Non-Bayesian Persuasion

Geoffroy de Clippel

Brown University

Xu Zhang

Hong Kong University of Science and Technology (Guangzhou) and Hong Kong University of Science and Technology

Following Kamenica and Gentzkow, this 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 into the revelation principle, the ranking of updating rules, when persuasion is beneficial to the sender, and when it is detrimental to the receiver. Our key result also extends to shed light on the question of robust persuasion.

I. Introduction

The past decade has seen much progress on the topic of information design. In a seminal paper, Kamenica and Gentzkow (2011) study how a rational agent (receiver, she) can be persuaded to take a desired action

We thank Pedro Dal Bó, Piotr Dworczak, Kfir Eliaz, Ignacio Esponda, Erik Eyster, Jack Fanning, Matthew Gentzkow, Wooyoung Lim, Elliot Lipnowski, Teddy Mekonnen, Xiaosheng Mu, Bobby Pakzad-Hurson, Alex Poterack, Louis Putterman, Kareen Rozen, Roberto Serrano, Joel Sobel, Alireza Tahbaz-Salehi, Neil Thakral, the editor in charge, and five referees for helpful comments. This paper was edited by Emir Kamenica.

Electronically published August 4, 2022

Journal of Political Economy, volume 130, number 10, October 2022.

https://doi.org/10.1086/720464

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 this 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's 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 that 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.

Suppose that a rational doctor has a patient's best interest at heart but knows that his patient is reluctant to change her belief upon hearing bad news. Which tests should he run to optimally 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 toward 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 Kamenica and Gentzkow's (2011) 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.

The 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 the receiver's belief. Under Bayesian updating, 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 the receiver is not Bayesian. Furthermore, the sender and receiver will now typically have different posteriors. This raises new, interesting questions when the sender's utility is state dependent, as his preferred

[1] The 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), Koehler (2002), Lindsey, Hertwig, and Gigerenzer (2002), Danziger, Levav, and Avnaim-Pesso (2011), and Kleinberg et al. (2017).

action also varies with information revealed by the experiment. Thus, we must characterize not only distributions of the receiver's posteriors induced by experiments but also distributions over sender-receiver posterior pairs (see sec. II).

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 III. Let μ_0 be the common prior (our approach also applies to noncommon priors; see below). 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. Kamenica and Gentzkow's (2011) concavification argument for optimal persuasion extends to such rules with the sender's indirect utility for Bayesian posteriors modified by factoring in the distortion function. This observation echoes the key insight of Alonso and Câmara (2016), who show that the concavification approach can be extended to problems of Bayesian persuasion with noncommon priors, because a simple transformation links the receiver's and the sender's posteriors.

Though straightforward, once the idea of a distortion function has been recognized, this methodology 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, in section IV we start with a characterization of these rules. We then provide a variety of examples encompassed by our notion, including conservative Bayesianism (Edwards 1968), α-β model (Grether 1980), divisible updating rules (Cripps 2018), and motivated updating models (e.g., optimal belief formation; Brunnermeier and Parker 2005) where posterior beliefs are formed endogenously. 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 the sender (a problem first studied in Alonso and Câmara [2016], as mentioned above). To complete the picture of the power and limits of our approach, we end section IV with a few nonexamples, that is, updating rules that do not systematically distort updated beliefs.[3]

[3] We should point out, however, that many conceptual contributions of the paper—the questions we raise and the definitions we introduce to tackle them—are valid beyond the class of rules that systematically distort updated beliefs.

The sender faces the problem of providing the right incentives for the receiver to take the desired action. The revelation principle (see, e.g., sec. VI in Myerson 1991) applies when the receiver is Bayesian: each signal admits an outcome-equivalent incentive compatible signal where realizations are suggested actions (Kamenica and Gentzkow 2011; Bergemann and Morris 2016a, 2016b, 2019). Section V starts with an observation that the revelation principle often fails when the receiver's updating rule does not systematically distort updated beliefs.[4] 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, for rules that systematically distort updated beliefs.

Kamenica and Gentzkow (2011) highlight that effective persuasion is possible in many problems despite the fact that the receiver is rational and knows the sender's intent to persuade her to take an action to his advantage. We note that departures from Bayes's rule need not equate to easier manipulation. For an extreme example, the worst for the sender is a stubborn receiver who never updates her belief. On the other extreme, the best for the 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 the sender fare as a function of the updating rule? Are some updating rules preferable to others? We tackle these questions in sections VI and VII.

Of course, by offering a method to compute the sender's optimal payoff, our extension of Kamenica and Gentzkow's (2011) concavification argument already provides a tool for addressing such questions in any given persuasion problem. However, when ranking updating rules on the basis of the sender's payoff, we can better understand the way specific

[4] Failures of the revelation principle in information design problems have been observed in recent studies. Lipnowski and Mathevet (2018) and Anunrojwong, Iyer, and Lingenbrink (2020) show that the revelation principle fails in Bayesian persuasion settings where the receiver's preference is nonlinear in her belief because of psychological or non-expected utility preferences, while Perez-Richet and Skreta (2021) introduce falsification to information design and find that lying costs also cause the revelation principle to fail. To our knowledge, we are the first to study the failure of the revelation principle when the receiver maximizes expected utility but is non-Bayesian.

belief updating rules systematically impact optimal persuasion by uncovering more robust comparisons—those that hold for a large class of persuasion problems sharing a common information structure. We develop these ideas in section VI and illustrate the concepts we define by means of examples. We provide a necessary and sufficient condition for unambiguous preference comparisons. Then we prove that, perhaps surprisingly, no two distinct rules that systematically distort updated beliefs (and whose distortion functions are one-to-one) 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 specific subclasses of problems. In particular, we study persuasion problems where the sender and receiver have common or purely opposed interests, where the sender's utility is state independent, and where the sender's goal is to get the receiver to switch her action.

Effective persuasion is possible if and only if there is a signal that gives the sender a strictly larger payoff than with the receiver's default action (optimal for the prior). Thanks to the overarching methodology outlined above and developed in section III, 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 Kamenica and Gentzkow 2011) to provide simpler, necessary, and sufficient conditions for when effective persuasion is possible. Section VII speaks to the robustness of Kamenica and Gentzkow's (2011) proposition 2 in that regard, as the result extends verbatim to most rules that systematically distort updated beliefs. Of course, their property that there is information that the sender would share must now take into account that the receiver's posterior after processing that information is a distortion of the Bayesian posterior. We apply this result to illustrate how mistakes in probabilistic inference can have striking implications on persuasion: when the receiver is subject to an overinference bias, persuasion becomes profitable even in the most challenging situations where the sender and receiver have purely opposed interests.

Giving the sender an opportunity to persuade can never make a Bayesian receiver worse off, but real life suggests that persuasion can be harmful to the receiver (e.g., consumers being too easily persuaded by advertising). In section VIII, we study the possibility of detrimental persuasion to a non-Bayesian receiver (in terms of her accurate ex ante welfare). Obviously, persuasion benefitting the sender is harmful to the receiver in the case of purely opposed interests, and optimal persuasion cannot harm in the case of purely common interests. When considering mixtures of conflicting and overlapping interests, whether optimal persuasion is harmful typically varies from problem to problem. We show that for numerous updating rules, persuasion is detrimental to the receiver in a simple problem where the sender wants her to take a certain action other than the status quo. It is also possible to find other updating rules retaining the property that optimal persuasion is never detrimental, which is true in the case of conservative Bayesianism. In fact, a receiver can sometimes be better off being conservative Bayesian than accurately updating beliefs.

Given the difficulty of concavifying general functions in the presence of multiple states, effort has been devoted to better understand the simpler case where the receiver's optimal action is measurable with respect to the state's expected value (see, e.g., Kamenica and Gentzkow 2011; Gentzkow and Kamenica 2016; Dworczak and Martini 2019). We show in section IX that these results extend for updating rules associated with affine distortion functions. Indeed, the original persuasion problem can then be proved to be equivalent to a Bayesian persuasion problem where the 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. Interestingly, the customer may actually fare better on average when she suffers from an unfavorable updating bias toward the product. We also revisit Kamenica and Gentzkow's (2011) lobbyist example and show that, contrary to the Bayesian benchmark, there are always circumstances under which optimal persuasion is partially revealing when the receiver's updating rule satisfies two rather mild properties.

Section X concludes by highlighting how our methodology for dealing with non-Bayesian updating extends in different directions. First, optimal persuasion value remains computable by concavification in problems where the 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, that the sender is concerned that the 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 the receiver's preferred actions for all beliefs in that neighborhood. Once again, the sender's optimal persuasion value remains computable by concavification and represents a guaranteed level of profit despite the risk of the receiver's mistakes. Alternatively, by allowing the sender to pick the receiver's posterior in the distortion correspondence, our approach extends to the case where the sender can nudge naive receivers who are affected by labels and the presentation of signal realizations.

Related literature.—Alonso and Câmara (2016) investigate optimal persuasion for a Bayesian receiver who has a different (full-support) prior than the sender. Galperti (2019) models changing worldviews, which also entail noncommon priors but with different supports. The receiver always adopts the same full-support prior (which is known to the sender and independent of the signal) after unexpected evidence and then updates on the basis of the new prior following Bayes's rule. Receivers in those papers could alternatively be interpreted as non-Bayesian with respect to the 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 the sender can achieve close to his first best in that case. We show that such updating rules do not systematically distort updated beliefs.

Some information processing models cannot be reduced to a belief updating rule that depends on only the prior and the signal. Bloedel and Segal (2018), Wei (2018), and Lipnowski, Mathevet, and Wei (2020) study optimal persuasion when the receiver is rationally inattentive—that is, the receiver is Bayesian but rationally decides how much information to acquire. Such a receiver's updated belief depends on her incentive and the entire signal structure. Eliaz, Spiegler, and Thysen (2021a, 2021b) analyze how the sender can manipulate a naive Bayesian receiver's misspecified beliefs by selective redaction of the message's content or by providing partial data for interpretation to enhance his persuasiveness. Neither the rational inattention models nor the Eliaz et al. (2021a, 2021b) models can be seen as systematically distorting updated beliefs even if we fix the receiver's incentive or the redaction (or interpretation) strategy, as the updated belief following one signal realization may depend on other realizations that could have occurred.

From a methodological perspective, our work further develops the belief-based approach to information design initiated by Kamenica and Gentzkow (2011) to accommodate a non-Bayesian receiver whose updating rule systematically distorts updated beliefs. In contrast, Bergemann and Morris (2016a, 2016b, 2019) and Taneva (2019) use the revelation principle to formulate a Myersonian approach to information design, based on the notion of Bayes correlated equilibrium. Mathevet, Perego, and Taneva (2020) highlight benefits of the belief-based approach for information provision in games. Our observation that the revelation principle often fails with a non-Bayesian receiver highlights the advantage of the belief-based approach in non-Bayesian persuasion problems.

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 to reveal a verifiable signal to an audience that suffers from base rate neglect. In contrast, we study a wide range of updating rules with a communication protocol where the 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. de Clippel, Saran, and Serrano (2019, 2021), Kneeland (2020), and Crawford (2021) study mechanism design with agents who need not have rational expectations. Our 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), Cripps (2018), Chauvin (2019), Augenblick and Rabin (2021), Dominiak, Kovach, and Tserenjigmid (2021), Kovach (2021), and Zhao (2022). Although we do not propose any axiom, the property of systematically distorting updated beliefs or the reduction property inherent to the revelation principle can help classify different updating rules.

II. General Framework

A state ω is drawn at random according to a full support distribution μ_0 on a finite set Ω. The 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 the receiver's chosen action from a compact set A. Neither player knows the state, but the 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.[5] Upon observing the realization s, the sender correctly updates his belief by applying Bayes's 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 Kamenica and Gentzkow (2011), the receiver may make mistakes in probabilistic inferences. Her posterior is denoted $\mu_s^R(\omega; \mu_0, \pi)$.

