such that

$$\begin{aligned}\mu_a^R(\omega;\mu_0,\pi') &= D_{\mu_0}(\mu_a^B(\omega;\mu_0,\pi')) \\ &= D_{\mu_0}(\sum_{s\in S^a}\lambda_s\mu_s^B(\omega;\mu_0,\pi)) \\ &= \sum_{s\in S^a}\gamma_s D_{\mu_0}(\mu_s^B(\omega;\mu_0,\pi)) \\ &= \sum_{s\in S^a}\gamma_s\mu_s^R(\omega;\mu_0,\pi)\end{aligned}$$

so we have

$$\begin{aligned}\hat{a}(\mu_a^R(\cdot;\mu_0,\pi')) &= \arg\max_{a'\in A}\sum_{\omega}u(a',\omega)\mu_a^R(\omega;\mu_0,\pi') \\ &= \arg\max_{a'\in A}\sum_{s\in S^a}\sum_{\omega}u(a',\omega)\gamma_s\mu_s^R(\omega;\mu_0,\pi)\end{aligned}$$

Since for all $s\in S^a$, a maximizes $\sum_\omega u(a',\omega)\mu_s^R(\omega;\mu_0,\pi)$, a should also maximize the convex combination of those terms, so $\hat{a}(\mu_a^R(\cdot;\mu_0,\pi')) = a$. This proves the sufficiency. □

Proof of Proposition 4. First, consider the case where μ_0 is a convex combination of ν_1 and ν_2. If $D_{\mu_0}(\mu_0)$ is not collinear with ν_1' and ν_2' (Case 1.1), we can find an action space A, a Receiver's utility function $u(a,\omega)$ and a signal π that induces ν_1' and ν_2' with $S = \{s_1,s_2\}$, such that $S^a = \{s_1,s_2\}$ for π, yet $a \neq \arg\max_{a'\in A}\sum_\omega u(a',\omega)\mu_a^R(\cdot;\mu_0,\pi') = \arg\max_{a'\in A}\sum_\omega u(a',\omega)D_{\mu_0}(\mu_0)(\omega)$ for the non-informative straightforward signal π', which is a contradiction to the revelation principle. If otherwise $D_{\mu_0}(\mu_0)$ is collinear with ν_1' and ν_2' (Case 1.2), then $\mu_0 \neq \lambda\nu_1 + (1-\lambda)\nu_2$ and $D_{\mu_0}(\mu_0)$ is not collinear with ν_1' (ν_2') and $\nu^* = D_{\mu_0}(\lambda\nu_1 + (1-\lambda)\nu_2)$. WLOG, assume μ_0 is a convex combination of ν_1 and $\lambda\nu_1 + (1-\lambda)\nu_2$, then we can choose a signal π with $S = \{s_1,s_2\}$ where s_1 induces Receiver's posterior ν_1' and s_2 induces ν^* to get the same contradiction as in Case 1.1.

When μ_0 is collinear with ν_1 and ν_2 but not a convex combination of ν_1 and ν_2, we can pick a ν_3 collinear with ν_1 and ν_2 such that μ_0 is a convex combination ν_1 (ν_2) and ν_3. If $D_{\mu_0}(\mu_0)$ is not collinear with ν_1' and ν_3' (Case 2.1), then this is essentially the same as Case 1.1. If $D_{\mu_0}(\mu_0)$ is collinear with ν_1' and ν_3' but ν^* is not collinear with ν_1' and ν_3' (Case 2.2.1), then this is essentially the same as Case 1.2. If both $D_{\mu_0}(\mu_0)$ and ν^* are collinear with ν_1' and ν_3' (Case 2.2.2), then $D_{\mu_0}(\mu_0)$ cannot be collinear with ν_2' and ν_3', so we are back at Case 1.1 with ν_1 substituted by ν_3.

