Bayesian Persuasion: A Game-Theoretic Overview

Sam Boardman Alan Gerdov Akshay Tondak

December 14, 2021

1 Introduction

The notion of persuasion is ubiquitous in real-world networks, with examples in consumer markets, the justice system, and political institutions. The way persuasion is carried out varies across these examples, but similarities can be abstracted by considering the knowledge, beliefs, and incentives of each agent—and the relationship between them—in such networks.

From this view, one agent, called the *sender*, attempts to persuade one or more other agents about the state of the world, called the *receivers*, to vote for a certain action. The sender does this by sending the receivers signals, which are a depiction of the state of the world, not necessarily corresponding to the true state of the world.

This framework becomes interesting in the case where there are some states of the world in which taking the recommended action is beneficial for the receivers, and others in which it is not, yet it is always beneficial for the sender. This creates a situation in which the interests of the sender and the receivers are only sometimes aligned, and so the signal from the sender cannot be blindly trusted by the receivers as an accurate representation of the state of the world. However, the chosen signal can still convey information about the likely state of the world, if the two coincide often enough. Both the sender (before observing the state of the world) and the receivers have a common prior belief about the state of the world, which is encoded as a probability distribution over states of the world. After observing the signals, the receivers update their belief via Bayes' rule, a mathematical formula that combines the information from the prior belief and the observed signal (such receivers are therefore called *Bayesian*).

The receivers choose their action based on these *posterior* (updated) beliefs to maximize their expected payoff over states of the world. Knowing this, the sender chooses a probability distribution over signals for each state of the world to maximize her expected payoff over states of the world before observing the state of the world. The foregoing set-up characterizes *Bayesian persuasion*.

Numerous questions arise when thinking about the nature, implementation, and outcomes of Bayesian persuasion settings. What are the stochastic dynamics of signals causing receivers to update their beliefs? How precisely do these posterior beliefs predict the true state of the world? How do the receivers aggregate their posterior beliefs into votes? Can the sender choose signals in a way that causes greater expected payoff than if the receivers voted based on their prior belief, i.e. without persuasion? How can the details of a Bayesian persuasion setting be chosen to enhance the outcome vis-à-vis these factors, bearing in mind that some may be mutually opposed? In order to answer these questions precisely, we formalize Bayesian persuasion using the notions and methodology of game theory.

2 Definitions, the Game, and the Solution Concept

Definition 1 (Conditional probability)**.** *Let P be a probability measure and let C and D be events with $P(D) \neq 0$. We say that the conditional probability of C given D is*

$$P(C|D) = \frac{P(C \cap D)}{P(D)}.$$

The idea is that, after observing an event D, one updates their probability of event C from $P(C)$ to $P(C|D)$ by only looking at the fraction of the time event C occurs when event D occurs. This creates a new probability measure $P(\cdot|D)$ by using the information implied by event D occurring. We can now state the following:

Theorem 2.1 (Bayes' rule)**.** *Let P be a probability measure and let C and D be events with $P(C), P(D) \neq 0$. Then,*

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)}.$$

This statement follows directly from substitution using the definition of $P(C|D)$ and $P(D|C)$. It provides a simple computational technique for calculating one conditional probability from another. In analogy with the above idea of updating, we say one engages in *Bayesian updating* and is *Bayesian* if they replace their prior probability measure $P(\cdot)$ with $P(\cdot|D)$ upon observing event D. This notion will allow receivers to update their beliefs about the state of the world upon viewing the signal sent by the sender. We return to this idea after delineating our game of study:

We model Bayesian persuasion as a dynamic game of incomplete information, in the next two sections referring to the work of Wang (2013):

- $N = \{\text{sender}, 1, \ldots, n\}$, the set of players;
- $t \in T = \{\alpha, \beta\}$, the state of the world;
- $p : T \to [0, 1]$, the common prior over T;
- $s \in S = \{a, b\}$, a signal;
- $A_{\text{sender}} = \{\{\pi(\cdot|t)\}_{t\in T} | \pi : T \to \Delta(S)\}$, a family of conditional distributions from which the signals s are drawn;
- $v_i : S \to \{A, B\}$, the vote of receiver i;
- $v(v_1 \ldots, v_n) \mapsto \{A, B\}$, a symmetric voting rule; and
- $u_{\text{sender}} : \{A, B\} \times T \to \mathbb{R}$, $u_i : \{A, B\} \times T \to \mathbb{R}$, the utility functions of the players.