Given the prior μ_0 and the 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 $\Sigma_{s\in S(\nu,\nu')}\Sigma_\omega \pi(s|\omega)\mu_0(\omega)$, where $S(\nu, \nu')$ is the set of signal realizations s such that $\nu = \mu_s^B(\cdot; \mu_0, \pi)$

[5] We assume that for all s, there exists ω such that $\pi(s|\omega) > 0$. Otherwise, s is simply dropped.

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

Given belief ν', the receiver picks an optimal action

$$\hat{a}(\nu') \in \operatorname*{argmax}_{a \in A} E_{\nu'} u(a, \omega),$$

and we assume that $\hat{a}(\nu')$ maximizes the sender's expected utility whenever the receiver is indifferent between some actions at ν'.[6]

To figure out the optimal signal and its value, the 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 the sender's utility should the posteriors be ν for himself and ν' for the receiver.

Remark 1. The problem further simplifies should the 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 the 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

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

[6] To fix ideas, we follow a logic akin to partial implementation in mechanism design: the sender, the information designer, can recommend which optimal action the receiver picks in response to signal realizations inducing knife-edge beliefs with multiple optimal responses. Equation (2) defines the sender's persuasion payoff under this most optimistic scenario. On the opposite, a full-implementation approach would focus on the sender's expected payoff when selecting the sender's worst action within the set of the receiver's optimal responses. An intermediate approach would view the sender as a player himself. Focusing on subgame-perfect equilibrium outcomes typically restricts the receiver's optimal response because beliefs at which the receiver has multiple optimal responses can be approximated by belief sequences where the receiver's optimal response is unique. Except for propositions 6*b* and 7, our results apply to all these scenarios, as they hold for any fixed reaction function. Footnote 18 explains how these two propositions extend to other tie-breaking rules under additional assumptions on the receiver's updating rule. For a general approach to information design with various solution concepts and outcome selection rules, see Mathevet, Perego, and Taneva (2020).

III. 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 Kamenica and Gentzkow (2011) is that, should the receiver be rational, a distribution τ^R of posteriors can arise if and only if it is *Bayes plausible* (see also Shmaya and Yariv 2009), 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 does not need to worry about the multitude of possible signals but only about the much simpler space of Bayes-plausible distributions of posteriors.

Suppose now that 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 of 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 Kamenica and Gentzkow (2011) to solve for optimal persuasion.

Say that μ^R *systematically distorts updated beliefs* if, for all full-support priors μ_0, there exists a *distortion function* $D_{\mu_0} : \Delta(\Omega) \rightarrow \Delta(\Omega)$ such that for all signals π and all signal realizations s, $\mu_s^R(\cdot; \mu_0, \pi) = D_{\mu_0}(\mu_s^B(\cdot; \mu_0, \pi))$. 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 function is 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 the receiver's posteriors that can arise under μ^R is characterized by the following *distorted Bayes plausibility* condition:

$$\sum_{\nu' \in supp(\tau^R)} D_{\mu 0}^{-1}(\nu') \tau^R(\nu') = \mu_0.$$

Indeed, the receiver has a posterior ν' with probability $\tau^R(\nu')$ if and only if the rational posterior is $D_{\mu_0}^{-1}(\nu')$. 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, the sender can generate a distribution τ over pairs of posteriors (ν, ν') (i.e., $\tau \in T(\mu_0, \mu^R)$) 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[7]

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

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

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

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 $[\text{CAV}(\check{v})](\mu_0)$. The sender benefits from persuasion if and only if $[\text{CAV}(\check{v})](\mu_0) > \hat{v}(\mu_0, \mu_0)$.

Unlike in Bayesian persuasion, to see whether the sender benefits from persuasion, we need to compare $[\text{CAV}(\check{v})](\mu_0)$ with $\hat{v}(\mu_0, \mu_0) = \Sigma_\omega \mu_0(\omega) v(\hat{a}(\mu_0), \omega)$ rather than $\check{v}(\mu_0) = \Sigma_\omega \mu_0(\omega) v(\hat{a}(D_{\mu_0}(\mu_0)), \omega)$. The former is the sender's default payoff with no persuasion, while the latter represents his payoff from sending an uninformative signal. The two coincide if the receiver is Bayesian but may differ if the receiver's belief can be modified by an uninformative signal ($D_{\mu_0}(\mu_0) \neq \mu_0$).[8]

IV. Characterization and Examples

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

[7] 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.

[8] In sec. IV.A, we give examples of both non-Bayesian updating rules with $D_{\mu_0}(\mu_0) = \mu_0$ and those with $D_{\mu_0}(\mu_0) \neq \mu_0$. When $D_{\mu_0}(\mu_0) \neq \mu_0$, which is the case if the receiver applies an affine distortion function with an ideal different from the prior (example 1) or if the receiver suffers base rate neglect (example 3), the sender can choose to do nothing and have the receiver hold belief μ_0 or to send an uninformative signal and induce $D_{\mu_0}(\mu_0) \neq \mu_0$.

Proposition 2. The updating rule μ^R systematically distorts updated beliefs if and only if, given any full-support prior μ_0, $\mu_s^R(\cdot;\mu_0,\pi) = \mu_{\hat{s}}^R(\cdot;\mu_0,\hat{\pi})$ for all signal realization pairs (π, s) and $(\hat{\pi}, \hat{s})$ such that the likelihood ratio $\hat{\pi}(\hat{s}|\omega)/\pi(s|\omega)$ is constant as a function of ω.[9]

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

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

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

Let us 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 (*neutrality*) and the set of other realizations that could have occurred (*independence of irrelevant signal realizations*). 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 us restrict attention, for simplicity, to signals π such that $\pi(s|\omega) > 0$ for all s and all ω. It follows from (1) that

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

for all s, ω, ω'. Hence, given μ_0, Bayes's rule defines a bijection (independent of π and s) between posterior $\mu_s^B(\cdot;\mu_0,\pi)$ and the set of all likelihood ratios $\{(\pi(s|\omega')/\pi(s|\omega))|\omega,\omega' \in \Omega\}$. 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 the receiver's and the sender's posteriors. In that case, if $\pi(s|\omega')/\pi(s|\omega) = \hat{\pi}(\hat{s}|\omega')/\hat{\pi}(\hat{s}|\omega)$ for all ω, ω', then the 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.

A. *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.

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

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

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

where $0 \le \chi \le 1$ 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 toward 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 the 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 the 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.[10] Such a trade-off is summarized by a belief-based utility, $\mathcal{U}(\hat{\nu}, \nu, \nu^*)$, which the receiver wants to maximize; then, the associated distortion function is

$$D_{\mu_0}^{MU}(\nu) = \underset{\hat{\nu} \in \Gamma(\nu)}{\operatorname{argmax}} \mathcal{U}(\hat{\nu}, \nu, \nu^*),$$

where $\Gamma(\nu) \subset \Delta(\Omega)$ is the set of the receiver's possible posteriors given the Bayesian one. We assume that 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 $\nu^* = \mu_0$.

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

The 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 $\chi^* = \text{argmax}_{\chi \in [0,1]} \mathcal{U}(D_{\mu_0}^{CB\chi}(\nu), \nu, \mu_0)$.

For a second example, consider a model of optimal belief formation in the spirit of Brunnermeier and Parker (2005), where the receiver chooses her belief when making her action choice following a realization. The optimal posterior belief $\hat{\nu}$ is determined by balancing the utility from wishful thinking encoded in the belief itself (an expected utility using $\hat{\nu}$ of the optimal action associated to $\hat{\nu}$) and the actual loss from taking a possibly suboptimal action (an expected utility using the accurate Bayesian posterior ν of the optimal action associated with $\hat{\nu}$). Solving this optimization problem delivers a distortion function that depends on the receiver's incentive u in the persuasion problem:

$$D_{\mu_0}^{BP}(\nu) = \underset{\hat{\nu} \in \Delta(\Omega)}{\text{argmax}}\ E_{\hat{\nu}} u(\hat{a}(\hat{\nu}), \omega) + E_{\nu} u(\hat{a}(\hat{\nu}), \omega).$$

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 2021; 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 whether 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

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

Therefore, it systematically distorts updated beliefs, with[11]

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

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

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

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

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

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

B. *Broader Examples*

Examples below involve a Bayesian receiver but depart from Kamenica and Gentzkow (2011) 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) relate to our general approach.

Example 5 (Bayesian updating with a different prior). A Bayesian persuasion problem where the 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 the receiver's updating rule is

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

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

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

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

This is exactly equation (6) in Alonso and Câmara (2016, proposition 1). Note that $D^{NCP}_{\mu_0}(\mu_0) = \mu^R_0 \neq \mu_0$.

A distortion function may exist even if the sender and receiver have noncommon priors with different supports. Galperti (2019) assumes that when an unexpected realization happens, the receiver first changes her prior to a full-support one, which is fixed and known to the sender, and then applies Bayesian updating. Since the receiver's prior depends on only whether the evidence is expected (not π itself), her posterior is pinned down by the Bayesian one. The corresponding distortion function differs from the above only in that the receiver's prior is either the original μ^R_0 or the full-support one, depending on whether the Bayesian posterior disproves μ^R_0 (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 (e.g., a feature accommodated in prospect theory). While perhaps updating beliefs accurately, the receiver may not use the Bayesian posterior correctly when deciding on which action to take but instead use $W(\mu^B_s(\cdot; \mu_0, \pi))$ 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 do not know much about the interplay between probability weighting and belief updating. In another scenario, the receiver might first subconsciously 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 $W(\mu^B_s(\cdot; W(\mu_0), \pi))$. It is straightforward to check, as for examples 4 and 5, that the condition in proposition 2 is satisfied.

C. *Nonexamples*

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 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, in which case her posterior is δ_ω, the Dirac probability measure defined for the revealed state:

$$\mu^R_s(\omega; \mu_0, \pi) = \begin{cases} 1 & \text{if } \pi(s|\omega) > 0 \text{ and for all } s', \exists\ !\ \omega' \text{ such that } \pi(s'|\omega') > 0, \\ 0 & \text{if } \pi(s|\omega) = 0 \text{ and for all } s', \exists\ !\ \omega' \text{ such that } \pi(s'|\omega') > 0, \\ \mu_0(\omega) & \text{otherwise.} \end{cases}$$

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 transformation). Given a function $f:[0,1]\times[0,1]\rightarrow\mathbb{R}_+$ such that $f(x, y) > 0$ if $x, y > 0$, 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's rule corresponds to the special case where $f(x, y) = xy$. For many other functions, however, μ^f is not homogeneous of degree zero in $\pi(s|\cdot)$ and thus does not systematically distort updated beliefs, by proposition 2.

Signal realizations are oftentimes multidimensional. A doctor, for instance, may run multiple tests to guide his patient; a prosecutor'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 $(s_1, \dots, s_K)$; that is, the 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 multisignals.

Example 9 (Information aggregation mistakes). Similar to the non-Bayesian social learning literature (see DeGroot 1974; Jadbabaie et al. 2012; Molavi, Tahbaz-Salehi, and Jadbabaie 2018),[12] the receiver treats each aspect of the signal realization in isolation (e.g., the receiver reads sections of a report one at a time, independently drawing inferences from each) and averages the K induced Bayesian posteriors:

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

where π_k is the marginal of π on dimension k for $k = 1, \dots, K$. We can apply proposition 2 to see that μ^{AVG} does not systematically distort updated beliefs. Suppose that 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 $(1 + 4\gamma)/(1 + 8\gamma)$, which does vary with γ.

