On the Bayesian Rational Assumption in Information Design

Wei Tang, Chien-Ju Ho
Washington University in St. Louis
{w.tang, chienju.ho}@wustl.edu

Abstract

We study the problem of information design in human-in-the-loop systems, where the *sender* (the system) aims to design an information disclosure policy to influence the *receiver* (the user) in making decisions. This problem is ubiquitous in systems with humans in the loop, e.g., recommendation systems might choose whether to present others' reviews to encourage users to follow recommendations, online retailers might choose which set of product features to present to persuade buyers to make the purchase. Among the flourish literature on information design, Bayesian persuasion has been one of the most prominent efforts in formalizing this problem and has spurred various research studies in both economics and computer science. While there has been significant progress in characterizing the optimal information disclosure policies and the corresponding computational complexity, one common assumption in this line of research is that the receiver is *Bayesian rational*, i.e., the receiver processes the information in a Bayesian manner and takes actions to maximize her expected utility. However, as empirically observed in the literature, this assumption might not be true in real-world scenarios. In this work, we relax this common Bayesian rational assumption in information design in the persuasion setting. In particular, we develop an alternative framework for information design based on discrete choice model and probability weighting to account for this relaxation. Moreover, we conduct online behavioral experiments on Amazon Mechanical Turk and demonstrate that our framework better explains real-world user behavior and leads to more effective information design policy.

Introduction

We study the problem of information design in human-in-the-loop systems, in which an informed *sender* (i.e., the system) aims to influence a *receiver* (i.e., humans in the system) in making decisions through designing information disclosure strategies. This problem is ubiquitous in our daily life. For example, online retailers might highlight a subset of product features to influence the buyers to make the purchases. Recommendation systems might selectively display other users' ratings to persuade users to take the recommendation. Public health officials might decide how to present vaccine information to encourage the general public to take vaccines to curb the pandemic. There have been various research efforts devoted to this problem from both economics (McCloskey and Klamer 1995; Rayo and Segal 2010; Gehlbach and Sonin 2014; Goldstein and Leitner 2018) and computer science (Dughmi and Xu 2019; Emek et al. 2014). Among the growing literature on the study of information design, the model of Bayesian persuasion proposed by Kamenica and Gentzkow (2011) is one of the most prominent ones and has inspired a body of studies. In this work, we also build on top of the framework of Bayesian persuasion and aim to relax the restrictive assumptions in their model.

In Bayesian persuasion, there are two players, a sender and a receiver.[1] The state of nature is randomly drawn from a distribution, with the prior known to both players. The sender has access to the realization of the state while the receiver does not. The sender can utilize the information advantage and selectively disclose information to the receiver to influence the receiver. Based on the prior information of the state and the information revealed by the sender, the receiver can take an action to maximize her own payoff, which depends on both the action and the realized state. The sender's objective also depends on the receiver's action, and the goal of the sender is to choose an information disclosure policy – which is determined before the state realization and is known to the receiver – to maximize his objective.

As an illustrative example, consider the scenario in which an online retailer (the sender) would like to persuade a buyer (the receiver) to make the purchase. The retailer's products are directly coming from the factory, and the product quality (the state of nature) is drawn from a distribution with known prior. The buyer's utility depends on both her purchase decision and the realized product quality, and the retailer's utility depends on the buyer's purchase decision. In order to persuade the buyers to purchase, the retailer can commit to perform (noisy) product inspections to reveal some information of the product quality (e.g., the inspection might signal the product quality is satisfactory with 80% chance if the quality

[1] In this paper, we use "he" to denote the sender and "she" to denote the receiver. Moreover, this work is motivated by scenarios of designing information for a population of users. Therefore, we use the term "receiver" to refer to a population of users, and sometimes we explicitly use the term "receivers".

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

of the product is indeed satisfactory and signal the product quality is unsatisfactory with 90% chance if the quality is indeed unsatisfactory). The retailer's goal is to find the optimal inspection policy to maximize the probability of selling the product to the buyer.

While Bayesian persuasion provides an elegant framework to address the above information design problem, it has made some restrictive assumptions. In particular, the receiver is assumed to be Bayesian rational, i.e., the receiver is able to form a posterior by incorporating the prior information and the signals revealed by the sender in a Bayesian manner, and then choose the action that maximizes her expected utility. However, as consistently observed in empirical studies (Axhausen and Gärling 1992; Svenson 1979; Loewenstein 1996; McFadden 2001), humans often systematically deviate from being Bayesian or being rational.

In this work, we explore the problem of information design with non-Bayesian-rational receiver. We develop an alternative framework to Bayesian persuasion that incorporates discrete choice model (McFadden 1981; Small 1987; Train 2009) and probability weighting (Wu and Gonzalez 1996; Rieger and Wang 2006; Prelec 1998) to model non-Bayesian-rational receiver. We formulate the problem of solving the optimal information disclosure policy under our model and characterize the properties of the optimal information disclosure policy. To showcase the difference of the two frameworks, we investigate the information policies derived from both frameworks in a simple baseline setting. We then conduct behavioral experiments on Amazon Mechanical Turk with 400 workers to examine the two frameworks. Our results demonstrate that our framework better aligns with the behavior with real-world humans and lead to a better information disclosure policy.

Related Work

Our work builds on top of the seminal work of Bayesian persuasion (Kamenica and Gentzkow 2011), which initiated a rich theoretical literature on communication game in which a sender can design information to persuade a receiver to take certain actions. Their work has inspired an active line of research in information design. (e.g., see the recent surveys by Kamenica 2019; Bergemann and Morris 2019). In this work, we extend this line of research on information design and focus on relaxing the assumption that the receiver is Bayesian rational through both developing an alternative framework and empirically examining human behavior.

Human models for decision making. In the problem of information design, the receiver needs to incorporate the information provided by the sender and make decisions accordingly. We can decompose this decision making process into two stages: 1) belief updating: how the receiver processes the information and updates her beliefs, and 2) decision making under uncertainty: how the receiver makes decisions with the updated belief. Since we are interested in settings in which receivers are human beings, in the following, we discuss existing human models for decision making in the above two stages.

For belief updating, Bayesian models have been the prominent model in algorithmic works (Tenenbaum 1999; Griffiths and Tenenbaum 2006; Chater, Tenenbaum, and Yuille 2006). However, it has also been consistently and widely observed in empirical studies that humans often deviate from being Bayesian (Kahneman and Tversky 1973; Tversky and Kahneman 1974; Axhausen and Gärling 1992; Svenson 1979; Loewenstein 1996; McFadden 2001). While there have been some alternative models in how humans process information to form their beliefs (Morris 1995; Sethi and Yildiz 2016; Mailath and Samuelson 2020; Wu and Gonzalez 1996; Rieger and Wang 2006; Prelec 1998), they are not widely adopted in algorithmic frameworks.

For decision-making under uncertainty, the commonly-used assumption is expected utility theory (von Neumann and Morgenstern 1944) which assumes humans take actions to maximize their expected utility. There is again a substantial body of work in behavioral economics in studying the systematic deviations of human behavior from expected utility theory. One important theory that summarizes these systematic biases is the *prospect theory* by Kahneman and Tversky (1979). Another commonly used theory, that accounts for the inherent randomness of human decision making by incorporating noises in the utility, is the discrete choice model (McFadden 1981; Small 1987; Train 2009).