We further stipulate that

$$u_{\text{sender}}(B, t) > u_{\text{sender}}(A, t),\ \forall t \in T$$

whereas

$$\forall 1 \leq i \leq n,\ u_i(B, \beta) > u_i(A, \beta);\ u_i(A, \alpha) > u_i(B, \alpha).$$

A few comments are in order with respect to the assumptions of the model and their purpose. First, the state of the world is cast as binary, as is the signal that can be conveyed to the receivers and the votes they ultimately cast. The intuition behind this is that the receivers would like to vote for one option, B, in one state of the world, β, and the other option, A, in the other state of the world, α. On the other hand, the sender would like the receivers to vote for $v = B$ regardless of which state of the world is realized. These preferences are reflected in the players' respective utility functions. The signal given to the receivers is drawn from $\pi(\cdot|\alpha)$ if the state of the world is α and $\pi(\cdot|\beta)$ if the state of the world is β; for example, given α is the state of the world, signal a occurs with probability $\pi(a|\alpha)$. Generally, we will think of signal a as suggesting that α is the state of the world: $\pi(a|\alpha) > \pi(a|\beta)$, which forces $\pi(b|\alpha) < \pi(b|\beta)$, and so we correspondingly think of signal b as suggesting that β is the state of the world. In principle, both of these could be equalities, but then the signals are chosen the same way in each state, and therefore do not reveal any information about the probable state of the world; this can be made precise using Bayes' rule. Also note that these particular associations are based on some notion of similarity between states of the world and signals, although reversing their interpretations would not affect our analysis of the game; Bayes' rule only considers the probability associated with each event.

We now present an explicit timeline for the game. First, the sender chooses a pair of conditional distributions for each state of the world from A_{sender}. Then, a phantom player, conceptualized as *Nature*, determines the state of the world by drawing from the common prior. A signal is drawn from the conditional distribution chosen by the sender for the present state of the world. The receivers observe the signal then cast their votes, with these votes and the present state of the world determining the utility obtained by the sender and the receivers.

A key observation is that unlike static games, this is a dynamic game where the sender has to choose their strategy first, followed by the receivers choosing their strategy second. Therefore, we consider a solution concept that qualifies Nash equilibrium: subgame perfect equilibrium. We state it below in the context of our Bayesian persuasion game and then review it:

Definition 2 (Subgame Perfect Equilibrium)**.** *We say that a strategy profile $(\pi^*(\cdot|\alpha), \pi^*(\cdot|\beta), \sigma_1^*, \ldots, \sigma_n^*)$ is a subgame perfect equilibrium of the above Bayesian persuasion game if:*

(1) $\forall t \in T$, $(\pi^*(\cdot|\alpha), \pi^*(\cdot|\beta))$ *maximizes* $E[u_{sender}(v,t)]$, *where* $v \equiv v(\sigma_1^*, \ldots, \sigma_n^*)$ *is stochastically determined by the signal drawn from the distributions* $(\pi^*(\cdot|\alpha), \pi^*(\cdot|\beta))$*; i.e.*

(2) $\forall 1 \leq i \leq n$, $\forall s \in S$, $\sigma_i^*(\pi^*(\cdot|t), s)$ *maximizes* $E[u_i(v,t)]$, *where the state of the world* t *is drawn from the Bayesian receivers' posterior beliefs* $\mu_i^*(\cdot|s)$ *upon viewing signal* s*:* $\mu_i^*(t|s) = \frac{\pi^*(s|t)p(t)}{\sum_{t' \in T} \pi^*(s|t')p(t')}$.