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

TABLE 1
SIGNAL π^{γ}

	a'	b'
	A. Likelihoods under State A	
a	γ	$1/4$
b	$1/4$	$1/2 - \gamma$
	B. Likelihoods under State B	
a	γ	0
b	0	$1 - \gamma$

EXAMPLE 10 (Correlation neglect). Alternatively, the 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:

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

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')—$(1 + 8\gamma + 16\gamma^2)/(1 + 8\gamma + 32\gamma^2)$—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 $(s_1, \dots, s_K)$ 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 from only 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.[13] With this model, τ^R induced by a signal has the same support as the Bayesian one but different likelihoods of posteriors, which makes it impossible to view it as a rule that systematically distorts updated beliefs. Another example is rational inattention (Bloedel and Segal 2018;

[13] 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.

Wei 2018; Lipnowski et al. 2020). Because of the optimal attention strategy, the receiver's updated belief depends on her incentive u and the entire signal structure. Hence, even if we allow μ^R to vary with u, it still violates independence of irrelevant signal realizations and thus does not systematically distort updated beliefs for each given u.

V. Revelation Principle

The sender wants the receiver to take some action, but the receiver is free to choose what she desires. Kamenica and Gentzkow (2011) establish a version of the revelation principle (Myerson 1991, sec. VI) for Bayesian persuasion: for any signal, the corresponding action recommendation is incentive compatible and outcome equivalent. 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 the receiver, that is, $S' \subset A$ and $\pi'(a|\omega) = \Sigma_{s\in S^a}\pi(s|\omega)$, where $S^a = \{s|\hat{a}(\mu_s^R(\cdot;\mu_0,\pi)) = a\}$ 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 the receiver's actions conditional on the state under π' is the same as under π. However, a non-Bayesian receiver may not always follow such action recommendations. We say that the revelation principle fails for a receiver (an updating rule) if there exists a persuasion problem and a signal such that the corresponding action recommendation is not incentive compatible.[14]

We start by observing that, broadly speaking, the revelation principle fails for rules that do not systematically distort updated beliefs. Though true more generally, we substantiate this statement below under some simplifying assumptions.

Proposition 3. Let μ^R be an updating rule such that (a) $\mu_s^R(\cdot;\mu_0,\pi)$ is a continuous function of the vector $[\pi(s|\omega)]_{\omega\in\Omega} \in \mathbb{R}_+^{|\Omega|} \setminus\{0\}$ and (b) the receiver is certain that state ω occurs after a realization s of some experiment if and only if s occurs with strictly positive probability only in state ω. The revelation principle fails if μ^R does not systematically distort updated beliefs.

Aside from continuity, assumption a means that updated beliefs associated with a signal realization s depends on only the state-dependent probability of s in the different states under the experiment, not on the description of s (*neutrality*) or the probability of other signal realizations (*independence of irrelevant signal realizations*).

[14] A similar definition appears in recent papers discussing failures of the revelation principle in persuasion problems (see, e.g., Lipnowski and Mathevet 2018; Anunrojwong et al. 2020).

The revelation principle may fail even for updating rules that systematically distort updated beliefs, as the following example illustrates.

EXAMPLE 11. Consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$, a uniform prior μ_0, and $A = \{a_1, a_2\}$. The 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 $\beta = 2$ (overinference).

With signal π and the receiver's utility function u shown in table 2, simple algebra confirms that for all $0 < \varphi < 0.5$ and all ρ such that $\varphi^2 < \rho < 2\varphi^2$, the receiver strictly prefers a_1 upon realizations s_1 and s_2 and strictly prefers a_2 upon s_3, that is, $S^{a_1} = \{s_1, s_2\}$, yet she would strictly prefer a_2 were s_1 and s_2 replaced by recommendation a_1. Therefore, the action recommendation is not incentive compatible, and the revelation principle fails.

A failure of the revelation principle does not immediately imply that all straightforward signals fail to be incentive compatible or optimal. We continue the example to illustrate three scenarios with different φ and ρ: (1) the optimal signal does require three realizations; (2) the optimal signal involves only two realizations but cannot simply recommend an action that the receiver will follow; and (3) the revelation principle fails, but the optimal signal gives an incentive-compatible action recommendation.

To illustrate the first two scenarios, suppose that the sender strictly prefers a_1 over a_2 in all states: $v(a_1, \omega) = 1$ and $v(a_2, \omega) = 0$ for all ω. Figure 1*A* showcases scenario 1 by depicting the sender's distorted indirect utility function $\check{v}$ when $\rho = 0.1$. We see from the concavification argument that an optimal signal in this persuasion problem must induce three Bayesian posteriors (as π with $\varphi = \sqrt{0.1}$ does).

Figure 1*B* depicts an example of scenario 2 with $\rho = 0.45$. Although the revelation principle fails at π with $\varphi = 0.5$, such a two-realization signal (as s_3 is dropped when $\varphi = 0.5$) inducing ν_1 and ν_2 is optimal. We see that optimality requires the receiver to take a_1 whatever signal realization, which cannot be done with an incentive-compatible straightforward signal.

To illustrate scenario 3, we take $\rho = 0.1$ as in figure 1*A* but add a third action a_3 in table 2: $u(a_3, \omega_1) = u(a_3, \omega_2) = -1$, and $u(a_3, \omega_3) = 2$.

TABLE 2
EXAMPLE WHERE REVELATION PRINCIPLE FAILS

	SIGNAL			RECEIVER'S UTILITY u	
	s_1	s_2	s_3	a_1	a_2
ω_1	1	0	0	ρ	0
ω_2	0	1	0	ρ	0
ω_3	φ	φ	$1 - 2\varphi$	0	1

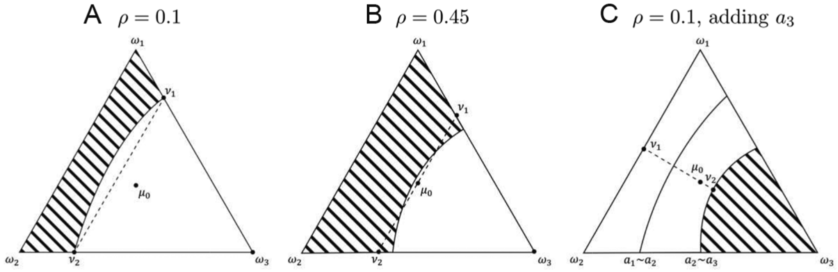

FIG. 1.—Illustration of sender's distorted indirect utility function $\hat{\nu}$. $\hat{\nu} = 1$ in the diagonally striped area (boundaries included); $\hat{\nu} = 0$ elsewhere.

Suppose now that the sender strictly prefers a_3 over a_1 and a_2 in all states: $v(a_3, \omega) = 1$ and $v(a_1, \omega) = v(a_2, \omega) = 0$ for all ω. As shown in figure 1*C*, the concavification argument implies that the sender's optimal expected payoff can be achieved by an incentive-compatible signal that either recommends a_1 (inducing $\nu_1 = (0.5, 0.5, 0)$, where the receiver strictly prefers a_1) or a_3 (inducing $\nu_2 = (1/(2 + \sqrt{2}), 1/(2 + \sqrt{2}), \sqrt{2}/(2 + \sqrt{2}))$, where the receiver is indifferent between a_2 and a_3), even though the revelation principle is violated at any π with $\sqrt{0.05} < \varphi < \sqrt{0.1}$.

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 the 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 appendix sec. A3.

PROPOSITION 4. 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 for all persuasion problems with prior μ_0.

Any affine distortion function certainly satisfies the condition in proposition 4 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

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

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 $D_{\mu_0}^{DU}$ is a composition of two projective transformations). Also note that the condition in proposition 4 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 4 is necessary: if the images of three collinear beliefs under D_{μ_0} are noncollinear, then the revelation principle fails in some persuasion problem involving D_{μ_0}. This result echoes the observation by Lipnowski and Mathevet (2018) and Anunrojwong et al. (2020) that the revelation principle fails in Bayesian persuasion settings where the receiver's preference is nonlinear in her Bayesian belief, which causes the set of Bayesian beliefs such that a given action best responses to the receiver's associated beliefs to be nonconvex.

PROPOSITION 5. Given $|\Omega| \geq 3$, if there exist two beliefs $\nu_1, \nu_2 \in \Delta(\Omega)$ 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 the revelation principle fails.

With proposition 5, it is easy to see that the revelation principle fails 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 a one-to-one distortion function D_{μ_0} for all persuasion problems with prior 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. For a straightforward example, consider a receiver who does not learn unless there is full disclosure (example 7). When $E_{\mu_0} v(\hat{a}(\mu_0), \omega) < E_{\mu_0} v(\hat{a}(\delta_\omega), \omega)$ (which is generically true if $u = v$), the sender will choose full disclosure, so optimal persuasion requires $|\Omega|$ signal realizations, which may be much larger than $|A|$. (See also example 11, scenario 1.)

Another result from Kamenica and Gentzkow (2011) 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.

VI. Which Updating Rule Is Preferable?

By now, we understand better how each persuasion problem generates a value for the sender. Receivers can then be ranked on the basis of 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's rule to update beliefs is the rational, correct thing to do. Yet it would be naive to conjecture that the sender can always profitably nudge a receiver who suffers from mistakes in probabilistic inferences. 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 the 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 the sender. How do more realistic updating rules compare with Bayesian updating and with each other?

Fix the set Ω of states of the world and the common prior μ_0.[15] Suppose Ann uses the updating rule μ^R, while Beth uses the updating rule $\hat{\mu}^R$. The sender *unambiguously prefers* Ann over Beth (or μ^R over $\hat{\mu}^R$, denoted $\mu^R \succcurlyeq \hat{\mu}^R$) if, for all action sets A, all utility functions (u, v), and all signals $\hat{\pi}$, 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 than or equal to his expected utility when using $\hat{\pi}$ in the modified persuasion problem with $\hat{\mu}^R$ replacing μ^R.[16] The comparison is strict $(\mu^R \succ \hat{\mu}^R)$ if the sender unambiguously prefers Ann over Beth but does not unambiguously prefer Beth over Ann.

What limits the sender in his information design problem is the set of distributions over sender-receiver posterior pairs that he can generate (see sec. II). We build on this insight to provide a necessary and sufficient

[15] We fix the initial information environment for expositional convenience when discussing the sender's preference over the receiver's updating rules, since distortion functions and sets of sender-receiver posterior pairs often vary with the prior. As illustrated in example 13, such a preference can reverse if the prior changes. A more stringent definition of unambiguous comparisons arises if one requires the preference to also hold for all priors. Note that all arguments below establishing an unambiguous comparison between two rules are valid independently of the prior, except for proposition 6*a*, which can be amended as follows: a prior-free unambiguous comparison is guaranteed under the stronger requirement that Ann is easier to persuade than Beth at all priors.

[16] Recall that (2) is defined with the supremum, which may not be achievable when the receiver is non-Bayesian. Therefore, we compare two updating rules in terms of the value that can actually be achieved by some signal in a persuasion problem. The definition simplifies when the supremum is reached, as it is the case, e.g., with continuous distortion functions. In that case, the sender unambiguously prefers Ann over Beth if his optimal persuasion payoff with Ann is larger than or equal to that with Beth.

condition for unambiguous preference comparisons. For this, 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. Also, say that a posterior $\nu' \neq \mu_0$ is *feasible for Ann* (or given μ^R) if it belongs to the support of some element of $T^R(\mu_0, \mu^R)$. Otherwise, it is said to be *unfeasible* for Ann.

PROPOSITION 6. (*a*) The sender unambiguously prefers Ann over Beth if Ann is easier to persuade than Beth; (*b*) if the sender unambiguously prefers Ann over Beth, then one cannot find a posterior that is feasible for Beth but not for Ann.

Part *a* follows at once from (2). Part *b* is proved by constructing, for any posterior that might be feasible for Beth but not for Ann, a persuasion problem where the sender gets a strictly larger payoff with Beth (see app. sec. A5).