In this work, to account for the receiver's deviation from being Bayesian rational, we adopt probability weighting function (Wu and Gonzalez 1996; Rieger and Wang 2006; Prelec 1998) for belief updating and discrete choice model (McFadden 1981; Small 1987; Train 2009) for decision making in our framework. We also examine whether our framework aligns with real-world human behavior through behavioral experiments. In addition, there have been some recent works that aim to incorporate human behavioral models in the computational framework. For example, Tang and Ho (2019) incorporate the herding bias during feedback generation in bandit learning. Kleinberg and Oren (2014) and Kleinberg, Oren, and Raghavan (2017) study the planning for time-consistent agents in an environment characterized by a graphical model. Our work aligns with this line of research that incorporates realistic human behavioral models in computation.

Behavioral experiments in information design. While there is a rich line of research on Bayesian persuasion, the amount of works on empirically investigating human behavior in information design is limited (Au and Li 2018; Aristidou, Coricelli, and Vostroknutov 2019; Fréchette, Lizzeri, and Perego 2019). Among these works, Au and Li (2018) incorporate reciprocity into the standard persuasion setting and conduct a laboratory experiment to validate their model on reciprocity. Aristidou, Coricelli, and Vostroknutov (2019) propose a unified framework to investigate the theoretical parallelism between information and mechanism design. Fréchette, Lizzeri, and Perego (2019) empirically examine different information design methods, including communications via cheap-talk, disclosure of verifiable information, and Bayesian persuasion. Our work departs from the above literature as we investigate the fundamental assumption of Bayesian rationality in human behavior. We create a decision-making scenario where the receiver is required to make a decision after seeing a signal realized according to some information disclosure policy to empirically measure how humans update their beliefs and make decisions.

Another closely-related work to ours is the one by de Clippel and Zhang (2019) who also relax the Bayesian assumption of receiver's behavior in persuasion. They theoretically study how receiver's mistakes in probabilistic inference impact optimal persuasion and characterize a large class of belief updating rules that the concavification method developed by Kamenica and Gentzkow (2011) can still be applied. However, their work focuses on theoretical characterization and the receiver is still assumed to be an expected utility maximizer. While in our model, we further relax this assumption by using a discrete choice model and empirically examine our models.

Theoretical Framework

In this section, we formalize the frameworks for the information design problem. We first describe the standard Bayesian persuasion framework that assumes Bayesian rational receiver. We then introduce our framework that relaxes the Bayesian rational assumption. In the later section, we compare the two frameworks on a simple baseline setting with two states and binary actions to showcase the differences of the frameworks. This simple baseline setting also motivates the design of our real-world behavioral experiments described in our experiment section.

Standard Framework: Bayesian Persuasion

We first describe the standard setting of Bayesian persuasion (Kamenica and Gentzkow 2011). In this setting, there are two players: a sender and a receiver. The goal of the sender is to design an information disclosure policy to persuade the receiver in taking actions to maximize the sender's objective.

Let the (payoff-relevant) state of the world be θ, which is drawn from a finite set Θ according to a prior distribution $\mu_0 \in \Delta(\Theta)$. The prior is common knowledge to all players. The receiver's utility is characterized by the function $u^R(a, \theta)$ which depends on the action she takes $a \in \mathcal{A}$ from a compact action set $\mathcal{A}$ and the state θ. The sender's utility is characterized by the function $u^S(a, \theta)$ that also depends on the receiver's action and the state.

Before observing the realization of the state, the sender can choose an information disclosure policy (π, Σ), which consists of a finite signal space Σ and a family of conditional distributions $\{\pi(\cdot|\theta)\}_{\theta \in \Theta}$ over $\sigma \in \Sigma$. This information disclosure policy is known to the receiver and specifies how the sender discloses information to the receiver. In particular, when a state $\theta \in \Theta$ is realized, the sender can observe the state but the receiver cannot. To influence the receiver's decision, the sender sends a signal σ, drawing from the conditional distribution $\pi(\cdot|\theta)$ specified in the information disclosure policy, to the receiver. The receiver forms her beliefs on the state of the world based on the prior and the signal provided by the sender. She then takes an action to maximize her own payoff.

In the Bayesian persuasion setting, it is assumed that the receiver is Bayesian rational, i.e., she updates her beliefs in a Bayesian manner and is an expected utility maximizer. Formally, upon seeing the signal realization σ from the sender, the receiver updates her belief, denoted by $\mu \in \Delta(\Theta)$, by applying Bayes' rule:

$$\mu(\theta|\sigma) = \frac{\pi(\sigma|\theta)\mu_0(\theta)}{\sum_{\theta' \in \Theta} \pi(\sigma|\theta')\mu_0(\theta')}. \tag{1}$$

Given the posterior belief μ, the receiver then chooses an action $a^* = a^*(\mu)$ that maximizes her expected payoff: $a^* \in \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} u^R(a, \theta)\mu(\theta)$.[2] As a key insight by Kamenica and Gentzkow (2011), the above two assumptions on the receiver's behavior allow the sender to reduce the problem of designing information disclosure policy to choosing a distribution of posterior beliefs that respects Bayes rule. Furthermore, a distribution $\tau \in \Delta(\Delta(\Theta))$ of posteriors can arise if and only if it is Bayes-plausible, i.e.,

$$\mathbb{E}_{\mu \sim \tau}[\mu] = \mu_0. \tag{2}$$

Therefore, it is without loss of generality to assume the set of available information disclosure policy to the sender is the set of Bayes-plausible distributions of posterior beliefs. By formulating the sender's *direct utility* $u^S(a, \theta)$, a function of the receiver's action, to an *indirect utility* $\widehat{u}^S(\mu)$, a function of Bayesian posteriors, the standard *concavification* argument can be applied to derive the optimal information design.

Our Framework: Persuading Non-Bayesian-Rational Receiver

In contrast to the assumptions made in Bayesian persuasion, the receiver may, in practice, exhibit systematic biases both in probabilistic inferences and in decision making. In the following discussion, we first incorporate the discrete choice model and probability weighting to model non-Bayesian-rational receiver. We then formulate the optimal information design problem under this receiver model.

Modeling non-Bayesian-rational receiver. We first relax the assumption that the receiver is an expected utility maximizer but still assume the receiver is Bayesian in updating the belief. Specifically, we leverage the discrete choice model (McFadden 2001), a commonly-used alternative of expected utility theory, to characterize the receiver's behavior when making her decision.

To provide informal intuitions, in expected utility theory, the receiver takes an action that maximizes her expected utility. When there is no ties in action utility, this action choice is deterministic. On the other hand, the discrete choice model accounts for the inherent randomness in human decision making and models the decision as a probabilistic process.

[2] In the persuasion literature, most work consider sender-preferred Subgame Perfect Equilibrium, where the receiver chooses the sender-preferred action when there are ties.

Specifically, in the discrete choice model, for each action $a \in \mathcal{A}$ the receiver can take, we add noise $\varepsilon(a)$ into the receiver's utility for taking action a. The receiver then takes an action that maximizes this noisy version of the utility. This noise captures several realistic aspects of human decision making, e.g., when there are additional inherent characteristics in the receiver's utility estimation that we cannot model, or when receiver is drawn from a population and individual differences need to be accounted for.