In this definition, (2) states that the receivers know the conditional distributions $(\pi^*(\cdot|\alpha), \pi^*(\cdot|\beta))$ the sender has selected *before* voting, and vote in order to maximize their expected utility by using this information, the signal drawn therefrom, and their prior beliefs, with each a separate component of Bayes' rule; in the denominator of the statement, $\pi^*(s)$ is written using the law of total probability so that these data may be substituted. Therefore, (1) implies that the sender must reveal their chosen conditional distributions before the receivers vote, and maximizes her expected utility based on how the receivers vote in (2). The idea is that, once the sender chooses her conditional distributions, a subgame is induced where the receivers have to maximize their expected utility *given* her choice, requiring the receivers to play a Nash equilibrium of this one-player game. The sender also maximizing her expected utility gives rise to a special type of Nash equilibrium of the entire Bayesian persuasion game. However, if we only tried to study arbitrary Nash equilibria of the Bayesian persuasion game, the receivers could try to deter the sender from choosing misleading signals by punishing such a choice with always voting A, regardless of the signal. This would always give the sender worse utility, but in state β, it would also give the receivers worse utility. Therefore, this is not a credible threat, and we avoid it by focusing on strategies forming a subgame perfect equilibria, where the receivers must maximize their expected utility under the sender's choice of conditional distributions; they cannot merely spite her for her strategy by voting in a way that makes the sender worse off if it also makes themselves worse off.

We are now prepared to present an example of such a Bayesian persuasion game.

Example (One-Shot Trial):
In this example, we take the sender to be a biotechnology company seeking commercial approval for a new procedure from $n = 15$ receivers—members of a regulatory committee. The procedure is potentially dangerous, and so each member of the committee votes either A, continue testing, or B, stop testing, after viewing just one signal s, the results of the procedure from one patient: a, success, or b, failure. The voting rule v is taken as majority-rule. There are two states of the world that describe the procedure: α, effective, and β, ineffective; both company and committee have a common belief about which is the case. However, regardless of which is the case, the company would always prefer to continue testing, whereas the committee prefers this only when the procedure is effective. The company's conditional distributions from which they draw the signal can be thought as reflecting implementation details of the procedure, e.g. they may design it so that it is always a success when the procedure is effective ($\pi(\text{success}|\text{effective}) = 1$) but still may appear to work even when the procedure is ineffective ($\pi(\text{success}|\text{ineffective}) > 0$). By law, the committee must observe these implementation details, and uses their technical expertise to deduce the conditional distributions corresponding to them.

This is an example of public persuasion, whose subgame perfect equilibria we analyze in the next section.

3 Multiple Receivers: Public and Private Persuasion

We analyze two general instances of the above Bayesian persuasion game: public persuasion and private persuasion.

3.1 Public Persuasion

Public persuasion is characterized by the sender communicating to all of the receivers simultaneously in a group setting. She generates just one signal from her chosen condition distributions, which is observed by all of the receivers, and each receiver knows that all of the other receivers observed that same signal (it is common knowledge). Wang (2013) makes a simplifying assumption to focus on the salient case. Let m be equal to the number of votes required under the voting rule for a particular decision (e.g. $m = \lceil \frac{n}{2} \rceil$ for majority-rule).

Assumption 1. *If the sender does not send any signal, with the receivers maximizing their expected utility under the common prior* p*, there are less than* m *receivers that vote for* B*:*

$$(1 - p(\alpha))u_i(A, \beta) + p(\alpha)u_i(A, \alpha) \geq (1 - p(\alpha))u_i(B, \beta) + p(\alpha)u_i(B, \alpha), m \leq i \leq n$$

Rearranging, the author rewrites this condition as $q_i \geq 1 - p(\alpha), m \leq i \leq n$, where we have assumed that the receivers are indexed such that $q_n \geq \cdots \geq q_1$. Each q_i represents a threshold doubt for receiver i; the larger it is, the more skeptical they are to vote for B, due to a combination of low probability that β is the state of the world and a high cost in utility for voting for B when α is true. Then under modest technical conditions, the following results holds.