REMARK 3. The strict counterpart of part *a* does not hold: being strictly easier to persuade need not imply a strict preference, as the next example shows. This is because not all distributions over posterior pairs are critical for optimal persuasion. But, combining parts *a* and *b*, we get the following: if Ann is easier to persuade and at least one posterior is feasible for Ann but not for Beth, then the sender unambiguously strictly prefers Ann over Beth.

EXAMPLE 12. Suppose that the 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 (an example of a rule that does not systematically distort updated beliefs). Suppose for the sake of this example that there happen 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 the receiver pays attention to realizations of all signals. Remember that, given any signal and Bayesian updating, the sender can achieve the same expected value using a signal with at most $|\Omega|$ signal realizations. Hence, the sender does not strictly prefer μ^B over μ^R.

Let us apply proposition 6. 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, $\mu^{CN} \succsim \mu^B$ by proposition 6*a*. Similarly, $\mu^{AVG} \succsim \mu^B$. These comparisons are strict thanks to remark 3. Indeed, it is easy to construct a posterior pair that is feasible for μ^{CN} (or μ^{AVG}) but not for μ^B.[17]

[17] In fact, Levy et al. (2018a, theorem 1) prove that the sender can approach his first-best payoff under μ^{CN} by using signals with sufficiently many components. Hence, the 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 posterior pairs remains rather limited in this case. Indeed, they satisfy the martingale property for the receiver.

When restricting to the unambiguous comparison among updating rules that systematically distort updated beliefs, we find the following negative result:[18]

Proposition 7. Let μ^R and $\hat{\mu}^R$ be two distinct rules that systematically distort updated beliefs. If the associated distortion functions D_{μ_0} and $\hat{D}_{\mu_0}$ are one-to-one, then neither $\mu^R \succcurlyeq \hat{\mu}^R$ nor $\hat{\mu}^R \succcurlyeq \mu^R$.

For instance, Bayesian updating is incomparable to conservative Bayesianism, which may seem counterintuitive, given that the former generates strictly more distributions of posteriors for the receiver than the latter. But, as illustrated in the next example, what matters for unambiguous preference comparisons are distributions over posterior pairs.

Example 13. A manager must decide whether to assign an employee 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.

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 suboptimal in the presence of Bob, as the manager would then pick either 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 2*A* depicts the situation.

Suppose instead that 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

[18] As pointed out in n. 6, results in propositions 6*b* and 7 rely on our partial implementation logic that favors the sender when the receiver has multiple optimal actions. This can be seen in these results' proofs, where an action a^* is optimal for the receiver at a single belief. It is not difficult to adjust the proof of proposition 6*b* to show the following whatever the receiver's optimal reaction function: suppose that Ann's updating rule is derived from a continuous distortion function. If the sender unambiguously prefers Ann over Beth, then one cannot find a posterior ν' that is feasible for Beth and a open ball of beliefs around ν' that are unfeasible for Ann. As for proposition 7, it holds whatever the receiver's optimal reaction function if we had the restriction that the distortion functions are continuous.

TABLE 3
Players' Payoffs

	Abe	Bob	No One
	A. Manager (Receiver)		
ω_1	1	2	0
ω_2	−2	−6	0
	B. Abe (Sender)		
ω_1	1	0	0
ω_2	−2	0	0

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 2*B* depicts the situation. Thus, Abe ends up strictly better off with a conservative Bayesian manager ($\chi = 1/10$) than with a rational manager ($\chi = 0$).

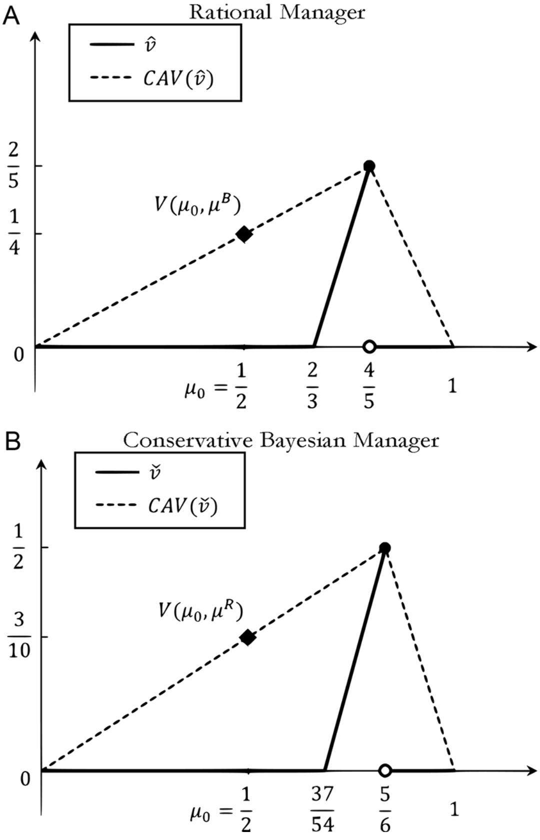

Fig. 2.—Optimal persuasion.

Note that Abe's preference over the manager's updating rules may reverse if the prior changes. For $\mu_0 = 9/10$, Abe's expected utility with a rational manager is $1/5$, and his expected utility with a conservative Bayesian manager ($\chi = 1/10$) is $33/190$, so Abe now strictly prefers the manager to be Bayesian.

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.

A. Common Interests

The sender and receiver have *ex post common interests* if they share in each state a common ranking of actions, that is, $u(a, \omega) \geq u(a', \omega)$ if and only if $v(a, \omega) \geq v(a', \omega)$, for all actions a, a' and all states ω. They have *purely common interests* if they share a common utility function, that is, $u = v$. Notice how the former definition encompasses more problems, as it leaves open the possibility of disagreement over the ranking of actions under risk. For instance, both a patient and her doctor may agree that a surgery is needed if and only if the patient has a certain condition while disagreeing on the probability threshold above which the surgery is desired when unsure whether the condition exists.

Over the class of persuasion problems with ex post common interests, the sender is indifferent between all rules that correctly update beliefs in fully revealing experiments (e.g., Bayesian or Grether's $\alpha - \beta$ rules) and ranks those strictly above any rule that cannot get the receiver to hold deterministic beliefs (e.g., conservative Bayesianism or some extreme forms of probability weighing). Indeed, the sender's first-best payoff obtains when his preferred action gets implemented in each state, which the receiver will do after learning the state if her preference ordering coincides with that of the sender in each state.

If we further restrict attention to problems with purely common interests ($u = v$), then the sender systematically prefers less conservative receivers in the following sense. Consider two distortion functions D_{μ_0} and $\hat{D}_{\mu_0}$ mapping any belief ν to a convex combination of ν and μ_0: for each belief ν, there exist $\chi(\nu), \hat{\chi}(\nu) \in [0, 1]$ such that[19]

$$\begin{cases} D_{\mu_0}(\nu) = \chi(\nu)\mu_0 + (1 - \chi(\nu))\nu, \\ \hat{D}_{\mu_0}(\nu) = \hat{\chi}(\nu)\mu_0 + (1 - \hat{\chi}(\nu))\nu. \end{cases} \tag{6}$$

This covers, for instance, the case of conservative Bayesianism and motivated conservative Bayesianism discussed in examples 1 and 2. Say that D_{μ_0} is less conservative than $\hat{D}_{\mu_0}$ if distorted beliefs are systematically farther

[19] $\chi(\nu)$ and $\hat{\chi}(\nu)$ may also depend on μ_0.

away from the prior, that is, $\chi(\nu) \leq \hat{\chi}(\nu)$ for all ν. To prove that the sender prefers D_{μ_0} over $\hat{D}_{\mu_0}$, we show that the sender's expected utility for a belief ν is no lower when the receiver takes her optimal action a for the distorted belief $D_{\mu_0}(\nu)$ instead of her optimal action $\hat{a}$ for the distorted belief $\hat{D}_{\mu_0}(\nu)$. The statement being obvious when $a = \hat{a}$, we can assume without loss of generality that $\chi(\nu) < \hat{\chi}(\nu)$. Simple computations reveal that

$$\nu = \hat{\gamma} D_{\mu_0}(\nu) - \gamma \hat{D}_{\mu_0}(\nu),$$

where $\gamma = \chi(\nu)/(\hat{\chi}(\nu) - \chi(\nu))$ and $\hat{\gamma} = \hat{\chi}(\nu)/(\hat{\chi}(\nu) - \chi(\nu))$. By linearity of expected utility in beliefs, we have that the jump in the sender's expected utility under ν when the receiver takes a instead of $\hat{a}$ is equal to $\hat{\gamma}$ times the difference of expected utility under $D_{\mu_0}(\nu)$ minus γ times the difference of expected utility under $\hat{D}_{\mu_0}(\nu)$. As claimed, this jump is nonnegative, as $\gamma, \hat{\gamma} \geq 0$, $u = v$, and the receiver chooses a over $\hat{a}$ under $D_{\mu_0}(\nu)$ and $\hat{a}$ over a under $\hat{D}_{\mu_0}(\nu)$.

B. *Purely Opposed Interests*

By contrast, Bayesian updating is least preferred among all updating rules when restricting attention to purely opposed interests ($u = -v$). Indeed, any action that improves a Bayesian receiver's welfare decreases the sender's welfare (sharing a common belief after each signal realization), in which case running no experiment at all is optimal. Clearly, the sender can make that same choice, securing the same payoff, whatever the receiver's updating rule. In fact, he can oftentimes do strictly better (see example 14).

C. *Sender's Utility Is State Independent*

Let us restrict attention now to problems where the sender's utility is state independent. Formally, the sender *unambiguously prefers* Ann over Beth *whatever his state-independent utility* 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 (Ω, μ_0, A, (u, v), μ^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. Stronger sufficient and the same necessary conditions apply in this case: (a') the sender unambiguously prefers Ann over Beth whatever his state-independent utility if $T^R(\mu_0, \hat{\mu}^R) \subseteq T^R(\mu_0, \mu^R)$ and (b') if the sender unambiguously prefers Ann over Beth whatever his state-independent utility, then one cannot find a posterior that is feasible for Beth but not for Ann. Part a' follows again from (3), using the fact that the sender's utility is state independent. Proposition 6b is proved with a state-independent utility function for the sender and hence remains valid.

Let us apply this result. Any belief on states can occur with positive probability under Bayesian updating (remember that μ_0 has full support). Hence, by b', updating rules that do not share this property cannot be unambiguously preferred to μ^B whatever the sender's state-independent utility. And it follows from d' that the sender unambiguously (strictly) prefers receivers who are closer to rationality under conservative Bayesianism, whatever his state-independent utility, because $T^R(\mu_0, \mu^{CB\chi})$ is strictly decreasing in χ. By contrast, these rules were incomparable in the absence of restrictions on the sender's utility (see proposition 7 and example 13).

D. *Getting the Receiver to Switch Action*

Finally, let us restrict attention to persuasion problems where the sender's goal is to get the receiver to take an action a_1 instead of sticking with the status quo a_0. Assume without loss of generality that $u(a_0, \omega) = v(a_0, \omega) = 0$ for all ω, in which case $\Sigma_\omega \mu_0(\omega) u(a_1, \omega) < 0$ (the receiver needs persuading to pick a_1) and $v(a_1, \omega) > 0$ for all ω (the sender prefers a_1 over a_0; note that his utility can still vary with the state, if desired). We call it a switch action problem. As in the case of purely common interests but for different reasons, the sender systematically prefers less conservative receivers. Consider D_{μ_0} and $\hat{D}_{\mu_0}$ as in (6), with $\chi(\nu) \leq \hat{\chi}(\nu)$ for all ν. To prove that the sender prefers D_{μ_0} over $\hat{D}_{\mu_0}$, we show that for each potential Bayesian posterior $\nu \in \Delta(\Omega)$, a receiver who is willing to take action a_1 for the distorted belief $\hat{D}_{\mu_0}(\nu)$ would also do so for the distorted belief $D_{\mu_0}(\nu)$. Simple computations reveal that

$$D_{\mu_0}(\nu) = \hat{\delta}\hat{D}_{\mu_0}(\nu) - \delta\mu_0,$$

where $\delta = (\hat{\chi}(\nu) - \chi(\nu))/(1 - \hat{\chi}(\nu)) \geq 0$ and $\hat{\delta} = (1 - \chi(\nu))/(1 - \hat{\chi}(\nu)) \geq 1$. By linearity of expected utility in beliefs, we have that the jump in the receiver's expected utility under $D_{\mu_0}(\nu)$ by taking a_1 instead of a_0 is equal to $\hat{\delta}$ times the difference of expected utility under $\hat{D}_{\mu_0}(\nu)$ minus δ times the difference of expected utility under μ_0. As claimed, this jump is non negative, as the receiver chooses a_1 over a_0 under $\hat{D}_{\mu_0}(\nu)$ and a_0 over a_1 under μ_0.