More formally, let $\mu \in \Delta(\Theta)$ denote the receiver's posterior induced by some signal realization. We define $\widehat{u}^R(a|\mu)$ as the noise-free expected utility for the receiver to choose action $a \in \mathcal{A}$ given the posterior belief μ, which can be written as $\widehat{u}^R(a|\mu) := \mathbb{E}_{\theta\sim\mu}\left[u^R(a,\theta)\right] = \sum_{\theta\in\Theta} u^R(a,\theta)\cdot\mu(\theta)$. In discrete choice model, the receiver takes actions based on the noisy version of the utility $\widetilde{u}^R(a|\mu)$, which can be written as

$$\widetilde{u}^R(a|\mu) := \beta \cdot \widehat{u}^R(a|\mu) + \varepsilon(a), \tag{3}$$

where $\epsilon(a)$ is the added noise and β is a parameter that tunes the relative strength of observable utility and the noises, e.g., when $\beta \to \infty$, the noise is negligible and the discrete choice model reduces to the standard expected utility theory.

Different choices of distributions of $\varepsilon(a)$ lead to different discrete choice models. In this work, we follow the commonly used Multinomial Logit (MNL) (McFadden et al. 1973) and assume that each $\varepsilon(a)$ is distributed independently, identically extreme value, where the CDF follows $F(\varepsilon(a)) = \exp(-\exp(-\varepsilon(a)))$.

Lemma 1 (McFadden et al. (1973)). *Given posterior belief μ, the probability that receiver chooses action a can then be derived as*

$$\Pr(a|\mu) = \frac{\exp\left(\beta\widehat{u}^R(a|\mu)\right)}{\sum_{a'}\exp\left(\beta\widehat{u}^R(a'|\mu)\right)}. \tag{4}$$

Proof. Define $v^R(a|\mu) = \beta\cdot\widehat{u}^R(a|\mu)$. By definition,

$$\begin{aligned}&\Pr(a|\mu)\\ =&\Pr\left(\widetilde{u}^R(a|\mu) > \widetilde{u}^R(a'|\mu),\ \forall a' \neq a\right)\\ =&\Pr\left(\varepsilon(a') < \varepsilon(a) + v^R(a|\mu) - v^R(a'|\mu),\ \forall a' \neq a\right).\end{aligned}$$

Since the ε's are independent, this cumulative distribution over all $a' \neq a$ is the product of the individual cumulative distributions:

$$\begin{aligned}&\Pr(a|\mu,\varepsilon(a))\\ =&\prod_{a'\neq a}\exp\left(-\exp(-(\varepsilon(a) + v^R(a|\mu) - v^R(a'|\mu)))\right).\end{aligned}$$

Since $\varepsilon(a)$ is not given, and so the choice probability is the integral of $\Pr(a|\mu,\varepsilon(a))$ over all values of $\varepsilon(a)$ weighted by its density

$$\begin{aligned}&\Pr(a|\mu)\\ =&\int\prod_{a'\neq a} e^{-e^{-(\varepsilon(a)+v^R(a|\mu)-v^R(a'|\mu))}} e^{-\varepsilon(a)} e^{-e^{-\varepsilon(a)}} d\varepsilon(a).\end{aligned}$$

Finally, by computing the integral over $\epsilon(a)$, we can obtain the closed-form expression (4). □

With the above lemma, we have a closed-form formulation specifying the distribution of actions the receiver will choose given her posterior belief under discrete choice model. We now relax the assumption that the receiver might not be Bayesian in updating her beliefs.

To account for non-Bayesian belief updating, we utilize the ideas of probability weighting and introduce a non-decreasing *prior-specific* probability distortion function $\omega(\cdot;\mu_0) : \Delta(\Theta) \to \Delta(\Theta)$ to capture the receiver's final belief on making her decision. This formulation helps explain the human biases in over-weighting or under-weighting the prior when performing beliefs updates. Now one can derive the following choice probabilities by incorporating the distorted posterior $\omega(\cdot;\mu_0)$ into (4):

$$\Pr(a|\omega(\mu;\mu_0)) = \frac{\exp\left(\beta\widehat{u}^R(a|\omega(\mu;\mu_0))\right)}{\sum_{a'}\exp\left(\beta\widehat{u}^R(a'|\omega(\mu;\mu_0))\right)}. \tag{5}$$

Many parametric forms of the probability weighting function have been proposed (Wu and Gonzalez 1996; Rieger and Wang 2006; Prelec 1998; Tversky and Kahneman 1992). For example, an affine probability distortion function (Gabaix 2019; Edwards 1968; Tversky and Wakker 1995) specifies a distorted posterior that falls in between a reference belief $\mu^* \in \Delta(\Theta)$ and Bayesian posterior μ: $\omega(\mu|\mu_0) = \gamma\mu^* + (1-\gamma)\mu$ where μ^* is allowed to vary with μ_0 and $\gamma \in [0,1]$ is a constant.

Optimal information design. With the modeling of the receiver, we now characterize the sender's optimal information design. To simplify the exposition, we mainly state the analysis when the receiver's behavior follows the discrete choice model defined in (4). The analysis for the model including probability weighting is similar. For notation simplicity, let $p(a|\mu) := \Pr(a|\mu)$ denote the the probability for the receiver to choose action $a \in \mathcal{A}$ when the posterior μ is induced. With this expression, we are now ready to characterize the sender's optimal information design problem:

Theorem 2. *Let μ_0 be the prior. Assume the receiver's behavior follows* (4) *when μ is the posterior. The sender's problem is equivalent to*

$$\begin{aligned}\max_{\tau\in\Delta(\Delta(\Theta))}\ & \mathbb{E}_{\mu\sim\tau}\left[\sum_{\theta\in\Theta}\mu(\theta)\sum_{a\in\mathcal{A}}p(a|\mu)u^S(a,\theta)\right]\\ \textit{s.t.}\ & \mathbb{E}_{\mu\sim\tau}[\mu] = \mu_0\end{aligned} \tag{6}$$

Proof. Let $\nu(\mu) = \{p(a|\mu)\}_{a\in\mathcal{A}} \in \Delta(\mathcal{A})$. Given a posterior μ and the corresponding $\nu(\mu)$, we can compute the sender's indirect expected utility $\widehat{u}^S(\mu)$ as a function of μ:

$$\begin{aligned}\widehat{u}^S(\mu) &= \mathbb{E}_{\theta\sim\mu}\left[\mathbb{E}_{a\sim\nu(\mu)}\left[u^S(a,\theta)\right]\right]\\ &= \sum_{\theta\in\Theta}\mu(\theta)\cdot\sum_{a\in\mathcal{A}}p(a|\mu)u^S(a,\theta).\end{aligned} \tag{7}$$

Given the prior μ_0, an information disclosure policy π generates a distribution $\tau \in \Delta(\Delta(\Theta))$ over Bayesian posteriors. It is known that, should the receiver be Bayesian, a distribution τ of posteriors is feasible iff it is Bayes-plausible (2).