Now suppose μ_0 is not collinear with ν_1 and ν_2. Pick any point ν_3 on the ray that goes from $\lambda\nu_1 + (1-\lambda)\nu_2$ through μ_0 such that there exists a Bayes plausible distribution of posteriors τ with $supp(\tau) = \{\nu_1, \nu_2, \nu_3\}$ and $\frac{\tau(\nu_1)}{\tau(\nu_2)} = \frac{\lambda}{1-\lambda}$. If $\nu_3' = D_{\mu_0}(\nu_3)$ is not collinear with ν_1' and ν_2' (Case 3.1), then there exists a convex region in $\Delta(\Omega)$ which contains ν_1' and ν_2' but not ν_3' nor ν^*. Therefore, we can find an action space A, a Receiver's utility function $u(a,\omega)$ and a signal π with $S = \{s_1, s_2, s_3\}$ where s_i induces ν_i and ν_i' for Sender and Receiver, such that $S^a = \{s_1, s_2\}$ for π yet $a \neq \arg\max_{a' \in A} \sum_\omega u(a',\omega)\mu_a^R(\cdot;\mu_0,\pi') = \arg\max_{a' \in A} \sum_\omega u(a',\omega)\nu^*(\omega)$, so the revelation principle fails. If ν_3' is collinear with ν_1' and ν_2', then we can find an action space A, a Receiver's utility function $u(a,\omega)$ and a signal π with $S = \{s_1, s_2, s_3\}$, such that $S^a = \{s_1, s_2, s_3\}$ for π, so π' should reveal no information. If $D_{\mu_0}(\mu_0)$ is not collinear with ν_1' and ν_2' (Case 3.2.1), we can choose u such that $a \neq \arg\max_{a' \in A} \sum_\omega u(a',\omega)\mu_a^R(\cdot;\mu_0,\pi') = \arg\max_{a' \in A} \sum_\omega u(a',\omega)D_{\mu_0}(\mu_0)(\omega)$, which contradicts the revelation principle. If otherwise $D_{\mu_0}(\mu_0)$ is collinear with ν_1' and ν_2' (Case 3.2.2), then $D_{\mu_0}(\mu_0) \neq \nu_3'$ cannot be collinear with ν_3' and ν^* while μ_0 is a convex combination of $\lambda\nu_1 + (1-\lambda)\nu_2$ and ν_3, so we can choose a signal π with $S = \{s_1, s_2\}$ where s_1 induces Receiver's posteriors ν^* and ν_3' and we are back at Case 1.2. This finishes the proof. □

Proof of Proposition 5. If D_{μ_0} is continuous, then $\check{v}$ is upper semicontinuous and bounded, so the proof of Proposition 7 in KG (2009) follows. Since $\check{v}$ is bounded for all D_{μ_0}, the proof of Proposition 9 in KG (2009) follows. □

Proof of Proposition 6. As pointed out earlier, (a) is obvious. As for (b) and (c), we simply prove that, should posterior ν' be feasible for Ann but not for Beth, there exists A and (u,v) such that Sender's persuasion value is strictly larger when facing Ann than Beth. A similar argument applies when Ann's and Beth's roles are reversed. Suppose first that $\nu'(\omega) < 1$, for all ω. Consider then the action set $A = \{a_\omega \mid \omega \in \Omega\} \cup \{a^*\}$. Sender has a state-independent utility function and cares only to have a^*: $v(a_\omega, \omega') = 0$ and $v(a^*,\omega) = 1$ for all ω, ω'. Receiver's utility is defined as follows: $u(a^*,\omega) = 0$, $u(a_\omega,\omega) = 1$ and $u(a_\omega,\omega') = -\nu'(\omega)/(\sum_{\omega'' \neq \omega} \nu'(\omega''))$ for all ω and all $\omega' \neq \omega$. Notice that, for all ω, Receiver's expected utility of a_ω is zero should her belief be ν'. For any other belief ν'', there exists a state ω such that $\nu''(\omega) > \nu'(\omega)$, which implies that her

expected utility of a_ω is strictly positive should her belief be ν'':

$$\sum_{\omega'} \nu''(\omega')u(a_\omega, \omega') = \nu''(\omega) - (\sum_{\omega' \neq \omega} \nu''(\omega')) \frac{\nu'(\omega)}{(\sum_{\omega'' \neq \omega} \nu'(\omega''))} > 0.$$