VII. When Does the Sender Benefit from Persuasion?

Kamenica and Gentzkow's (2011) analysis highlights the following fact. By designing the right experiment, the sender can oftentimes nudge the receiver's decision to his advantage, even though the receiver is rational and aware of the sender's intent to persuade. How will the scope for persuasion change when the receiver does not update beliefs rationally?

We understand by now that mistakes in probabilistic inferences need not make persuasion easier. Thanks to section VI, we can say this: if the sender unambiguously prefers μ^R over $\hat{\mu}^R$ given a reference class of persuasion problems and benefits from persuasion given $\hat{\mu}^R$ for a problem in that class, then so does he given μ^R. Given the incompleteness of unambiguous preferences, more effort is required to understand circumstances under which persuasion is profitable.

In fact, we already have a characterization result: persuasion is profitable if and only if $[\mathrm{CAV}(\check{v})](\mu_0) > \hat{v}(\mu_0, \mu_0)$ (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 Kamenica and Gentzkow (2011). To address it, they propose a simpler condition that characterizes profitable persuasion for almost all priors 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 satisfies some regularity conditions.

As in Kamenica and Gentzkow (2011), we say the receiver's *preference is discrete* at belief ν' if the receiver's expected utility from her preferred action $\hat{a}(\nu')$ is bounded away from her expected utility from any other action, that is, if there is an $\varepsilon > 0$ such that $\forall\ a \neq \hat{a}(\nu')$, $E_{\nu'} u(\hat{a}(\nu'), \omega) > E_{\nu'} u(a, \omega) + \varepsilon$. This is copied verbatim from Kamenica and Gentzkow (2011), as it corresponds to a joint restriction on the receiver's belief and utility, which has nothing to do with how the receiver updates her beliefs. Clearly, the receiver's preference is discrete at almost all beliefs ν' when A is finite.

We say *there is information the 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 the sender had in his possession private information in the form of a signal realization that led him to believe ν, then he would prefer sharing that information with the receiver (leading her to believe $D_{\mu_0}(\nu)$) rather than having her act on the basis of the prior. This extends Kamenica and Gentzkow's (2011) property to reflect the fact that the receiver's posterior is distorted.

We say that the distortion function D_{μ_0} is *regular* if it is continuous and $D_{\mu_0}(\mu_0) = \mu_0$. With the three definitions above, we can now prove the following.

Proposition 8. Fix the prior μ_0 and any updating rule that systematically distorts updated beliefs with a regular distortion function. The two following properties hold. (*a*) If there is no information the sender would share at μ_0, then the sender does not benefit from persuasion. (*b*) If there is information the sender would share and the receiver's preference is discrete at μ_0 (which is generically true when A is finite), then the sender benefits from persuasion.

We see that Kamenica and Gentzkow's (2011) proposition 2 is quite robust. While regularity is not needed for part *a*, we show in the online appendix (by means of counterexamples) that part *b* does not extend to irregular distortion functions (and, a fortiori, more general updating rules). Let us apply this result to illustrate how mistakes in probabilistic inferences can have striking implications on persuasion. Intuitively, persuasion is most challenging in problems with purely opposed interests, and indeed, the sender cannot benefit from persuasion when the receiver is rational. Yet persuasion has a strictly positive value in essentially all problems of purely opposed interests when the receiver is subject to an overinference bias modeled as $\alpha = 1$ and $\beta > 1$ in Grether's rule.

Example 14. Consider a persuasion problem with purely opposed interests. In order to apply part *b* of the above proposition, we impose the mild requirements that A is finite and that the receiver's preference is discrete at the prior. To make the problem interesting, we also assume that $\hat{a}(\mu_0)$ is not the receiver's most preferred action at all beliefs. Suppose that the receiver is subject to an overinference bias modeled as $\alpha = 1$ and $\beta > 1$ in Grether's rule. We establish that persuasion has a strictly positive value to the sender (and hence strictly negative value to the receiver). Following the assumptions, there exists a state ω such that $\hat{a}(\mu_0)$ is strictly inferior to another action for the receiver when she believes the state is ω. For $0 < x \leq 1 - \mu_0(\omega)$, let $\nu_x \in \Delta(\Omega)$ be the modification of the prior that adds x to the probability of ω and keeps the relative likelihood of other states: $\nu_x(\omega) = \mu_0(\omega) + x$ and $\nu_x(\omega') = \mu_0(\omega')\{1 - [x/(1 - \mu_0(\omega))]\}$ for all $\omega' \neq \omega$. We show in the online appendix that there exists a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f(x) > x$ for all x and the distorted belief associated with ν_x is $\nu_{f(x)}$. This matches the intuition that $\beta > 1$ corresponds to overinference: the distorted belief associated with ν_{x} places even more probability on ω while keeping the same relative likelihood of other states. Let x^* be the infimum of the numbers $x > 0$ for which $\hat{a}(\mu_0)$ is strictly inferior to some other action given $\nu_{f(x)}$, and let $a^* \in A$ be such that a^* is the receiver's optimal action given $\nu_{f(x^*+\varepsilon)}$ for all $\varepsilon > 0$ small enough, then $E_{\nu_{f(x^*)}}[u(a^*, \omega) - u(\hat{a}(\mu_0), \omega)] = 0$. But we also have that $E_{\mu_0}[u(a^*, \omega) - u(\hat{a}(\mu_0), \omega)] < 0$. By linearity of expected utility in beliefs, $E_{\nu_{x^*}}[u(a^*, \omega) - u(\hat{a}(\mu_0), \omega)] < 0$, since ν_{x^*} can be written as a strict convex combination of μ_0 and $\nu_{f(x^*)}$ (remember that $x^* < f(x^*)$). Thus, when the accurate updated belief is ν_{x^*}, the receiver takes an action that gives the sender an expected payoff strictly larger than what he gets with $\hat{a}(\mu_0)$ (remember that $v = -u$). Hence, there is information the sender would share, and persuasion has a strictly positive value for the sender by the above proposition.

VIII. On the Possibility of Detrimental Persuasion

Giving the sender an opportunity to persuade is never detrimental to a Bayesian receiver. But real-life experiences suggest that persuasion can be harmful to the receiver (in terms of her accurate ex-ante welfare). This could take the form, for instance, of consumers being too easily persuaded of buying. By definition, persuasion benefitting the sender is harmful to the receiver in the case of purely opposed interests (see example 14). And, obviously, optimal persuasion cannot harm in the case of purely common interests. When considering mixtures of conflicting and overlapping interests, whether optimal persuasion is harmful typically varies from problem to problem. A couple of further observations are worth mentioning though.

Given an updating rule and a class of persuasion problems, one may ask whether there exists a problem in that class for which optimal persuasion is harmful to the receiver. Of course, if there exists such a problem in a small reference class, then so does it in larger classes. With that in mind, our first result is formulated for a rather small class of problems with mixed interests: for numerous updating rules, one can find a situation where the receiver is harmed in a simple switch action problem (see sec. VI.D).

Proposition 9. Suppose that there exist $\omega \in \Omega$ and $\mu_0, \nu \in \Delta(\Omega)$ such that D_{μ_0} is continuous, μ_0 is a strict convex combination of ν and δ_ω, and $(D_{\mu_0}(\nu) - \nu) \cdot (\delta_\omega - \nu) < 0$. Then optimal persuasion is harmful to the receiver in some switch action problem. This happens, for instance, when considering overinference in Grether's model ($\beta > 1$, whatever α is) or non-Bayesian rules associated with continuous distortion functions independent of μ_0.

With only two states, the assumption in the above proposition simply means that some updated belief ν under which the probability of ω is reduced compared with μ_0 further decreases the probability of ω once distorted. The assumption in the above proposition could be seen as an extension of this idea to any number of states. Figure 3 provides an illustration for three states: ν falls on the ray going through the prior and an extreme point of the probability simplex. The distorted belief $D_{\mu_0}(\nu)$ falls further beyond the hyperplane that is orthogonal to that ray and passes through ν.

It is possible, however, to find other updating rules retaining the property that optimal persuasion is never detrimental, as the next result shows.

Proposition 10. Optimal persuasion is never detrimental to the receiver in case of conservative Bayesianism.

In fact, conservative Bayesianism can sometimes leave the receiver with a larger expected payoff than when accurately updating beliefs.[20] Think of

[20] This echoes Sobel's (2013, 326–27) 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."

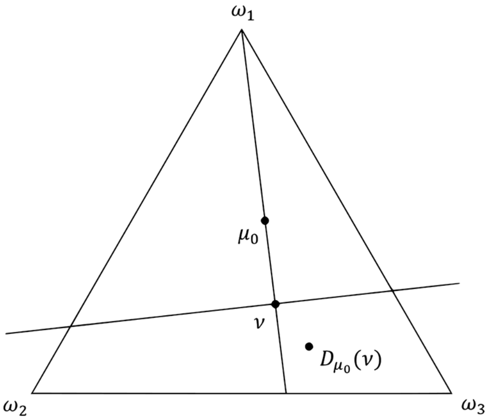

FIG. 3.—Illustration of proposition 9 conditions.

switch action problems, for instance. A Bayesian receiver does not benefit from optimal persuasion, as updated beliefs either confirm her choice of the status quo or leave her indifferent between the status quo and the alternative. By contrast, under conservative Bayesianism, such indifference means a strictly positive payoff under the true underlying posterior (as the increase of payoff that leads from a preference for the status quo under prior to indifference under the distorted belief is further amplified when considering the accurate posterior). That being said, it is not true that more conservatism is always better. Indeed, being persuaded becomes almost impossible if the receiver is extremely conservative in her updating, meaning that the receiver is far too rarely persuaded to choose the alternative over the status quo.

IX. Receiver's Action Varies with Expected State

When the state space is large, the standard concavification method has limited applicability. For this reason, Kamenica and Gentzkow (2011; also subsequent papers, including Gentzkow and Kamenica [2016] and Dworczak and Martini [2019]) extend their analysis to persuasion problems where the sender's preference is state independent and the 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 the receiver's beliefs ν'.

In this section, we establish that such techniques further extend to accommodate updating rules that systematically distort updated beliefs with affine distortion functions.[21] Indeed, we show next that the sender's

[21] 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.

optimal signals can be found by analyzing a modified problem where the receiver is rational but her utility is distorted.

Proposition 11. Suppose that μ^R systematically distorts updated beliefs with affine distortion functions $(D_{\mu_0})_{\mu_0 \in \Delta(\Omega)}$. 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 $\delta_\omega : \tilde{u}(a, \omega) = E_{D_{\mu_0}(\delta_\omega)} u(a, \cdot)$.