Now the sender's expected utility can be written as a function of the receiver's choices and the probability measure τ, we obtain the stated reformulation of the sender's problem. □

Note that the problem (6) can be further simplified when the sender's utility is state-independent, i.e., $u^S(a,\theta) = u^S(a), \forall\theta \in \Theta$, which is a common assumption in the persuasion literature. Indeed, we have the objective $\mathbb{E}_{\mu\sim\tau}\left[\sum_{a\in\mathcal{A}} p(a|\mu)u^S(a)\right]$ in (6). By writing the sender's problem as a function of the induced Bayesian posterior, then (6) can be addressed using the tools developed by Kamenica and Gentzkow (2011). In particular, for an arbitrary real-valued function $u : \Delta(\Theta) \to [0,1]$, let u^{cc} be the concave closure of u,

$$u^{\mathsf{cc}}(\mu) = \sup\{z|(\mu,z) \in \mathsf{co}(u)\}, \tag{8}$$

where $\mathsf{co}(u)$ is the convex hull of the graph of u.

Proposition 3. *The sender's expected utility under an optimal policy is $\widehat{u}^{\mathsf{cc}}(\mu_0)$, where $\widehat{u}$ is defined in* (7).

The above analysis can also be applied to deal with settings in which the receiver distorts the probabilities through a probability weighting function. In particular, the results in Theorem 2 still hold with the only difference being that the choice probabilities in (6) will accordingly correspond to (5). We can simplify the sender's problem (6) to the following optimization problem with a distorted Bayes-plausibility constraint:

Proposition 4. *Let μ_0 be the prior, μ be the Bayesian posterior and μ^R be the receiver's non-Bayesian posterior. Assuming the receiver's behavior follows* (5) *with the probability weighting function $\omega(\cdot|\mu_0) : \Delta(\Theta) \to \Delta(\Theta)$. The sender's problem is equivalent to* [3]

$$\max_{\tau\in\Delta(\Delta(\Theta)\times\Delta(\Theta))} \mathbb{E}_{(\mu,\mu^R)\sim\tau}\left[\sum_{\theta\in\Theta}\mu(\theta)\sum_{a\in\mathcal{A}} p\left(a|\mu^R\right)u^S(a,\theta)\right]$$

$$\text{s.t. } \mathbb{E}_{\mu^R\sim\tau^R}\left[\omega^{-1}\left(\mu^R|\mu_0\right)\right] = \mu_0,$$

where $\tau^R = \int_\mu \tau(\mu,\cdot)d\mu$.

Proof. Given the receiver's belief μ^R, let $\nu(\mu^R) = \{p(a|\mu^R)\}_{a\in\mathcal{A}} \in \Delta(\mathcal{A})$. Together with the Bayesian posterior μ, we have the following sender's indirect utility

$$\begin{aligned}\widehat{u}^S(\mu,\mu^R) &= \mathbb{E}_{\theta\sim\mu}\left[\mathbb{E}_{a\sim\nu(\mu^R)}[u^S(a,\theta)]\right]\\ &= \sum_{\theta\in\Theta}\mu(\theta)\cdot\sum_{a\in\mathcal{A}} p(a|\mu^R)\mu^S(a,\theta).\end{aligned}$$

Recall that μ^R is the result of the mapping of probability weighting function $\omega(\cdot|\mu_0)$ from the Bayesian posterior μ. As the mapping $\omega(\cdot|\mu_0)$ is invertible and μ satisfies Bayes-Plausibility, μ^R must satisfy $\mathbb{E}_{\mu^R\sim\tau^R}\left[\omega^{-1}\left(\mu^R|\mu_0\right)\right] = \mu_0$. Thus, we can achieve the above sender's reformulated optimization problem. □

Similarly, when the sender's utility is state-independent, sender's problem can be further simplified as $\max_{\tau\in\Delta(\Delta(\Theta))}\mathbb{E}_{\mu^R\sim\tau^R}\left[\sum_{a\in\mathcal{A}} p(a|\mu^R)u^S(a)\right]$ with the distorted Bayesian-plausibility constraint.

A Baseline Setting with Two States and Binary Actions

In the above section, we formulate the information design problem for both the standard framework of Bayesian persuasion and our framework of persuading non-Bayesian-rational receiver. To instantiate the discussion and comparison, in this section, we consider a simple setting with two states and binary actions, a variant of the leading example in Kamenica and Gentzkow (2011), to demonstrate the differences of the two frameworks. This setting also motivates our experiment design as presented in the next section.

Consider a world with two states $\Theta = \{\mathsf{X},\mathsf{Y}\}$, where state X happens with probability $\mu_0 \in [0,1]$ and state Y happens with probability $1-\mu_0$. The receiver can choose from two actions $\mathcal{A} = \{a_{\mathsf{X}}, a_{\mathsf{Y}}\}$. The utility of the sender and the receiver both depend on the receiver's action and the realized state and have been summarized in Table 1.

Payoff	State X	State Y
Receiver chooses a_{X}	Receiver: 1. Sender: 1	Receiver: 0. Sender: 1
Receiver chooses a_{Y}	Receiver: 0. Sender: 0	Receiver: 1. Sender: 0

Table 1: Payoff structure.

In this payoff structure, the receiver aims to select the action that matches the state (i.e., select action $a_{\mathsf{X}}/a_{\mathsf{Y}}$ for state X/Y), while the sender wishes to persuade the receiver to select action a_{X}.

Optimal information design with Bayesian-rational receiver. In the following discussion, we use μ to denote the posterior probability of state X. If the receiver is Bayesian rational, whenever the receiver sees a signal that induces a posterior $\mu \geq 0.5$, the receiver's best response is to choose action a_{X}. In other words, the receiver's response is a simple step function in posterior beliefs (the receiver chooses action a_{X} when $\mu \geq 0.5$ and action a_{Y} when $\mu < 0.5$). Given the receiver's behavior, the optimal information disclosure policy can be achieved with only 2 signals, represented using $\{\mathsf{R},\mathsf{B}\}$, and the policy can be specified as below. [4]

Proposition 5 (Optimal policy assuming Bayesian rational receiver (Kamenica and Gentzkow 2011))**.** *When the prior $\mu_0 < 0.5$, an optimal information disclosure policy exists and satisfies:*

- *when state X is realized, always sends signal R;*
- *when state Y is realized, with prob. $\frac{\mu_0}{1-\mu_0}$ sends signal R, and with prob. $1-\frac{\mu_0}{1-\mu_0}$ sends signal B.*

When $\mu_0 \geq 0.5$, an uninformative information disclosure policy is the optimal policy.

[3] Including probability weighting in our model is essentially the same as distorting updated beliefs proposed by de Clippel and Zhang (2019). We can show that a distorted version of Bayes-plausibility holds, and therefore the standard concavification technique to derive optimal information design can be applied.

[4] We choose $\{\mathsf{R},\mathsf{B}\}$ as signal notations mainly for the consistency of our experiment presentation in our experiment section.

Below is the intuition of the optimal policy. When $\mu_0 \geq 0.5$, when deploying an uninformative information policy, the receiver's posterior is the same as prior, and she will always choose action a_X, and therefore an uninformative information policy is the optimal policy. When $\mu_0 < 0.5$, recall that the goal of the sender is to persuade the receiver to choose a_X when the prior of state Y is larger than half. In the optimal information policy, when the state is X, the sender wants to reveal the true information to encourage the receiver to choose a_X. When the state is Y, the sender wants to make the receiver have indifferent beliefs between the state to maximize the chance the receiver chooses a_X. The above policy generates two possible posteriors: $\mu = 0.5$ with probability $2\mu_0$ on seeing signal R, and $\mu = 0$ with probability $1 - 2\mu_0$ on seeing signal B.