Theorem 3.1 (Subgame Perfect Equilibrium in Public Persuasion)**.** *Under public persuasion, the sender induces a subgame perfect equilibrium by choosing conditional distributions π^*_{pub} characterized by*

$$\pi^*_{pub}(b|\alpha) = \frac{(1-q_m)(1-p(\alpha))}{q_m p(\alpha)}, \pi^*_{pub}(b|\beta) = 1.$$

Proof. We summarize the proof given by the author. The first thing to note is that each receiver votes sincerely, i.e. in a way that maximizes his expected utility by combining just his own observed signal with the prior distribution via Bayes' rule; this is intuitive because each receiver observes the same signal in public persuasion. The sender has to choose $x = \pi(b|\alpha)$ and $y = \pi(b|\beta)$. Technical conditions ensure that no receiver votes for B if the signal is a; this is largely a normalization condition that forces the sender to choose conditional distributions such that β is more likely under b than a. Then, the sender only wants to maximize the probability that the signal b is observed, subject to credibility constraints about $\mu(\beta|b) \equiv \frac{y(1-p(\alpha))}{xp(\alpha)+y(1-p(\alpha))}$ with respect to the threshold doubts of the receivers:

$$\max_{(x,y)\in[0,1]^2} p(\alpha)x + (1-p(\alpha))y$$

$$\text{subject to} \frac{y(1-p(\alpha))}{xp(\alpha)+y(1-p(\alpha))} \geq q_m.$$

The existence of a solution is guaranteed by nice properties of this problem (e.g. continuity, compactness). Using the method of Lagrange multiplier gives the result. □

The author also shows that there are exactly m voters who vote for B when $s = b$, which is intuitive because we sought to set $\mu(\beta|b) \geq q_m$ as to convince only the m receivers with the smallest threshold values; convincing more would require a more accurate signal, reducing the proportion of the time b is realized and the vote is B. It is not surprising, then, that the above optimization makes this inequality sharp.

3.2 Private Persuasion

Under private persuasion, the sender communicates to all of the receivers separately, generating one signal from her chosen conditional distributions for each receiver. Each receiver only sees this signal, but it is common knowledge that each receiver sees a different signal realization. The signals are independent. Consider first:

Lemma 3.2. *Under private persuasion, if all receivers vote sincerely, the sender maximizes her expected utility by choosing π_{sin}, given by*

$$\pi_{sin}(b|\alpha) = \frac{(1-q_t)(1-p(\alpha))}{q_t p(\alpha)}, \pi_{sin}(b|\beta) = 1,$$

where $q_t \geq q_m$ is a parameter.

This time, the sender needs to convince some number of voters $t \geq m$ to vote for B upon observing β, as it is quite unlikely that all m receivers with the lowest threshold doubt observe the signal b. Moreover, we cannot assume the receivers vote sincerely, because there is additional information about the state of the world expressed by other receivers' signals, which they do not see. But each receiver can still condition on what these signals must be in the hypothetical scenario they are the *pivotal vote*, which is the only case in which a receiver (given the votes of the others') can actually change the vote. Essentially, if a receiver would vote one way under sincere voting but anticipates being the tie-breaking vote, other receivers probably received a different signal, corresponding to a different inference of the state of the world via Bayes' rule (modulo differing threshold doubts). This makes receivers more conservative in trusting their signal, which is less favorable to the sender since she is trying to persuade the receivers (via signals) that the state of the world is β. More formally:

Theorem 3.3 (Subgame Perfect Equilibrium in Private Persuasion)**.** *Under private persuasion, the sender induces a subgame perfect equilibrium by choosing conditional distributions π^*_{pri}, which have no closed-form expression but satisfy $\pi^*_{pri}(b|\alpha) < \pi_{sin}(b|\alpha), \pi^*_{pri}(b|\beta) \leq 1$*

Proof. The proof is similar to that of Theorem 3.1, except that the optimization problem is formulated using the binomial distribution to reason about the number of times a particular signal is drawn (since the signals are independent), and the modified threshold voting behavior elucidated above. Then the stated estimates on the solution of the problem can be obtained. □