Thus a^* is optimal for Receiver if and only if her belief is ν'. This implies that Sender can achieve a strictly positive value of persuasion with Ann, but not with Beth. A similar, simpler argument applies if $\nu'(\omega) = 1$ for some ω: keeping a^* and its associated payoffs, simply define one additional action a that gives Sender a zero payoff, and Receiver a payoff of 1 except for state ω, in which case her payoff is zero. □

Proof of Proposition 7. Let D_{μ_0} (resp., $\hat{D}_{\mu_0}$) be the distortion function associated to μ^R (resp., $\hat{\mu}^R$). Since these two updating rules are distinct, there exists a probability distribution ν such that $\nu' = D_{\mu_0}(\nu) \neq \hat{D}_{\mu_0}(\nu)$. We now construct a persuasion problem where Sender gets a strictly higher persuasion payoff when facing μ^R rather than $\hat{\mu}^R$. A similar construction provides an example where the comparison is reversed. If $\nu' \notin T^R(\mu_0, \hat{\mu}^R)$, then the proof of Proposition 6 provides such a persuasion problem. Suppose $\nu' \in T^R(\mu_0, \hat{\mu}^R)$ and let $\hat{\nu} = (\hat{D}_{\mu_0})^{-1}(\nu')$. Consider the same action set and utility functions (A, u, v) as in the proof of Proposition 6, except for the following: $v(a^*, \omega^*) = 1$ and $v(a^*, \omega) = -x$ for all $\omega \neq \omega^*$, where ω^* such that $\nu(\omega^*) > \hat{\nu}(\omega^*)$, and x is any number strictly in between $\frac{\nu(\omega^*)}{1-\nu(\omega^*)}$ and $\frac{\hat{\nu}(\omega^*)}{1-\hat{\nu}(\omega^*)}$. As in Proposition 6, Sender's payoff is zero whenever Receiver's posterior is different from ν'. Given μ^R, the rational belief associated to that Receiver's posterior is ν, in which case Sender gets a strictly positive expected payoff. Things are different, however, when Receiver updates beliefs according to $\hat{\mu}^R$: the rational belief associated to it is now $\hat{\nu}$, in which case Sender gets a strictly negative expected payoff. In that case, Sender's optimal persuasion payoff is zero, which is strictly inferior than what he gets when Receiver updates according to μ^R. □

Proof of Proposition 8. For the first part, suppose there is no information Sender would share at μ_0, then for any ν, $\check{v}(\nu) \leq \hat{v}(\nu, \mu_0) = E_\nu v(\hat{a}(\mu_0), \omega)$. Given a signal π that induces some τ, its value is

$$\begin{aligned}
\sum_{s \in S} \tau_s \check{v}(\mu_s^B) &\leq \sum_{s \in S} \tau_s \sum_{\omega \in \Omega} v(\hat{a}(\mu_0), \omega) \mu_s^B(\omega) \\
&= \sum_{\omega \in \Omega} v(\hat{a}(\mu_0), \omega) \sum_{s \in S} \tau_s \mu_s^B(\omega) \\
&= \sum_{\omega \in \Omega} v(\hat{a}(\mu_0), \omega) \mu_0(\omega) \\
&= \hat{v}(\mu_0, \mu_0).
\end{aligned}$$

Thus Sender does not benefit from persuasion.