Optimal information design with non-Bayesian-rational receiver. When the receiver is not Bayesian rational, the receiver's probability of choose a_X is not a step-function of posterior μ as in Proposition 5. Instead, as described in our framework, it is a smoothed continuous function as below:

$$p(a = a_\mathsf{X}|\mu) = \frac{\exp(\beta\mu)}{\exp(\beta(1-\mu)) + \exp(\beta\mu)}. \qquad (9)$$

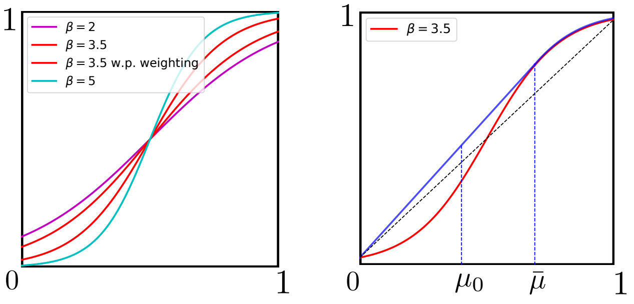
We can also derive the sender's optimal information design when the receiver follows the model (9). In particular, since the sender obtains zero utility when the receiver chooses action a_Y, the sender's indirect utility $\widehat{u}^S(\mu)$ as a function of posterior μ is simply $\widehat{u}^S(\mu) = p(a = a_\mathsf{X}|\mu)$. A concavification argument allows us to characterize the following optimal information disclosure policy:

Proposition 6 (Optimal policy assuming non-Bayesian-rational receiver)**.** *Let $p(\mu) := p(a = a_\mathsf{X}|\mu)$ and let $\bar{\mu}$ be the unique solution of $\bar{\mu}p'(\bar{\mu}) = p(\bar{\mu}) - p(0)$. Given prior $\mu_0 \leq \bar{\mu}$, an optimal information disclosure policy exists and satisfies*

- *when state* X *is realized, always sends signal* R*;*
- *when state* Y *is realized, with prob. $\frac{\mu_0(1-\bar{\mu})}{(1-\mu_0)\bar{\mu}}$ sends signal* R*, and with other prob. sends signal* B*.*

When $\mu_0 > \bar{\mu}$, an uninformative information disclosure policy is the optimal policy.

The optimal information policy shares a similar structure as the one when the receiver is Bayesian rational (Proposition 5). However, the threshold $\bar{\mu}$, that characterizes when an uninformative policy is not optimal, and the probability for sending signal B when the realized state is Y are different and are influenced by the receiver model. Furthermore, as β in the receiver model (9) increases, the shape of $p(\mu)$, the probability for the receiver to choose action a_X given posterior μ, is more towards a step function with breaking point at 0.5 and thus $\bar{\mu}$ is smaller. Intuitively, larger β implies that the impact of unobserved component $\varepsilon(a)$ is smaller on the receiver's utility, and thus the receiver is more towards an expected utility maximizer. The above discussion is graphically illustrated in Figure 1. The analysis when including the probability weighting is similar.



(a) (b)

Figure 1: Left: Various shapes of $\widehat{u}^S(\mu)$ (or $p(\mu)$) and $\widehat{u}^S(\omega(\mu))$ (or $p(\omega(\mu))$) with an affine distorting function ω where $\gamma = 0.3, \mu^* = 0.5$. Right: Red line is the concavification $\widehat{u}^{\mathsf{cc}}(\mu)$ for $\widehat{u}^S(\mu)$.

Real-World Experiment

Our discussion in the previous sections demonstrates the different predictions on the receiver's behavior and the optimal information disclosure policy when we consider different receiver models. In this section, we describe the setup and results of our real-world behavioral experiments to examine these predictions. The experiment has been approved by IRB at Washington University.

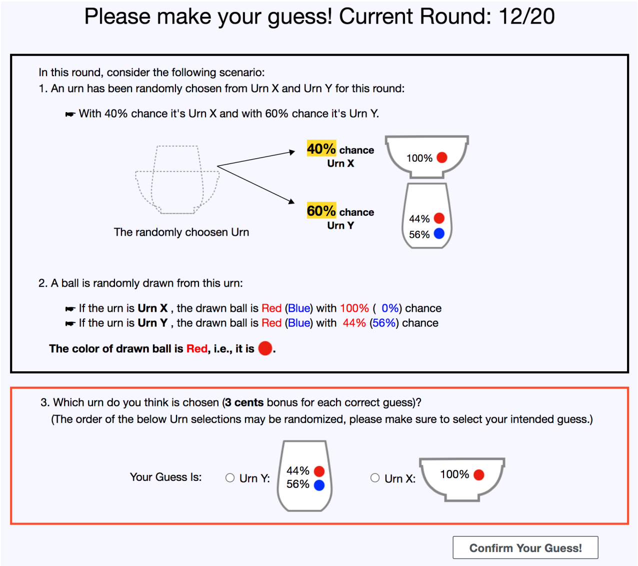
In our experiment, we recruit online workers to answer a series of questions. In each question, workers are asked to perform a probabilistic-inference and decision-making task. We design the questions in a way that we can control the prior and the information structure and then observe workers' corresponding actions. Moreover, given a prior and a realized signal from the information policy, we are able to derive the corresponding induced Bayesian posterior (calculated using Bayes rule). We are interested in examining the following two questions:

- **Q1:** Are workers Bayesian?
  To examine whether workers are Bayesian, we can design two scenarios that lead to the same induced posterior but have different priors and information policies. If workers are Bayesian, their decisions should depend only on the posterior, and we should observe the same worker behavior on the two scenarios.
- **Q2:** Are workers rational?
  To examine whether workers are rational, we can create scenarios that lead to different posteriors. If workers are rational, we should observe workers' behavior follows a step function over the induced posteriors.

Experiment Setup

We recruited 400 unique workers from MTurk, where each worker is required to complete 20 questions. We offer a \$0.5 base payment, and each worker may also receive a bonus payment of up to \$0.6 (the bonus rule will be explained shortly). The bonus amount is chosen to be large enough so workers are motivated to perform well. The average hourly rate is around \$12.15.

Task. Our goal is to evaluate the receiver's behavior. Therefore, we play the role of the sender and have all recruited workers play as the receiver. Each worker needs to complete 20 questions as described below.



Please make your guess! Current Round: 12/20

In this round, consider the following scenario:
1. An urn has been randomly chosen from Urn X and Urn Y for this round:

☛ With 40% chance it's Urn X and with 60% chance it's Urn Y.

2. A ball is randomly drawn from this urn:

☛ If the urn is **Urn X** , the drawn ball is Red (Blue) with 100% (0%) chance
☛ If the urn is **Urn Y** , the drawn ball is Red (Blue) with 44% (56%) chance

The color of drawn ball is Red, i.e., it is ●.

3. Which urn do you think is chosen (**3 cents** bonus for each correct guess)?
(The order of the below Urn selections may be randomized, please make sure to select your intended guess.)

Your Guess Is: ○ Urn Y: 44% ● 56% ● ○ Urn X: 100% ●

Confirm Your Guess!

Figure 2: The task interface.