Since the sender's utility remains unchanged when transforming the original problem $(\Omega, \mu_0, A, (u, v), \mu^R)$ into its fictitious variant $(\Omega, \mu_0, A, (\tilde{u}, v), \mu^B)$, the sender's expected utility from the optimal signals coincides in both problems. The two problems are thus identical for the sender but for a small difference. The above result captures the best the sender can achieve by using some signal, but of course he also has the option to use no signal at all (recall the discussion at the end of sec. III). 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 toward a cause, the receiver may interpret uninformative evidence about its value as positive news. To account for this possibility, we must keep in mind that the sender benefits from persuasion if and only if his expected utility from the optimal signal identified in proposition 11 is strictly larger than $\hat{v}(\mu_0, \mu_0)$, the sender's expected utility in the absence of persuasion. To summarize, the 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 an illustration, recall that 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 us revisit, for instance, Kamenica and Gentzkow's (2011) example B in section V.

Example 15 (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 that 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 us apply proposition 11 to a consumer with $D_{\mu_0}^{\chi,\nu^*}$. With 0 denoting not buying and 1 denoting buying, the receiver's modified utility is given by

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

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

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

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

Using the terminology introduced for Bayesian persuasion in Kamenica and Gentzkow's (2011) section IV, we find that it is easy to check that preferences are more aligned when the outside option's value goes down. Thus, the sender's optimal profit goes up when $E_{\nu^*}[\omega]$ increases. Let us 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 the receiver less likely to buy, that is, if $E_{\nu^*}[\omega] < \underline{u}$. In that case, the 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 $1 \geq \chi > (1-\underline{u})/(1-E_{\nu^*}[\omega])$. 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 $1 \geq \chi \geq (u - E_{\mu_0}[\omega])/(E_{\nu^*}[\omega] - E_{\mu_0}[\omega])$.

We now turn our attention to consumer's accurate ex ante welfare. For this, we must know what the optimal persuasion strategy is, which is doable thanks to proposition 11 and corollary 2 of Dworczak and Martini (2019). It is optimal to reveal whether ω is below or above u^* where $E[\omega|\omega \geq u^*] = (\underline{u} - \chi E_{\nu^*}[\omega])/(1-\chi)$ (remember that the revelation principle holds, and hence two signal realizations are sufficient). With such a signal, the consumer's accurate ex ante 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), which is equal to her (actual) outside option value. This echoes proposition 9 but in a continuous state space and provides another example where persuasion is detrimental to the receiver with mistakes in probabilistic inferences. However, when the

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

Next, we revisit Kamenica and Gentzkow's (2011) example A in section V, with a lobbying group aiming to persuade a benevolent politician. Many rules that systematically distort updated beliefs are continuous and share the following two properties: (1) they do not distort degenerate posteriors (i.e., δ_ω is a fixed point of the distortion function for each ω) and (2) one can find a belief whose expected value is strictly smaller than that of its distortion.[22] We show that persuasion has a positive value for a larger set of parameters when the receiver uses a continuous rule satisfying these two rather mild properties instead of Bayesian updating. Furthermore, contrary to the Bayesian benchmark, each such rule admits a range of parameters under which optimal persuasion is partially—but not fully—revealing.

Example 16. Suppose $A = \Omega = [0, 1]$, the receiver is a benevolent politician with $u = -(a - \omega)^2$, and the sender is a biased lobbying group with $v = -(a - a^*(\omega))^2$, where ω represents the socially optimal policy and $a^*(\omega) = \gamma\omega + (1 - \gamma)\omega^*$ captures the lobbying group's biased goal (for some $\omega^* > 1$ and $\gamma \in [0, 1]$). With a rational politician, Kamenica and Gentzkow (2011) show that optimal persuasion is fully revealing when it has a strictly positive value, which happens if and only if $\gamma > 1/2$. Hence, any rule satisfying property 1 above retains the property of having a strictly positive value over that range of parameters. We now construct a simple experiment with a strictly positive value when $\gamma = 1/2$ for all continuous updating rules satisfying both properties 1 and 2. By continuity, this implies that the sender benefits from persuasion for a range of γ's below 1/2. Of course, this payoff is not achievable using fully revealing experiments, as the resulting distribution over posterior pairs coincides with that of a Bayesian politician, given property 1.[23]

[22] For instance, take Grether's rule. Assume $\alpha = 1$ to focus on biases of over- and underinference (as base rate neglect would behave similarly to a case of noncommon prior, which is covered in Alonso and Câmara [2016]). These two properties are satisfied for any $\beta \neq 1$. In other words, Bayesian updating is the only rule in this class that does not satisfy both properties (of course, it violates property 2).

[23] Even for γ strictly above but close enough to 1/2, the optimal experiment will have to be partially revealing, as the optimal value of fully revealing experiments goes down to zero as γ tends to 1/2, while there are partially revealing experiments that have a positive value at $\gamma = 1/2$.

Let ν be a belief such that $\bar{a} = E_\nu[\omega] < E_{\hat{\nu}}[\omega]$, where $\hat{\nu} = D_{\mu_0}(\nu)$, and define a number $\pi(s|\omega) \in [0, 1]$ for each $\omega \in [0, 1]$ such that $\mu_s^B(\omega; \mu_0, \pi) = \nu(\omega)$. In addition to s, consider a signal realization s_ω with $\pi(s_\omega|\omega) = 1 - \pi(s|\omega)$ for each ω such that $\pi(s|\omega) < 1$. This way, π is a well-defined signal, with at most two realizations (s and s_ω) occurring with strictly positive probability under any state ω. By construction, the rational posterior is ν after the realization s. Thus, the receiver takes action $\hat{a} = E_{\hat{\nu}}[\omega]$ after s. The sender's expected payoff in that event is $E_\nu[-(\hat{a} - a^*(\omega))^2]$, which is strictly larger than his expected payoff $E_\nu[-(\bar{a} - a^*(\omega))^2]$ for a Bayesian receiver (because $\bar{a} < \hat{a} < \omega^*$ and $\gamma = 1/2$). The receiver has the posterior δ_ω after the realization s_ω and takes action ω. The sender's payoff in that event coincides with what he would get with a Bayesian receiver. By Kamenica and Gentzkow (2011), all experiments give the sender the same expected payoff with a Bayesian receiver when $\gamma = 1/2$, but the experiment we designed gives him a strictly larger payoff when the receiver uses a continuous rule satisfying properties 1 and 2. Hence, persuasion has a strictly positive value in that case, as claimed.

X. Extensions

A. Belief over Updating Rules

Suppose that the sender is unsure about the 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: $(\mu_s^B(\cdot; \mu_0, \pi), \mu_s^R(\cdot; \mu_0, \pi))$ arises with probability $\lambda(\mu^R)\Sigma_\omega \pi(s|\omega)\mu_0(\omega)$. 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-receiver public information case in Kamenica and Gentzkow's (2011) section VI.B.

When focusing on rules that systematically distort updated beliefs, we can think of λ as a distribution over distortion functions. For instance, the sender may use Grether (1980) to model the 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 the sender's indirect utility for Bayesian posteriors. Indeed, the 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).[24] Proposition 1 extends, and optimal persuasion requires at most $|\Omega|$ signal realizations.

B. Distortion Correspondence

Suppose that the sender models the receiver's probabilistic inferences using Bayesian updating as a benchmark but fears that she may make some limited mistakes. For instance, if the receiver does not make careful computations, her perceived posteriors may be in the ballpark of—but not always equal to—the correct posterior. To fix ideas, suppose that the sender is confident that the receiver's posterior falls within distance ε of the Bayesian posterior ν but is concerned though that any posterior within the ball $B(\nu, \varepsilon)$ is possible. The sender may then be interested in finding a signal that guarantees him a good profit whatever the mistakes the receiver may make within those bounds. Hence, the 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, the 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*.[25] By selecting the worst posterior for the 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.[26] Because the information

[24] 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}.

[25] 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.

[26] Hu and Weng (2018), Dworczak and Pavan (2020), and Kosterina (2020) also study robustness in information design problems. However, the uncertainty the sender faces in these papers is over the receiver's exogenous private information, not the receiver's rationality in belief updating.

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 zero as soon as $\hat{v}$ is continuous.[27]

Robust persuasion has intuitive implications. Consider Kamenica and Gentzkow's (2011) 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 ξ^* larger than the prior (i.e., 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 ξ^*. 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 that 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 $\xi^* + \varepsilon$. Indeed, v_ε remains a step function as in the Bayesian case but with a threshold at $\xi^* + \varepsilon$ instead of ξ^*.

With a distortion correspondence, one may also consider the case where the sender can pick the best posterior for himself within $B(\nu, \varepsilon)$ by substituting the inf operator in $v_\varepsilon(\nu)$ with sup; then intuitively, a receiver is unambiguously more gullible than another if her ε is larger. This captures the idea that the sender can nudge naive receivers by labeling or presenting the content of signal realizations differently, like a prosecutor using a variety of rhetorical tricks to present evidence. This idea is related to recent papers on falsification in Bayesian persuasion (Lipnowski, Ravid, and Shishkin 2019; Li and Lim 2020; Guo and Shmaya 2021; Nguyen and Tan 2021; Perez-Richet and Skreta 2021), except that the sender can induce in our case multiple posteriors by choosing a right narrative and that the receiver is unaware of being nudged in our approach. We hope that this direction will be further explored in the future.

Appendix

A1. 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 $\mu_s^R(\cdot; \mu_0, \pi) = D_{\mu_0}(\mu_s^B(\cdot; \mu_0, \pi))$. To do this, we apply the assumption for sufficiency to show that $\mu_s^R(\cdot; \mu_0, \pi) = \mu_s^R(\cdot; \mu_0, \hat{\pi}_\nu)$, where $\nu = \mu_s^B(\cdot; \mu_0, \pi)$. Notice that, by Bayes rule, $\pi(s|\omega) = 0$ if and only if $\nu(\omega) = 0$,

[27] Hu and Weng (2018) also show such continuity as the sender's uncertainty about the receiver's private information vanishes.

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 $\hat{\pi}(\hat{s}|\omega)/\pi(s|\omega)$ 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 $\hat{\pi}(\hat{s}|\omega)/\pi(s|\omega)$ 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 $\mu_s^R(\cdot; \mu_0, \pi) = \mu_{\hat{s}}^R(\cdot; \mu_0, \hat{\pi})$ or, equivalently, that $\hat{D}_{\mu_0}(\mu_s^B(\cdot; \mu_0, \pi)) = \hat{D}_{\mu_0}(\mu_{\hat{s}}^B(\cdot; \mu_0, \hat{\pi}))$. To establish this last equality, we simply check that $\mu_s^B(\cdot; \mu_0, \pi) = \mu_{\hat{s}}^B(\cdot; \mu_0, \hat{\pi})$. 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 $\hat{D}_{\mu_0} = D_{\mu_0}$. QED

A2. Proof of Proposition 3

By proposition 2, we can find two signal realization pairs (π, s) and $(\hat{\pi}, \hat{s})$ such that (*a*) the likelihood ratio $\hat{\pi}(\hat{s}|\omega)/\pi(s|\omega)$ is constant over the set of ω's for which $\pi(s|\omega) > 0$, (*b*) $\hat{\pi}(\hat{s}|\omega) = 0$ if and only if $\pi(s|\omega) = 0$, and yet (*c*) $\mu_s^R(\cdot; \mu_0, \pi) \neq \mu_{\hat{s}}^R(\cdot; \mu_0, \hat{\pi})$. By assumption *b* on μ^R, this implies in particular that s ($\hat{s}$) occurs with strictly positive probability for at least two states under π ($\hat{\pi}$) and that $\mu_s^R(\cdot; \mu_0, \pi)$ ($\mu_{\hat{s}}^R(\cdot; \mu_0, \hat{\pi})$) places strictly positive probability on at least two states.