3.3 Analyzing the Settings

Using that $q_t \geq q_m$, we have $\pi^*_{pub}(a|\alpha) < \pi^*_{pri}(a|\alpha)$ and $\pi^*_{pri}(b|\beta) \leq \pi^*_{pub}(b|\beta)$, which corroborates our above analysis of the difficulties of the sender in persuading the receivers under private persuasion vis-à-vis public persuasion. Relatedly, in the above subgame perfect equilibria, the sender's expected utility is strictly worse under private persuasion than public persuasion. However, both provide better utility than no persuasion, in which case the receivers vote based on the common prior (inducing action A, per Assumption 1). In some sense, the sender is able to design a signal in a way that conveys to the receivers when β is probably the state of the world (for which they would want to vote B), still adding some noise to the signal so that the receivers sometimes vote for B even when α is the state of the world. As expected by how the conditional distributions differ, the receivers vote for the "wrong" decision (i.e. A in state β or B in state α) with higher probability under public persuasion. The author shows that these differences are robust to adding multiple independent signal realizations in public persuasion, with the sender merely generating unfavorable signals less often to mitigate the probability the receivers believe they are in state α when a is drawn.

3.4 Applications, Limitations, and Future Considerations

Our analysis reveals that, assuming a subgame perfect equilibrium is played, the sender receives higher expected utility under public persuasion compared with private persuasion, and the author even shows that she attains her maximum utility under his Bayesian persuasion set-up. This gives the sender an incentive to orchestrate her persuasion such that it aligns with the public persuasion context. Moreover, this does not require coordinating with all the receivers; with the advent of the internet, in many less-regulated settings the sender could easily transmit such a signal in an online manner, as each receiver still observes the same signal and hence does not benefit in communicating with the other receivers. On the other hand, under public persuasion the receivers vote for the option that gives worse expected utility with higher probability, which heuristically suggests that they tend to receive worse expected utility, although a sharper characterization of their performance would be preferable in future work. However, if the receivers simply attain utility 1 for vote A in α and vote B in β (0 otherwise), maximization of utility coincides with voting for the option for which the corresponding state occurs with the greater probability, in which case these error probabilities correspond to the probability of incorrectly appraising the state of the world; then, minimizing this sum of probabilities would be intrinsically appealing for epistemic reasons, justifying private persuasion. On the other hand, the sender does not have a closed-form expression for their optimal conditional probabilities, only upper bounds, and so the prescribed subgame perfect equilibrium may be difficult to implement. Perhaps its computation (especially for a given instance of the game, to be repeated with distinct receivers) could be approximated by having the sender use a learning algorithm, which may still give good error guarantees, although per the above discussion, this is only likely to happen is the sender is mandated to render private persuasion.

It should also be noted that, regardless of which setting is chosen, the sender faces complications if the receivers solicit more signal realizations *after* the sender has chosen her signal mechanism. Above, the sender was able to achieve the same expected utility under public persuasion with an arbitrary number of independent signal draws because she was able to adjust her signal mechanism accordingly. As the number of independent signal draws tends to infinity, by the strong law of large numbers, the proportion of signals that are α tends to the probability of that signal under whichever state of the world the players are in, which will be different for the two states of the world (unless the sender chooses identical conditional distributions, which by symmetry do not change the receivers' beliefs); there are also algorithms in this vein that state the number of trials needed to guarantee a correct conclusion about the state of the world with desired probability (e.g. see the Chernoff bound). Such a tactic would allow the receivers to vote for the "correct" decision with arbitrarily high probability, leaving the sender to get their preferred action B only with probability arbitrarily close to β. Naturally, the sender wants to avoid this by only giving information from a pre-determined number of signals, but receivers may try to circumvent the set-up described by Wang (2013).

Another drawback of the paper is that it only considers the expected utility of each player, which is the framework for subgame perfect equilibria, but misconstrues the incentives of risk-averse and risk-seeking players, who, respectively, tend to have lower and higher tolerance for variation in the utility that they receive.

In the paper, the author eventually allows S to be a continuous space, $[0, 1]$. Likewise, it would be interesting to allow $T = [0, 1]$, with perhaps the spaces corresponding to each other in an even more nuanced way. For instance, the signal could represent the efficacy of a drug in a clinical trial, whereas the state of the world represents the true efficacy of a drug; inferring the latter from the former is still amenable to Bayesian methods, e.g. maximum likelihood estimation. On the other hand, it seems somewhat unrealistic that the sender would be able to generate signal mechanisms with arbitrary probabilities in $[0, 1]$, as they only seem to have finitely many choices for the mechanism they set up in most real-world contexts. The sender and the receiver may also not agree on the actual distributions that describe how the signals are generated, which makes it difficult for the sender to anticipate how the receivers will maximize their expected utility. The

sender could try to convey this information, but has an incentive to manipulate the receivers' posterior beliefs so they conclude the state of the world is β; in other words, this introduces a sub-Bayesian-persuasion-game, which perhaps no base case of this recursive difficulty.