For the second part, since there is information Sender would share at μ_0, $\check{v}(\nu_h) > \hat{v}(\nu_h, \mu_0)$. As in KG, since Receiver's preference is discrete at μ_0, there exists $\delta > 0$ such that all μ in a δ-ball around μ_0 (denoted as B_δ), $\hat{a}(\mu) = \hat{a}(\mu_0)$. $D_{\mu_0}(\mu_0) = \mu_0$ and its continuity at μ_0 imply that there exists $\phi > 0$, such that all μ in a ϕ-ball around μ_0 (denoted as B_ϕ), $D_{\mu_0}(\mu) \subset B_\delta$. Given μ_0 is in the interior of $\Delta(\Omega)$, there exists a belief ν_l on the ray from ν_h through μ_0 such that $\nu_l \in B_\phi$. Let $\mu_0 = \gamma\nu_l + (1-\gamma)\nu_h$ for some $0 < \gamma < 1$, then there exists some signal π that induces the distribution of joint posteriors τ with $\tau(\nu_l, D_{\mu_0}(\nu_l)) = \gamma$ and $\tau(\nu_h, D_{\mu_0}(\nu_h)) = 1-\gamma$. Therefore

$$\begin{aligned}
E_\tau[\hat{v}(\nu, \nu')] &= \gamma\check{v}(\nu_l) + (1-\gamma)\check{v}(\nu_h) \\
&> \gamma\hat{v}(\nu_l, \mu_0) + (1-\gamma)\hat{v}(\nu_h, \mu_0) \\
&= \sum_{\omega\in\Omega} v(\hat{a}(\mu_0), \omega)[\gamma\nu_l(\omega) + (1-\gamma)\nu_h(\omega)] \\
&= \sum_{\omega\in\Omega} v(\hat{a}(\mu_0), \omega)\mu_0(\omega) \\
&= \hat{v}(\mu_0, \mu_0).
\end{aligned}$$

This finishes the proof. □

Proof of Proposition 9. Given any $\nu \in \Omega(\Delta)$, $\nu(A) = \int_{\omega\in\Omega} \delta_\omega(A)d\nu(\omega)$ for all (Lebesgue) measurable set $A \in \Omega$, where δ_ω is the Dirac measure, so we can rewrite ν as an integral

$$\nu = \int_{\omega\in\Omega} \delta_\omega d\nu(\omega).$$

If the distortion function D_{μ_0} is affine, i.e., $D_{\mu_0}(\lambda\nu_1 + (1-\lambda)\nu_2) = \lambda D_{\mu_0}(\nu_1) + (1-\lambda)D_{\mu_0}(\nu_2)$ for all $\lambda \in [0,1]$ and $\nu_1 \neq \nu_2 \in \Delta(\Omega)$, then

$$D_{\mu_0}(\nu) = D_{\mu_0}\left(\int_{\omega\in\Omega} \delta_\omega d\nu(\omega)\right) = \int_{\omega\in\Omega} D_{\mu_0}(\delta_\omega)d\nu(\omega).$$

Define $u'(a,\omega) = E_{D_{\mu_0}(\delta_\omega)}u(a,\omega')$, then by Leadbetter, Cambanis, and Pipiras (2014) Lemma 7.2.2,

$$\begin{aligned}
E_{D_{\mu_0}(\nu)}u(a,\omega') &= \int_{\omega'\in\Omega} u(a,\omega')dD_{\mu_0}(\nu)(\omega') \\
&= \int_{\omega\in\Omega}\left(\int_{\omega'\in\Omega} u(a,\omega')dD_{\mu_0}(\delta_\omega)\right)d\nu(\omega) \\
&= \int_{\omega\in\Omega} E_{D_{\mu_0}(\delta_\omega)}u(a,\omega')d\nu(\omega) \\
&= E_\nu u'(a,\omega).
\end{aligned}$$

Therefore, $\check{a}(\nu) \equiv \arg\max_{a\in A} E_\nu u'(a,\omega) = \hat{a}(D_{\mu_0}(\nu))$. Sender's modified payoff function $\check{v}(\nu) = E_\nu v(\check{a}(\nu))$ is indeed his reduced form payoff function in the Bayesian persuasion problem $(\Omega, \mu_0, A, (u', v), \mu^B)$. □