Switching the roles of π and $\hat{\pi}$ if needed, we can assume without loss of generality that $\hat{\pi}(\hat{s}|\omega) \geq \pi(s|\omega)$ for all ω. By continuity of μ^R, we can assume without loss of generality that the constant number $\hat{\pi}(\hat{s}|\omega)/\pi(s|\omega)$ described in assumption *a* is rational. In fact, we can assume without loss of generality that the ratio is an integer K. Say the ratio is i/j with $i > j$. Then consider a new experiment π' with the same set of signal realizations as π, and $\pi'(s|\omega) = \pi(s|\omega)/j$ for all ω. If $\mu_s^R(\cdot; \mu_0, \pi) \neq \mu_s^R(\cdot; \mu_0, \pi')$, then we can pursue the reasoning with π in the role of $\hat{\pi}$ and π' in the role of π ($K = j$ in this case). Otherwise, we simply pursue the reasoning with $\hat{\pi}$ and π' in the role of π ($K = i$ in this case).

Consider now the experiment π^* with signal realizations $\{s_1, \dots, s_K\} \cup \{s_\omega|\omega \in \Omega\}$, with $\pi^*(s_k|\cdot) = \pi(s|\cdot)$ for each $k = 1, \dots, K$, $\pi^*(s_\omega|\omega) = 1 - K\pi(s|\omega)$, and $\pi^*(s_{\omega'}|\omega) = 0$ for all $\omega \neq \omega'$. By assumption *a* on μ^R, $\mu_{s_k}^R(\cdot; \mu_0, \pi^*) = \mu_s^R(\cdot; \mu_0, \pi)$ for each k. Define now two actions a and d along with utility vectors

$u(a, \cdot)$ and $u(a', \cdot)$, such that a gives a strictly larger utility than a' given $\mu_s^R(\cdot; \mu_0, \pi)$, and vice versa given $\mu_{\hat{s}}^R(\cdot; \mu_0, \hat{\pi})$ (which is possible since these two beliefs are distinct). For each ω, define an action a_ω and a utility vector $u(a_\omega, \cdot)$ such that $u(a_\omega, \omega) = \max\{u(a, \omega), u(a', \omega)\} + 1$ and $u(a_\omega, \omega')$ is very small for each $\omega \neq \omega'$. This way, we get that a_ω is best for the receiver given s_ω (by assumption *b*) and that each a_ω is worse than both a and a' at both $\mu_s^R(\cdot; \mu_0, \pi)$ and $\mu_{\hat{s}}^R(\cdot; \mu_0, \hat{\pi})$ (since they place strictly positive probability on at least two states). In particular, a is best for the receiver given $\mu_s^R(\cdot; \mu_0, \pi)$. The reduced experiment associated with π^* has a unique signal realization, call it s_a, associated with all s_k, as they all lead the receiver to take the same action a. The probability of s_a under this reduced experiment coincides with that of $\hat{s}$ under $\hat{\pi}$. Hence, by assumption *a*, the receiver's updated belief given s_a is $\mu_{\hat{s}}^R(\cdot; \mu_0, \hat{\pi})$. But a' is best for the receiver given that belief, in contradiction to the revelation principle. QED

A3. Proof of Proposition 4

Given an arbitrary compact action set A, an arbitrary utility function $u(a, \omega)$ for the 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 the receiver's posterior after observing realization s. Define a straightforward signal with $S' = A$ and $\pi'(a|\omega) = \Sigma_{s \in S^a} \pi(s|\omega)$. 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 $\gamma_s \geq 0$ for all $s \in S^a$, $\Sigma_{s \in S^a}\gamma_s = 1$, and

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

so we have

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

Since for all $s \in S^a$, a maximizes $\Sigma_\omega u(a',\omega)\mu_s^R(\omega;\mu_0,\pi)$, 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. QED

A4. Proof of Proposition 5

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 $a \neq \operatorname{argmax}_{a'\in A}\Sigma_\omega u(a',\omega)\mu_a^R(\cdot;\mu_0,\pi') = \operatorname{argmax}_{a'\in A}\Sigma_\omega u(a',\omega)D_{\mu_0}(\mu_0)(\omega)$ for the noninformative 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 $\nu^* = D_{\mu_0}(\lambda\nu_1 + (1-\lambda)\nu_2)$. Without loss of generality, assume that 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 the 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 $\tau(\nu_1)/\tau(\nu_2) = \lambda/(1-\lambda)$. 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 $\Delta(\Omega)$ that contains ν_1' and ν_2' but not ν_3' or 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 the sender and receiver, such that $S^a = \{s_1, s_2\}$ for π yet $a \neq \arg\max_{a'\in A}\Sigma_\omega u(a',\omega)\mu_a^R(\cdot;\mu_0,\pi') = \arg\max_{a'\in A}\Sigma_\omega u(a',\omega)\nu^*(\omega)$, 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 $a \neq \arg\max_{a'\in A}\Sigma_\omega u(a',\omega)\mu_a^R(\cdot;\mu_0,\pi') = \operatorname{argmax}_{a'\in A}\Sigma_\omega u(a',\omega)D_{\mu_0}(\mu_0)(\omega)$, 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 the receiver's posteriors ν^* and ν_3', and we are back at case 1.2. This finishes the proof. QED

A5. *Proof of Proposition 6*

Property a follows from (2). As for the necessary condition presented in property b, assume that there is a posterior ν' feasible for Beth but not for Ann. We construct a persuasion problem where the sender gets a strictly larger payoff with Beth. Suppose first that $\hat{\nu}(\omega) < 1$ for all ω. Consider then the action set $A = \{a_\omega | \omega \in \Omega\} \cup \{a^*\}$. The 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'. The receiver's utility is defined as follows: $u(a^*, \omega) = 0$, $u(a_\omega, \omega) = 1$, and $u(a_\omega, \omega') = -\hat{\nu}(\omega)/(\Sigma_{\omega''\neq\omega}\hat{\nu}(\omega''))$ for all ω and all $\omega' \neq \omega$. Notice that for all ω, the receiver's expected utility of a_ω is zero should her belief be $\hat{\nu}$. For any other belief ν'', there exists a state ω such that $\nu''(\omega) > \hat{\nu}(\omega)$, which implies that her expected utility of a_ω is strictly positive should her belief be ν'':

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

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

A6. *Proof of Proposition 7*

Let D_{μ_0} ($\hat{D}_{\mu_0}$) be the distortion function associated with μ^R ($\hat{\mu}^R$). 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 the 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*b* 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*b* 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 $\nu(\omega^*)/(1 - \nu(\omega^*))$ and $\hat{\nu}(\omega^*)/(1 - \hat{\nu}(\omega^*))$. As in proposition 6*b* the sender's payoff is zero whenever the receiver's posterior is different from ν'. Given μ^R, the rational belief associated with that receiver's posterior is ν, in which case the sender gets a strictly positive expected payoff. Things are different, however, when the receiver updates beliefs according to $\hat{\mu}^R$: the rational belief associated with it is now $\hat{\nu}$, in which case the sender gets a strictly negative expected payoff. In that case, the sender's optimal persuasion payoff is zero, which is strictly inferior to what he gets when the receiver updates according to μ^R. QED

A7. *Proof of Proposition 8*

For the first part, suppose that there is no information the 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

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

Thus, the sender does not benefit from persuasion.

For the second part, since there is information the sender would share at μ_0, $\check{v}(\nu_h) > \hat{v}(\nu_h,\mu_0)$. As in Kamenica and Gentzkow (2011), since the 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 that 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. QED

A8. Proof of Proposition 9

Let $u(a_0,\omega') = v(a_0,\omega') = 0$, $v(a_1,\omega') = 1$, and

$$u(a_1,\omega') = \nu(\omega') - \delta_\omega(\omega') + (\delta_\omega - \nu)\cdot\nu$$

for all ω'. Notice that the receiver's utility for a_1 at ω is $-(1-\nu(\omega))^2 - \Sigma_{\omega'\neq\omega}(\nu(\omega'))^2$, which is strictly negative. Also, her expected utility for a_1 given ν is zero. Given that μ_0 is a strict convex combination of ν and δ_ω, it must be that the receiver's utility for a_1 is strictly negative when her belief is the prior. Hence, she picks a_0 in that case and gets a zero payoff.

Let τ^S be the distribution over Bayesian posteriors achieving optimal persuasion. Let P_i be the subset of posteriors ν' in its support so that the receiver does pick a_i given $D_{\mu_0}(\nu')$. The sender's expected payoff is $\tau^S(P_1)$. Since $\mu_0 = \lambda\nu + (1-\lambda)\delta_\omega$, the sender also has the option to select a signal that triggers the Bayesian posterior ν with probability λ and δ_ω with probability $1-\lambda$. When ν occurs, the receiver's belief, $D_{\mu_0}(\nu)$, makes her pick a_1 since $(D_{\mu_0}(\nu) - \nu)\cdot(\delta_\omega - \nu) < 0$. The sender's expected payoff is at least λ in that case. We now prove that $\lambda \geq \tau^S(P_1)$ if the receiver's accurate expected welfare under τ^S is nonnegative.

Notice that $(1-\lambda)(\delta_\omega - \mu_0) = \lambda(\mu_0 - \nu)$. Performing an inner product with $\delta_\omega - \nu$ and rearranging terms, we obtain

$$\frac{1}{\lambda} - 1 = \frac{(\mu_0 - \nu)\cdot(\delta_\omega - \nu)}{(\delta_\omega - \mu_0)\cdot(\delta_\omega - \nu)}.$$

By the martingale property of Bayesian updating, $\mu_0 = \tau^S(P_1)\bar{\nu}_1 + (1-\tau^S(P_1))\bar{\nu}_0$, where $\bar{\nu}_i = \Sigma_{\nu\in P_i}(\tau^S(\nu)/\tau^S(P_i))\nu$. Similar computations reveal that

$$\frac{1}{\tau^S(P_1)} - 1 = \frac{(\mu_0 - \bar{\nu}_1)\cdot(\delta_\omega - \nu)}{(\bar{\nu}_0 - \mu_0)\cdot(\delta_\omega - \nu)}.$$

Thus, $\lambda \geq \tau^S(P_1)$ if and only if

$$\frac{(\delta_\omega - \mu_0)\cdot(\delta_\omega - \nu)}{(\mu_0 - \nu)\cdot(\delta_\omega - \nu)} \geq \frac{(\bar{\nu}_0 - \mu_0)\cdot(\delta_\omega - \nu)}{(\mu_0 - \bar{\nu}_1)\cdot(\delta_\omega - \nu)}.$$

Clearly, $x\cdot(\delta_\omega - \nu)$ is maximized over $\Delta(\Omega)$ when $x = \delta_\omega$. Hence, the numerator on the left-hand side is larger than its counterpart on the right-hand side. It remains to show that the opposite inequality holds for the denominators, or that $(\nu - \bar{\nu}_1)\cdot(\delta_\omega - \nu) \geq 0$. This is precisely the condition for the receiver's true expected welfare under τ^S to be nonnegative.

We have thus shown that optimal persuasion giving the receiver a nonnegative expected payoff gives the sender an expected payoff that is smaller or equal to what he gets by selecting the signal triggering ν with probability λ and δ_ω otherwise. But the sender can do better, thereby contradicting the fact that optimal persuasion leaves the receiver with a nonnegative expected payoff: he can select another signal triggering δ_ω and a ν' that is closer to μ_0 on the segment between μ_0 and ν. By the martingale property, the weight on ν' has to be strictly larger than λ. And if ν' is close enough to ν, the receiver will still want to take a_1 given $D_{\mu_0}(\nu')$ because D_{μ_0} is continuous and her expected utility for a_1 under $D_{\mu_0}(\nu')$ is strictly positive. Hence, optimal persuasion must leave the receiver with a strictly negative expected payoff.

Having proved the main result, we now check that the classes of updating rules mentioned in the statement indeed verify the assumptions. We start with Grether's rule. Suppose that $\mu_0(\omega) = \mu_0(\omega') = 1/2$ (while other states have zero probability, which we will adjust in a moment). Then, under Grether's rule, the belief ν defined by $\nu(\omega) = 1/4$ and $\nu(\omega') = 3/4$ is distorted into $\hat{\nu}$, defined by $\hat{\nu}(\omega) = 1/(1+3^\beta)$ and $\hat{\nu}(\omega') = 1 - \hat{\nu}(\omega)$ (independently of α). Thus, ω' further increases in probability under $\hat{\nu}$ compared with ν for $\beta > 1$, and it is easy to see that the assumptions of the main result are satisfied. But priors are assumed to have full support throughout the paper. Since Grether's rule is continuous, clearly there will be a prior near μ_0 and a posterior near ν for which the assumptions of the main result hold.