In each question, as shown in Figure 2, workers are informed that there are two urns, Urn X and Urn Y. At the beginning of the question, an urn is randomly drawn according to the prior distribution that is known to the workers. Each urn contains certain fraction of red balls and blue balls. The ball composition of each urn is also shown to workers. After an urn is realized, we choose a ball uniformly at random from this realized urn. The color of the drawn ball is then disclosed to the worker. Upon seeing the color, the worker is required to make guess on which urn is realized.

This experiment setup is designed to capture human decision-making process. The two urns represent the world state. The ball composition is the information disclosure policy. When seeing the realized ball, the workers update their prior beliefs (the prior of urn drawing) with additional information (realized ball drawn according to the commonly known ball compositions in urns) and make decisions (guessing which is the realized urn).

Bonus rule. For each correct guess (i.e., worker's guess matches the realized urn), worker receives a bonus of $\$0.03$, thus each subject will receive at most $\$0.6$ in the game. The bonus for correct guess on Urn X and Urn Y is the same to match the setting in section about our baseline setting.

Treatment design. To answer our research questions, we conducted a randomized behavioral experiment. The experiment consists of two treatments, which differ in the prior distribution of the state. In the *high prior* treatment, we fixed the prior to be $(0.4, 0.6)$, while in the *low prior* treatment, the prior is fixed as $(0.2, 0.8)$. We then design eight ball compositions in urns (corresponding to information disclosure policies) such that, conditional on the realization of a red ball draw, the Bayesian posterior would be $(0.2, 0.3, \ldots, 0.9)$ for both treatments. The detailed setup of our ball composition is included in Table 2. For each arriving worker, she is randomly assigned to one of the treatments and needs to answer 20 questions. Each question corresponds to a ball composition. Each ball composition is repeated 2 to 3 times and the order of the question and the options are all randomized to alleviate any potential position bias.

This treatment design enables us to answer both research questions Q1 and Q2. Since we control the ball compositions so that both treatments lead to the same set of Bayesian posteriors (conditional on red ball draw), by comparing the worker behavior between the two treatments, we can answer Q1. Since the prior is fixed in each treatment, by examining the behavior with different induced posterior in the same treatment, we can answer Q2.

Experiment Results

Among the 400 recruited workers, 199 workers were randomly assigned to the high prior $(0.4, 0.6)$ treatment and 201 workers were randomly assigned to the low prior $(0.2, 0.8)$ treatment. For the self-reported population demographic for the participants, there are 41.5% female, 71.25% under 40 years old, and over 90% of the participants reported to have at least college degrees.

Receiver's behavior. We first report the receiver's behavior on both treatments. Note that if workers are Bayesian rational, we should expect to see workers taking the same actions for any fixed posterior no matter which treatment they are in. In addition, workers' behavior should follows a step function within each treatment, with workers choosing urn X when the posterior is larger than 0.5 and choosing urn Y otherwise.

The results, as shown in Figure 3, show that worker behavior has significantly deviated from the model of Bayesian rationality. In particular, the differences between the two treatments demonstrate that workers are not updating their beliefs in a Bayesian manner. The sigmoid-shape curve in workers' behavior demonstrates that worker behavior aligns better with the discrete choice model instead of the expected utility theory (which leads to a step function).

Fitting receiver behavior to our framework. Next we examine how well our framework explains the empirical worker behavior by fitting the empirical observations to our model as described in Equation (5). For the probability weighting function ω, we choose a simple but an intuitive affine probability weighting function. In addition, since the data quality by online workers have known to be inconsistent (Jagabathula, Subramanian, and Venkataraman 2017; Ipeirotis, Provost, and Wang 2010), when fitting the data to models, we consider the case that there is a $(1-\alpha)$ fraction of workers who might always be random guessing (choosing urn X with 0.5 chance).

ball composition	prior $(0.2, 0.8)$	prior $(0.4, 0.6)$
posterior $(0.2, 0.8)$	$(100\%, 0\%, 100\%, 0\%)$	$(37\%, 63\%, 100\%, 0\%)$
posterior $(0.3, 0.7)$	$(100\%, 0\%, 58\%, 42\%)$	$(64\%, 36\%, 100\%, 0\%)$
posterior $(0.4, 0.6)$	$(100\%, 0\%, 37\%, 63\%)$	$(100\%, 0\%, 100\%, 0\%)$
posterior $(0.5, 0.5)$	$(100\%, 0\%, 25\%, 75\%)$	$(100\%, 0\%, 67\%, 33\%)$
posterior $(0.6, 0.4)$	$(100\%, 0\%, 17\%, 83\%)$	$(100\%, 0\%, 44\%, 56\%)$
posterior $(0.7, 0.3)$	$(100\%, 0\%, 11\%, 89\%)$	$(100\%, 0\%, 29\%, 71\%)$
posterior $(0.8, 0.2)$	$(100\%, 0\%, 6\%, 94\%)$	$(100\%, 0\%, 17\%, 83\%)$
posterior $(0.9, 0.1)$	$(100\%, 0\%, 3\%, 97\%)$	$(100\%, 0\%, 7\%, 93\%)$

Table 2: Ball compositions for different prior and different posterior on seeing red ball. In each cell, the first two numbers correspond to the fraction of red balls and blue balls in Urn X, and the last two numbers correspond to the fraction of red balls and blue balls in Urn Y.

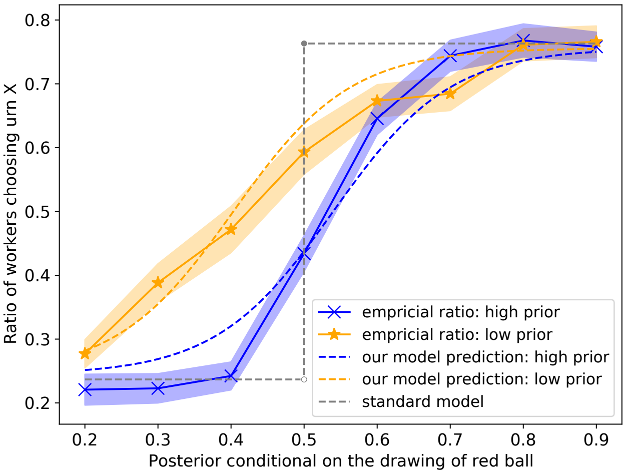


Figure 3: The solid lines represent the percentage of workers that choose Urn X conditional on a red ball realization. Shaded regions correspond to the regions of plus/minus one standard error. Dashed lines correspond to fitted models in our framework.

The fitted curves are also included in Figure 3. Compared with the step function as predicted with the Bayesian rationality assumption, our model aligns better with real-world human behavior.

Details on the model evaluation. Recall that our model, after including an affine probability distorting function $\omega(\cdot|\mu_0)$ and α-fraction random workers, is defined as follows

$$p(a = a_{\mathsf{X}}|\omega(\mu|\mu_0)) \\ = \frac{\alpha \exp(\beta\omega(\mu|\mu_0))}{\exp(\beta(1-\omega(\mu|\mu_0))) + \exp(\beta\omega(\mu|\mu_0))} + (1-\alpha)0.5,$$

where $\omega(\mu|\mu_0) = \gamma\mu^* + (1-\gamma)\mu$, $\alpha, \gamma \in [0, 1], \beta > 0$ and μ^* is a reference belief that may depend on the prior information. Using non-linear least squares, we jointly optimize the parameters of function $p(a = a_{\mathsf{X}}|\omega(\mu|\mu_0))$, while ensuring parameters (α, β, γ) to be the same for both treatments and allowing $\mu^* \in [0, 1]$ to vary with the prior, to be fitted to the data of both treatments. To assure for fair comparisons, we also include the prediction if we assume α fraction of workers are Bayesian rational (see gray dashed line in Figure 3). Recall that in Proposition 5, the response of Bayesian rational workers is a step function $p(a = a_{\mathsf{X}}|\mu) = \mathbf{1}\{\mu \geq 0.5\}$. Thus, with $(1-\alpha)$ fraction random workers, the prediction should be characterized by $p(a = a_{\mathsf{X}}|\mu) = \alpha\mathbf{1}\{\mu \geq 0.5\} + (1-\alpha)0.5$.