The predictions of the paper are sharp in the sense that there is a unique subgame perfect equilibrium in each instance of the above Bayesian persuasion game, but it may be difficult for players who are not fully rational to implement the above calculations (e.g. requiring calculus-based optimization), and receivers, especially if non-experts in a given domain (e.g. politicians), may instead succumb to mental heuristics that view the sender with undue skepticism, even though their incentives are aligned in state β, in which the signal b is disproportionately likely to appear, thus increasing the probability that β truly is the state of the world through Bayes' rule when b is observed.

4 Single Receiver Persuasion

With almost 700 cited sources/references, Kamenica and Gentzkow (2011), heavily focus on the mathematics and applications of Single Receiver Persuasion. They first provide context on how we see "persuasion" in our day-to-day lives but highlight how it is especially relevant to consider its influence in "advertising, courts, lobbying, financial disclosure, and political campaigns, among many other economic activities." The authors initially provide the relevant definitions and this context through an interesting example of a prosecutor persuading the judge of a guilty verdict. They then demonstrate the effectiveness of transmission from sender to receiver and how it depends on the correlation of their preferences. Conversely to other works, these authors conclude that aligning these preferences in equilibrium decreases the amount of communication between sender and receiver. They provide a mathematical model of persuasion where they consider various propositions and assumptions along with their respective conclusions. They then apply these results to examples of lobbying and free product trials. Lastly, they extend their results to cases where the receiver has private information and cases where there are multiple receivers.

4.1 Key Assumption

1. The main assumption of the model is that the sender cannot distort or conceal information from the state of the world once the true realization of the world is known.
2. If the receiver is indifferent between some actions at a given belief, then she takes the decision which maximizes the sender's utility.
3. Equilibrium means **sender-preferred sub-game equilibrium**.
4. There are at least 2 actions in receiver's action space
5. Any signal that induces the same action distribution for the receiver is of same value to the sender. If we know that what is the expected utility for the sender corresponding to a posterior distribution on receiver's actions, we can assign the value to sender's signal. These is a one-to-one relationship between posterior distribution and value of sender's action.

4.2 Notation Setup

- $\omega \in \Omega$ State of the world
- $a \in A$ Sender's and/or receiver's Action.
- $u(a, \omega)$ Receiver's utility
- $v(a, \omega)$ Sender's Utility
- $\tau \in \Delta(\Delta(\Omega))$ Distribution of posteriors.
- $\mu_0 \in int(\Delta(\omega))$ Prior probability (Sender and receiver share this belief)
- π where $\pi(.|\omega))_{\omega \in \Omega}$ Signal
- μ_s Receiver's posterior
- $a^*(\mu_s) = arg\,max_{a \in A} E[u(a, \omega)]$ Receiver's action based on the posterior

4.2.1 Properties of Optimal Signal

1. The sender's least preferred state is chosen with absolute surety.
2. When the receiver picks the action which is sender's least likely action, he/she is indifferent between the two choices.

Definition 3 (Bayes' Plausible)**.** *A distribution of posteriors is Bayes plausible if the expected posterior probability equals the prior :* $\sum_{Supp4.2.1(\tau)} \mu\tau(\mu) = \mu_0$ [1]

Definition 4 (Straightforward Signal)**.** *A signal is one which makes the receiver pick a recommended action.*

4.3 Examples

4.3.1 Judge-Prosecutor Example

In this example, sender is a prosecutor and receiver is a judge. The state of the world is the guilt of the defendant. Choice of signal is prosecutor's actions like : witness questions, DNA tests, cross questions, etc.