Let us now turn our attention to distortion functions that are independent of the prior (see, e.g., example 1 when $\nu^* \neq \mu_0$ and example 6 when probability weighting is not applied to the prior). Let ν be such that $D(\nu) \neq \nu$. By continuity, we can assume without loss that ν is not an extreme point of the probability simplex. We now claim that there exists ω such that $(D(\nu) - \nu) \cdot (\delta_\omega - \nu) < 0$. Otherwise, $(D(\nu) - \nu) \cdot (\delta_\omega - \nu) \geq 0$ for all ω. But ν is a convex combination of the δ_ω's. The weighted sum of the numbers on the left-hand side of these inequalities, using the weights for such a decomposition of ν, is zero. Hence, it must be that $(D(\nu) - \nu) \cdot (\delta_\omega - \nu) = 0$ for all ω. The system of equations $(x - \nu) \cdot (\delta_\omega - \nu) = 0$ for all ω, with $x \in \Delta(\Omega)$, admits a unique solution. Since ν satisfies them, $D(\nu) \neq \nu$ cannot, a contradiction. Having established this way the existence of a state ω such that $(D(\nu) - \nu) \cdot (\delta_\omega - \nu) < 0$, we can fulfill all the assumptions of the main result by selecting any prior that is a strict convex combination of ν and δ_ω. QED

A9. Proof of Proposition 10

Let a_0 be the action the receiver takes under μ_0. Let τ^R be the distribution over the receiver's posteriors achieving optimal persuasion. Then, given any belief ν in the support, the receiver prefers to take her optimal action $\hat{a}(\nu)$ over a_0. Following the reasoning presented when discussing switch action problems in section VI (letting D_{μ_0} be Bayesian updating in this case), we get that the receiver would also prefer $\hat{a}(\nu)$ over a_0 under the correct updated belief that got distorted into ν. Hence, optimal persuasion cannot be detrimental for the receiver's true welfare. QED

A10. Proof of Proposition 11

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, 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 $\nu_1 \neq \nu_2 \in \Delta(\Omega)$, 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 lemma 7.2.2 in Leadbetter, Cambanis, and Pipiras (2014),

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

Therefore, $\check{a}(\nu) \equiv \arg\max_{a \in A} E_\nu u'(a, \omega) = \hat{a}(D_{\mu_0}(\nu))$. The 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 $(\Omega, \mu_0\ A, (u', v), \mu^B)$. QED

References

Alonso, R., and O. Câmara. 2016. “Bayesian Persuasion with Heterogeneous Priors.” *J. Econ. Theory* 165:672–706.

Anunrojwong, J., K. Iyer, and D. Lingenbrink. 2020. “Persuading Risk-Conscious Agents: A Geometric Approach.” Working paper.

Augenblick, N., and M. Rabin. 2021. “Belief Movement, Uncertainty Reduction, and Rational Updating.” *Q.J.E.* 136 (2): 933–85.

Aumann, R. J., M. Maschler, and R. E. Stearns. 1995. *Repeated Games with Incomplete Information.* Cambridge, MA: 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 *Handbook of Behavioral Economics: Foundations and Applications*, vol. 2, edited by B. D. Bernheim, S. DellaVigna, and D. Laibson, 69–186. Amsterdam: North-Holland.

Benjamin, D. J., M. Rabin, and C. Raymond. 2016. “A Model of Nonbelief in the Law of Large Numbers.” *J. European Econ. Assoc.* 14 (2): 515–44.

Bergemann, D., and S. Morris. 2016a. “Bayes Correlated Equilibrium and the Comparison of Information Structures in Games.” *Theoretical Econ.* 11 (2): 487–522.

———. 2016b. “Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium.” *A.E.R.* 106 (5): 586–91.

———. 2019. “Information Design: A Unified Perspective.” *J. Econ. Literature* 57 (1): 44–95.

Bloedel, A. W., and I. R. Segal. 2018. “Persuasion with Rational Inattention.” Working paper.

Brunnermeier, M. K., and J. A. Parker. 2005. “Optimal Expectations.” *A.E.R.* 95 (4): 1092–118.

Camerer, C. 1998. “Bounded Rationality in Individual Decision Making.” *Experimental Econ.* 1 (2): 163–83.

Chauvin, K. P. 2019. “Euclidean Properties of Bayesian Updating.” Working paper.

Crawford, V. P. 2021. “Efficient Mechanisms for Level-*k* Bilateral Trading.” *Games and Econ. Behavior* 127:80–101.

Cripps, M. W. 2018. “Divisible Updating.” Working paper.

Danziger, S., J. Levav, and L. Avnaim-Pesso. 2011. “Extraneous Factors in Judicial Decisions.” *Proc. Nat. Acad. Sci. USA* 108 (17): 6889–92.

de Clippel, G. 2014. “Behavioral Implementation.” *A.E.R.* 104 (10): 2975–3002.

de Clippel, G., R. Saran, and R. Serrano. 2019. “Level-*k* Mechanism Design.” *Rev. Econ. Studies* 86 (3): 1207–27.

———. 2021. “Continuous Level-*k* Mechanism Design.” Working paper.

DeGroot, M. H. 1974. “Reaching a Consensus.” *J. American Statis. Assoc.* 69 (345): 118–21.

Dominiak, A., M. Kovach, and G. Tserenjigmid. 2021. “Minimum Distance Belief Updating with General Information.” Working paper.

Dworczak, P., and G. Martini. 2019. “The Simple Economics of Optimal Persuasion.” *J.P.E.* 127 (5): 1993–2048.

Dworczak, P., and A. Pavan. 2020. “Preparing for the Worst but Hoping for the Best: Robust (Bayesian) Persuasion.” Working paper.

Edwards, W. 1968. “Conservatism in Human Information Processing.” In *Formal Representation of Human Judgment*, edited by B. Kleinmuntz, 17–52. New York: Wiley.

Eliaz, K., R. Spiegler, and H. C. Thysen. 2021a. “Persuasion with Endogenous Misspecified Beliefs.” *European Econ. Rev.* 134:103712.

———. 2021b. “Strategic Interpretations.” *J. Econ. Theory* 192:105192.

Epstein, L. G. 2006. “An Axiomatic Model of Non-Bayesian Updating.” *Rev. Econ. Studies* 73 (2): 413–36.

Epstein, L. G., J. Noor, and A. Sandroni. 2008. “Non-Bayesian Updating: A Theoretical Framework.” *Theoretical Econ.* 3 (2): 193–229.

Gabaix, X. 2019. “Behavioral Inattention.” In *Handbook of Behavioral Economics: Foundations and Applications*, vol. 2, edited by B. D. Bernheim, S. DellaVigna, and D. Laibson, 261–343. Amsterdam: North-Holland.

Galperti, S. 2019. “Persuasion: The Art of Changing Worldviews.” *A.E.R.* 109 (3): 996–1031.

Gentzkow, M., and E. Kamenica. 2016. “A Rothschild-Stiglitz Approach to Bayesian Persuasion.” *A.E.R.* 106 (5): 597–601.

Grether, D. M. 1980. “Bayes Rule as a Descriptive Model: The Representativeness Heuristic.” *Q.J.E.* 95 (3): 537–57.

Guo, Y., and E. Shmaya. 2021. “Costly Miscalibration.” *Theoretical Econ.* 16 (2): 477–506.

Guthrie, C., J. J. Rachlinski, and A. J. Wistrich. 2001. “Inside the Judicial Mind.” *Cornell Law Rev.* 86 (4): 814.

———. 2007. “Blinking on the Bench: How Judges Decide Cases.” *Cornell Law Rev.* 93:1.

Hagmann, D., and G. Loewenstein. 2017. “Persuasion with Motivated Beliefs.” Working paper.

Hu, J., and X. Weng. 2018. “Robust Persuasion of a Privately Informed Receiver.” Working paper.

Jadbabaie, A., P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. 2012. “Non-Bayesian Social Learning.” *Games and Econ. Behavior* 76 (1): 210–25.

Kahneman, D., and A. Tversky. 1973. “On the Psychology of Prediction.” *Psychological Rev.* 80 (4): 237–51.

Kamenica, E., and M. Gentzkow. 2011. “Bayesian Persuasion.” *A.E.R.* 101 (6): 2590–615.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2017. “Human Decisions and Machine Predictions.” *Q.J.E.* 133 (1): 237–93.

Kneeland, T. 2020. “Mechanism Design with Level-*k* Types: Theory and an Application to Bilateral Trade.” Working paper, Univ. Coll. London.

Koehler, J. J. 2002. “When Do Courts Think Base Rate Statistics Are Relevant?” *Jurimetrics*, 373–402.

Kosterina, S. 2020. “Persuasion with Unknown Beliefs.” Working paper.

Kovach, M. 2021. “Conservative Updating.” Working paper.

Leadbetter, R., S. Cambanis, and V. Pipiras. 2014. *A Basic Course in Measure and Probability: Theory for Applications.* Cambridge: Cambridge Univ. 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.” Working paper, London School Econ.

———. 2018b. “Persuasion with Correlation Neglect: Media Power via Correlation of News Content.” CEPR Discussion Paper no. DP12640, Center Econ. Policy Res, London.

Li, R., and W. Lim. 2020. “Persuasion with Strategic Reporting.” Working paper.

Lindsey, S., R. Hertwig, and G. Gigerenzer. 2002. “Communicating Statistical DNA Evidence.” *Jurimetrics* 43:147–63.

Lipnowski, E., and L. Mathevet. 2018. “Disclosure to a Psychological Audience.” *American Econ. J. Microeconomics* 10 (4): 67–93.

Lipnowski, E., L. Mathevet, and D. Wei. 2020. “Attention Management.” *A.E.R. Insights* 2 (1): 17–32.

Lipnowski, E., D. Ravid, and D. Shishkin. 2019. “Persuasion via Weak Institutions.” Working paper.

Mathevet, L., J. Perego, and I. Taneva. 2020. “On Information Design in Games.” *J.P.E.* 128 (4): 1370–404.

Molavi, P., A. Tahbaz-Salehi, and A. Jadbabaie. 2018. “A Theory of Non-Bayesian Social Learning.” *Econometrica* 86 (2): 445–90.

Myerson, R. B. 1991. *Game Theory: Analysis of Confllict.* Cambridge, MA: Harvard Univ. Press.

Nguyen, A., and T. Y. Tan. 2021. “Bayesian Persuasion with Costly Messages.” *J. Econ. Theory* 193:105212.

Oury, M., and O. Tercieux. 2012. “Continuous Implementation.” *Econometrica* 80 (4): 1605–37.

Perez-Richet, E., and V. Skreta. 2021. “Test Design under Falsification.” CEPR Discussion Paper no. DP15627, Center Econ. Policy Res, London.

Rabin, M., and J. L. Schrag. 1999. “First Impressions Matter: A Model of Confirmatory Bias.” *Q.J.E.* 114 (1): 37–82.

Shmaya, E., and L. Yariv. 2009. “Foundations for Bayesian Updating.” Working paper, California Inst. Tech.

Sobel, J. 2013. “Giving and Receiving Advice.” *Advances Econ. and Econometrics* 1:305–41.

Taneva, I. 2019. “Information Design.” *American Econ. J. Microeconomics* 11 (4): 151–85.

Wei, D. 2018. “Persuasion under Costly Learning.” Working paper.

Zhao, C. 2022. “Pseudo-Bayesian Updating.” *Theoretical Econ.* 17 (1): 253–89.