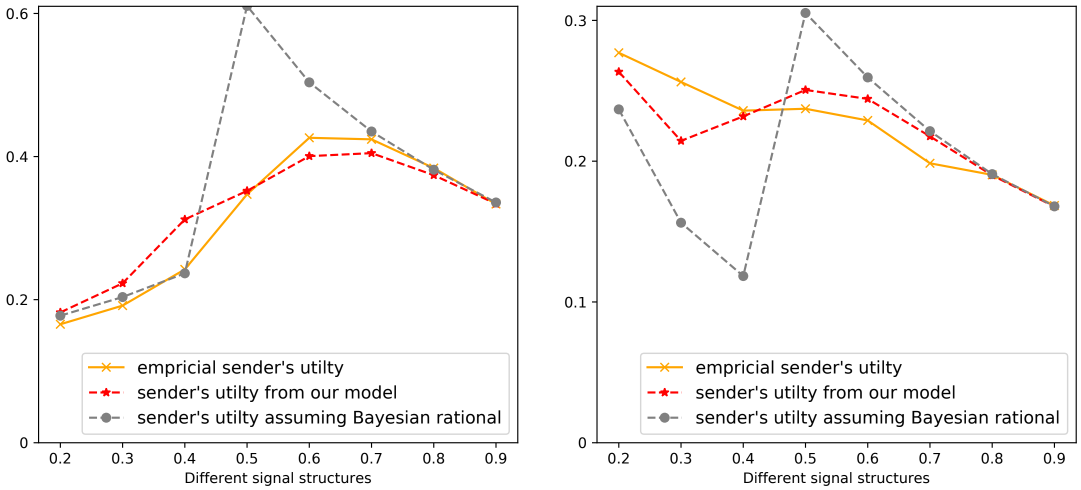
To evaluate how well each model fits the data, we use 5-fold cross-validation to estimate the out-of-sample prediction error of the model. In particular, we split the available data randomly into 5 equally-sized disjoint subsets. In each iteration, we choose one subset as the test data and the remaining subsets as training data to find out the model parameters. The out-of-sample performance is then evaluated on the chosen test data. After iterating all subsets, we compute the average out-of-sample error across 5 test sets.

The evaluation errors, computed via the sum of squared residuals, together with the errors if we assume workers are Bayesian rational, are shown in Table 3. The results demonstrate that our framework explains the real human behavior better than Bayesian persuasion does.

	error using our model	error assuming Bayesian rational
prior $(0.2, 0.8)$	0.0506	0.1230
prior $(0.4, 0.6)$	0.0417	0.1231

Table 3: 5-fold cross validation error (computed via the sum of squared residuals) for the models in Figure 3.

Implication to information design. Finally, we discuss the impacts of receiver models to the information design problem. In particular, note that each of the ball composition of urns corresponds to an information disclosure policy. For each policy, we can compute the expected utility for each receiver model by assuming the receiver takes action follows the model prediction. In addition, given the data collected



(a) High prior.

(b) Low prior.

Figure 4: Comparisons between the empirical sender's utility collected in data, sender's utility predicted by our model, and the sender's utility predicted by assuming workers are Bayesian rational.

by our experiments, we can compute the empirical average utility achieved by each policy (i.e., multiply the empirical ratio of workers choosing Urn X by the probability of red ball realization) and use it as the ground truth for comparison.

The results, as shown in Figure 4, demonstrate that our model (fitted with data) makes a much more accurate prediction (red dashed line) on the empirical average utility (orange line) of different information disclosure policies than the one predicted by Bayesian persuasion (gray dashed line). For example, for the high prior treatment, different from the peak at $(0.5, 0.5)$ when assuming workers are Bayesian rational, the empirical data shows that the empirical optimal information disclosure policy is generating a posterior between $(0.6, 0.4)$ and $(0.7, 0.3)$ when seeing a red ball, and this is also reflected in our model prediction. Similar results can also be found for the low prior treatment.

Discussions and Future Work

In this section, we discuss the limitations of our current results and potential future directions.

Generalizability of our framework and experimental results. While our work has been one of the few empirical studies in examining human behavior in the persuasion literature, similar to prior work, our experiment is constructed on a more abstract setup (i.e., utilizing the urn and ball drawing problem). Developing a more realistic experimental setup that depicts real-world scenarios (e.g., how a seller selectively discloses product information to persuade the buyer to make the purchase decision) and/or conducting more extensive experiments (e.g., including more priors and posteriors, recruiting more workers) would help better understand and model real human behavior.

In addition, our current experiments has limited to a simple form of information presentation. It is therefore not trivial to claim that our findings hold for different presentations of information structure. In particular, in our experiment design, for each combination of prior and target posterior, we identify an information disclosure policy (i.e., a particular set of ball compositions in urns) that induces the target posterior from the prior when receiver sees a red ball. In our design, almost all ball compositions have 100% red balls in Urn X except two compositions in the upper right of Table 2. There are several benefits for this style of composition. First, it aligns with the optimal information design as derived in Proposition 5, i.e., the sender always sends a signal R when Urn X is realized. For the two compositions that it is not feasible to have 100% red balls in Urn X, we choose to make Urn Y to contain 100% red balls to ensure that our ball compositions are consistent among different tasks, i.e., at least one urn has 100% red balls. Second, we believe the simplicity of these compositions also helps to alleviate human's cognitive burden when processing signal information. However, despite the above mentioned benefits, it limits the generalizability of our findings outside of this particular form of information presentation. Note that there are essentially infinite number of different ball compositions that we can use to induce the same target posterior. For example, in Figure 2, given the prior $(0.4, 0.6)$, any ball composition $(x, 1-x, y, 1-y)$ that satisfies $\frac{0.4x}{0.4x+0.6y} = 0.6, x, y \in [0, 1]$ can induce a posterior $(0.6, 0.4)$ whenever a worker sees a red ball. Understanding the impacts of different signal presentations has practical importance and would be an important future research direction.

We have considered a particular set of behavioral models, i.e., discrete choice model and probability weighting, to relax the Bayesian rational assumption. While these models have been well-examined in the literature, there have also been other models of human decision making to relax the assumption of Bayesian rationality. Empirically understanding whether and when other models are suitable and how different models impact the information design problem requires more future studies from both theoretical and experimental investigations.

Algorithmic solutions for information design. In Section , we develop an alternative framework to model the receiver's behavior and formulate the sender's optimization problem. In the section about our baseline setting, we then demonstrate that in a simple baseline setup with two states and binary actions, we can obtain a closed form of optimal information structure. The natural next question to ask is that whether we can develop an algorithmic procedure to obtain the optimal information design for general settings in these frameworks.