1. **Set-up**
   Given that on average, a judge acquits a defendant 30% of the time and convicts the defendant 70% of the time. A prosecutor wants (ideally) to make the judge give the verdict as guilty, irrespective of whether the defendant is innocent or guilty. In game theory terminology, the judge gets a $utility(u^{4.2})$ of 1 if he/she gives the correct verdict i.e. guilty when guilty and innocent when innocent. On the other hand, the prosecutor gets a $utility(v^{4.2:})$ of 1 when the defendant is charged guilty.
2. **Aim**
   Can the prosecutor change the judge's information world so much so that the judge, on average, convicts any defendant more number of times than he/she acquits the defendant? The answer is yes, and we'll see how. The prosecutor is required to conduct an investigation and also is required by the law to report the investigation fully and honestly. However, there are no prohibitions on how he conducts the investigation. This means, he/she can question witnesses who makes his/her case stronger and gather evidence which is solely against the defendant. Following the research, the defendant signals the verdict of the investigation with probability distribution:

   - When innocent: $\pi(i/innocent)$ OR $\pi(g/innocent)$
   - When guilty: $\pi(i/guilty)$ OR $\pi(g/guilty)$

   where i is the case when the investigation's resulting signal is "innocent" and g the case when resulting signal is "guilty".
3. **Optimal Strategy** The optimal strategy for the prosecutor to signal "innocent" and "guilty" are with conditional probabilities :

   $$\pi(i/innocent) = 4/7 \qquad \pi(i/guilty) = 0 \qquad \pi(g/innocent) = 3/7 \qquad \pi(g/guilty) = 1$$

   Here i and g are $straight - forward$[4] signals since they make the judge take the actions recommended by the prosecutor. In the prosecutor's ideal world where the judge blindly follows the results of the investigation, the prosecutor can signal "always guilty" which would make the judge always signal guilty. But in a persuasion setup, the receiver(judge) is aware of the fact that sender's signal is biased and are intended to make the action go in sender's favor. Thus, the prosecutor comes up with sort of a "clever" distribution where the judge is partially in belief that the investigation is honest and unbiased to a large extent.

   With these numbers, the judge verdicts "guilty" 60% of the time despite having a prior probability of charging "guilty" only 30% of the time. :

   $$\pi(g/innocent) + \pi(g/guilty) = (0.3 * 0.7)/0.7 + (0.3 * 0.3)/0.3 = \mathbf{0.6} \qquad \text{(BAYES' RULE)}$$

4.3.2 Lobbying

Figueiredo et al. (2014) define lobbying as the transfer of information in private meetings and venues between interest groups and politicians, their staff, and agents in order to influence and **persuade** decision making. Figueiredo et al. (2014) mention that in 2012, nearly $3.5 billion were spent in lobbying the federal government while other interest group affiliated to political parties spent another $1.55 billion during the 2011-2012 election cycle. With such huge investments from interest groups toward pushing forward their agenda through organized persuasion, it becomes all the more important to question how much / what methods of lobbying are ethical. Quantifying lobbying through persuasion models by setting parameters on information communication becomes crucial in such scenarios. Going with the single receiver persuasion

[1] supp is short for support.
The support of a mixed strategy s_i for a player i is the set of pure strategies $\{a_i | s_i(a_i) > 0\}$

model that Kamenica and Gentzkow (2011) illustrates, the commitment assumption for this example would require the lobbyist's information to be fully available to the politician (receiver) and he/she can make **verifiable claims** on the information.
To model such a scenario, let's say the lobbyist's preferred action is modeled by $a^* = \alpha\omega + (1-\alpha)\omega^*$ but is biased toward ω^*. A.1 shows that that $\alpha > \frac{1}{2}$, the preferences of sender and receiver's align, and that full disclosure of information becomes immediately optimal. When $\alpha < \frac{1}{2}$, the preferences are different and concealing information becomes optimal. One important conclusion here is that the lobbyist either chooses to reveal all of the information or does not reveal it at all. Thus in conclusion, when the communication is modeled in a way that whole information is required to be published and need to be made "verifiably true", it often makes sense for the interest groups (example: big pharmacy, tobacco, etc.) not to invest in such studies as this would be counter productive for pushing their interests further. Moreover, since such studies are still funded by big corporations, it hints toward the fact that the receiver (politician/consumer) is not fully rational.

4.4 Related Work

Kamenica (2019) himself explores Bayesian persuasion to answer some further questions like: **1.** Schools can improve students' job prospects if they give strict evaluations on exams. **2.** Google can de-congest roads by providing incorrect data to users showing traffic when there is not. In essence, what is the optimal information that should be revealed by these senders? Some other works in the field of persuasion model the problem by relaxing some of the pre-conditions taken by Kamenica and Gentzkow (2011). Castiglioni et al. (2020) argue that the condition that the sender needs to know receiver's payoff/utility is too strict a condition to model some real life scenarios.

4.5 Conclusion and Critical Comments

The model study of Bayesian persuasion introduced by Kamenica and Gentzkow (2011) has been instrumental in applying the infrastructure to some real world scenarios. Crowd voting is one examples of it and has been studied in detail by Alonso and Câmara (2016). Online advertisement is also another field of application and Badanidiyuru et al. (2018) explore how Bayesian persuasion can be applied to it. Kamenica (2019) articulate the three ways to induce someone to do something as either to incentivize by affecting one's marginal utility (payment, coercion, complementary goods) or through engendering an outcome by facilitating the decision making process for the receiver via technology. Lastly, through altering the state of the world by changing one's belief system through persuasion. Focusing on the third way, they remind the reader that their studies set up both the sender who is doing the persuasion and the receiver who is receiving it as rational Bayesians. The model is reasonable in situations where the model can be applied to. But the model does make some strict assumptions like the fact that the sender needs to know receiver's utility function before making a decision on the signal which might not be reasonable in many real world scenarios. There can be cases where the sender meets a receiver whose 'type' is unknown and there could be multiple rounds of such interactions between a sender and multiple receivers. For example, A defender trying to design an optimal strategy to protect critical resource (maybe) from a hacker. In this situation, the defender needs to design the best optimal strategy irrespective of the type and strategy of the attacker known before-hand. Castiglioni et al. (2020) model this situation through *Stackelberg games* where the sender (often called leader) signals first and then receivers (called followers) follow. Thus by taking some pre-conditions, Kamenica and Gentzkow (2011) miss out on certain scenarios which have thoroughly been explored by other researchers.

References

Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11):3590–3605, November 2016. doi: 10.1257/aer.20140737. URL https://www.aeaweb.org/articles?id=10.1257/aer.20140737.

Ashwinkumar Badanidiyuru, Kshipra Bhawalkar, and Haifeng Xu. Targeting and signaling in ad auctions. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, page 2545–2563, USA, 2018. Society for Industrial and Applied Mathematics. ISBN 9781611975031.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16188–16198. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ba5451d3c91a0f982f103cdbe249bc78-Paper.pdf.

de Figueiredo, John M., Richter, and Brian Kelleher. Advancing the empirical research on lobbying. *Annual Review of Political Science*, 17(1):163–185, 2014. doi: 10.1146/annurev-polisci-100711-135308. URL https://doi.org/10.1146/annurev-polisci-100711-135308.

Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272, 2019. doi: 10.1146/annurev-economics-080218-025739. URL https://doi.org/10.1146/annurev-economics-080218-025739.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011. doi: 10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Yun Wang. Bayesian persuasion with multiple receivers. June 2013. doi: 10.2139/ssrn.2625399. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2625399.

A Appendix

A.1 Optimal Strategy in Lobbying

Kamenica and Gentzkow (2011) show that the lobbying mechanism can be modelled as follows:
Interest group's preferred action : $a^* = \alpha\omega + (1-\alpha)\omega^*$ with $\alpha \in [0,1]$
Politician's payoff : $u = (a-\omega)^2$ Lobbyist's payoff $v = -(a-a^*)^2$
Also, $\hat{a}(\mu) = E_\mu[\omega]$
Simply putting the value of a^* in lobbyists's payoff gives :

$\hat{v}(\mu) = -(1-\alpha)^2\omega^{*2} + 1(1-\alpha)^2\omega^* E_\mu[\omega] - \alpha^2 E_\mu[\omega^2]$

Differentiating this with respect to α and setting to zero shows that the function is

- Strictly concave when $\alpha > \frac{1}{2}$
- Strictly convex when $\alpha < \frac{1}{2}$