Note that if the receiver is Bayesian rational, for a general information design problem, there have been earlier works (Dughmi and Xu 2019) showing that it is #P-hard to exactly compute the expected sender utility for the optimal information structure. One interesting future direction is to explore whether the earlier computational complexity results still hold in our framework. More specifically, can we identify a polynomial-time algorithm to derive the optimal information disclosure policy, as defined in (6).

Potential negative societal impacts. Lastly, we would like to highlight the potential negative societal impacts of the usage of information design. When the sender's objective is to maximize the social welfare or to improve the quality of the receiver's action, the impacts of information design could be positive to the receiver and beneficial to the society. However, in our work and in almost the entire literature on Bayesian persuasion, we have often focused on how to identify an optimal information disclosure policy that maximizes the sender's payoff. Since the sender often represents the advantageous party (e.g., the government, the company, the platform, etc) that has access to more information, when the interests of the sender do not align with the interests of the receiver, optimizing the sender's utility could lead to potential negative social impacts to the receivers, who are often the general public. In other words, with ill-specified objective in information design, the sender could utilize the information advantage and create significant negative impacts. It is therefore also important to consider the impacts and the potential regulations on information design.

Conclusion

This paper investigates the information design problem in human-in-the-loop systems. We extend the standard framework of Bayesian persuasion and relax the common Bayesian rational assumption on the receiver's behavior. In particular, we develop an alternative framework based on discrete choice model and probability weighting to account for this relaxation. We characterize the property of this framework and show the optimal information design can also be formulated as a constrained optimization problem. To demonstrate the difference of the two frameworks, we utilize a simple baseline setting with two states and binary actions and provide the optimal information disclosure policies derived from both frameworks. Finally, we conduct behavioral experiments on Amazon Mechanical Turk to examine the behavioral models and the policies derived by both frameworks. The results demonstrate that our proposed framework better explains user behavior and leads to better information design policies. Our results showcase the importance of understanding human behavior before applying the results on information design research to the field.

Acknowledgments

We thank the anonymous reviewers for the helpful comments. This work is supported in part by the Office of Naval Research Grant N00014-20-1-2240.

References

Aristidou, A.; Coricelli, G.; and Vostroknutov, A. 2019. Incentives or Persuasion? An Experimental Investigation. *GSBE Research Memoranda* (012).

Au, P. H.; and Li, K. K. 2018. Bayesian persuasion and reciprocity: theory and experiment. *Available at SSRN 3191203* .

Axhausen, K. W.; and Gärling, T. 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews* 12(4): 323–341.

Bergemann, D.; and Morris, S. 2019. Information design: A unified perspective. *Journal of Economic Literature* 57(1): 44–95.

Chater, N.; Tenenbaum, J. B.; and Yuille, A. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences* 10(7): 287–291.

de Clippel, G.; and Zhang, X. 2019. Non-bayesian persuasion. Technical report, Technical report, Working Paper.

Dughmi, S.; and Xu, H. 2019. Algorithmic bayesian persuasion. *SIAM Journal on Computing* (0): STOC16–68.

Edwards, W. 1968. Conservatism in human information processing. *Formal representation of human judgment* .

Emek, Y.; Feldman, M.; Gamzu, I.; PaesLeme, R.; and Tennenholtz, M. 2014. Signaling schemes for revenue maximization. *ACM Transactions on Economics and Computation (TEAC)* 2(2): 1–19.

Fréchette, G. R.; Lizzeri, A.; and Perego, J. 2019. Rules and commitment in communication: An experimental analysis. Technical report, National Bureau of Economic Research.

Gabaix, X. 2019. Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, 261–343. Elsevier.

Gehlbach, S.; and Sonin, K. 2014. Government control of the media. *Journal of public Economics* 118: 163–171.

Goldstein, I.; and Leitner, Y. 2018. Stress tests and information disclosure. *Journal of Economic Theory* 177: 34–69.

Griffiths, T. L.; and Tenenbaum, J. B. 2006. Optimal Predictions in Everyday Cognition. *Psychological Science* 17(9): 767–773.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67.

Jagabathula, S.; Subramanian, L.; and Venkataraman, A. 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *The Journal of Machine Learning Research* 18(1): 3233–3299.

Kahneman, D.; and Tversky, A. 1973. On the psychology of prediction. *Psychological Review* 80: 237–251.

Kahneman, D.; and Tversky, A. 1979. Prospect theory: An analysis of decisions under risk. *Econometrica* 263–291.

Kamenica, E. 2019. Bayesian persuasion and information design. *Annual Review of Economics* 11: 249–272.

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

Kleinberg, J.; and Oren, S. 2014. Time-inconsistent planning: a computational problem in behavioral economics. In *Proceedings of the fifteenth ACM conference on Economics and computation*, 547–564.

Kleinberg, J.; Oren, S.; and Raghavan, M. 2017. Planning with multiple biases. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, 567–584.

Loewenstein, G. 1996. Out of control: Visceral influences on behavior. *Organizational behavior and human decision processes* 65(3): 272–292.

Mailath, G. J.; and Samuelson, L. 2020. Learning under Diverse World Views: Model-Based Inference. *American Economic Review* 110(5): 1464–1501. doi:10.1257/aer.20190080. URL https://www.aeaweb.org/articles?id=10.1257/aer.20190080.

McCloskey, D.; and Klamer, A. 1995. One quarter of GDP is persuasion. *The American Economic Review* 85(2): 191–195.

McFadden, D. 1981. Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications* 198272.

McFadden, D. 2001. Economic choices. *American economic review* 91(3): 351–378.

McFadden, D.; et al. 1973. Conditional logit analysis of qualitative choice behavior .

Morris, S. 1995. The Common Prior Assumption in Economic Theory. *Economics and Philosophy* 11(2): 227–253. doi:10.1017/S0266267100003382.

Prelec, D. 1998. The probability weighting function. *Econometrica* 497–527.

Rayo, L.; and Segal, I. 2010. Optimal information disclosure. *Journal of political Economy* 118(5): 949–987.

Rieger, M. O.; and Wang, M. 2006. Cumulative prospect theory and the St. Petersburg paradox. *Economic Theory* 28(3): 665–679.

Sethi, R.; and Yildiz, M. 2016. Communication With Unknown Perspectives. *Econometrica* 84(6): 2029–2069.

Small, K. A. 1987. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society* 409–424.

Svenson, O. 1979. Process descriptions of decision making. *Organizational behavior and human performance* 23(1): 86–112.

Tang, W.; and Ho, C.-J. 2019. Bandit Learning with Biased Human Feedback. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1324–1332.

Tenenbaum, J. 1999. Bayesian Modeling of Human Concept Learning. In *Advances in Neural Information Processing Systems*. MIT Press.

Train, K. E. 2009. *Discrete choice methods with simulation*. Cambridge university press.

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185(4157): 1124–1131. doi:10.1126/science.185.4157.1124.

Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty* 5(4): 297–323.

Tversky, A.; and Wakker, P. 1995. Risk attitudes and decision weights. *Econometrica: Journal of the Econometric Society* 1255–1280.

von Neumann, J.; and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Wu, G.; and Gonzalez, R. 1996. Curvature of the probability weighting function. *Management science* 42(12): 1676–1690.