Persuasion with Verifiable Information[*]

Maria Titova[†]

May 2022

Abstract

The large and growing applied Bayesian persuasion literature is sometimes criticized for assuming that the sender can commit to an experiment that reveals a signal based on the realized state of the world. This paper shows that if the sender's preferences are state-independent, the receiver is choosing between two actions, and the state space is sufficiently rich, then the sender reaches the full-commitment payoff in an equilibrium of the disclosure game with verifiable information. The latter setup is more natural in the applications to judicial systems, electoral campaigns, product advertising, financial disclosure, and job market signaling.

KEYWORDS: persuasion, value of commitment

JEL CLASSIFICATION: D72, D82, D83, M37

[*]I am very grateful to Renee Bowen and Joel Sobel for their guidance and support. I thank S. Nageeb Ali, Arjada Bardhi, Simone Galperti, Remy Levin, Aleksandr Levkun, Elliot Lipnowski, Denis Shishkin, Joel Watson, Alex Weiss, and the audience members at Amsterdam, Bern, BSE, CIDE, Princeton, Manchester, Vanderbilt, UCSD, SWET 2018, YES 2019, EGSC 2019, CSQIEP 2020, and SITE 2021. All errors are my own.

[†]Department of Economics, Vanderbilt University. E-mail: maria.titova@vanderbilt.edu.

1. INTRODUCTION

Suppose a sender is privately informed about the state of the world and would like to convince a receiver to take his favorite action. The sender does not have commitment power, but his messages are verifiable. On average, what is the best outcome that the sender can hope for?

Persuasion with verifiable information plays an essential role in courtrooms, electoral campaigns, product advertising, financial disclosure, job market signaling, and many other economic situations. For example, in a courtroom, the prosecutor tries to persuade the judge to convict the defendant by selectively presenting inculpatory evidence. In an electoral campaign, a politician carefully chooses which campaign promises he can credibly make in order to win over voters. In advertising, a firm convinces consumers to purchase its product by highlighting only specific product characteristics. In finance, a CEO divulges only certain financial statements and indicators to board members in order to obtain higher compensation. In a labor market, a job candidate lists specific certifications in order to make his application more attractive to an employer.

I consider the following formal model of persuasion with verifiable information. There is an underlying continuous space of possible states of the world, which is a unit interval. The sender is fully informed about the state of the world (he knows the point on the unit interval), but his preferences do not depend on it. Metaphorically, I say that the sender wants the receiver to approve his proposal. The receiver is uninformed about the state of the world, which to her is payoff-relevant. The sender communicates with the receiver using verifiable messages. Each messages is a subset of the unit interval, interpreted as a statement about the state of the world. Verifiability means that the message contains the truth (the true state of the world), but not necessarily the whole truth (it may contain other things, as well). The receiver chooses between two options: to approve or reject the proposal.

How does the sender convince the receiver using verifiable information? Rather than looking at the sender's messages and the receiver's beliefs, I focus on what the receiver does in each state of the world. Since she chooses between two options, we can partition the state space into two subsets: the states in which she approves the proposal and the states in which she rejects it. My first result says that a subset of the state space is an equilibrium set of approved states if and only if it satisfies two constraints. The first is the *sender's incentive-compatibility (IC) constraint*, which ensures that the sender does not wish to deviate toward a fully informative strategy that induces the receiver to act as if she knows the state of the world. Conveniently, that is the only deviation of the sender that needs to be checked. The second constraint is the *receiver's obedience constraint*, which ensures that the receiver approves the proposal whenever her expected net payoff from approval is non-negative.

In the sender's least preferred equilibrium, his ex-ante odds of approval are minimized across all equilibria. The receiver learns whether she would approve under complete information, and makes a fully informed choice. This is the equilibrium in which *full unraveling* takes place.

In the sender's most preferred equilibrium, his odds of approval are maximized subject to the receiver's obedience constraint. Specifically, the sender pools the "good" states, in which the receiver prefers to approve, with some of the "bad" states, in which the receiver prefers to reject. The solution is characterized by a cutoff value of the receiver's net payoff from approval: she approves whenever it is not too negative. When the receiver approves, her obedience constraint binds, and she is indifferent between approval and rejection. The sender improves his ex-ante payoff over full disclosure because the receiver approves in some of the "bad" states. In fact, in his most preferred equilibrium, the sender reaches the commitment payoff. This observation bridges the gap between the verifiable disclosure literature introduced by Milgrom (1981) and Grossman (1981), and the Bayesian persuasion literature pioneered by Kamenica and Gentzkow (2011). The sender does not need ex-ante commitment power; he can persuade the receiver with evidence alone.

The Value of Commitment

To see why the sender does not benefit from commitment, let us revisit the canonical example from Kamenica and Gentzkow (2011), in which a prosecutor wishes to persuade a judge to convict a defendant. In their setup, the state of the world is binary: the defendant could be guilty or innocent. The judge prosecutes if the probability that the defendant is guilty is above 0.5. The prior that the defendant is guilty is 0.3. The authors write, "The judge knows 70 percent of defendants are innocent, yet she convicts 60 percent of them!"[1] How does this happen? When the defendant is guilty (with the prior probability of 0.3), the judge convicts them for sure. When they are innocent (with the prior probability of 0.7), the judge convicts them with probability $3/7$.

Can we replicate this outcome if the prosecutor knows the true state of the world, does not have ex-ante commitment power, but can send verifiable messages, which have to include the true state? The answer is no. When the defendant is guilty, the prosecutor has two available messages: "The defendant is guilty" (the fully revealing message) and "The defendant could be innocent or guilty" (the fully uninformative message). Notice that the second message is available in both states (i.e., when the defendant is innocent or guilty). If the prosecutor sends this message when the defendant is guilty and the judge convicts, then the prosecutor would also send that message when the defendant is innocent. As a result, there is no incentive-compatible way for the prosecutor to get the judge to convict all the guilty defendants, and some of the innocent ones.

Now consider the following continuous interpretation of the same situation. Suppose that, instead of being innocent or guilty, the defendant has some level of guilt between 0 (0% guilty) and 1 (100% guilty). The prior is uniform on $[0, 1]$. The judge wishes to convict the defendants who are at least 70% guilty. Note that from the judge's point of view, the prior that the defendant is *sufficiently* (over 70%) guilty is 0.3. At the same time, the judge's best response under incomplete information is to convict whenever the probability that the defendant is sufficiently guilty exceeds 0.5. Thus, the continuous interpretation and the example of Kamenica and Gentzkow (2011) tell the same story.

With a richer state space comes a richer message space: now, the prosecutor can send any message that includes the actual state. In particular, the prosecutor can send the message $[0.4, 1]$ whenever the defendant is at least 40% guilty and the message $[0, 0.4]$ otherwise. When the judge hears the former message, she concludes that the defendant is, on average, 70% guilty and convicts. The prosecutor has no profitable deviations from this strategy. When the defendant is indeed at least 40% guilty, the prosecutor receives

[1] See Kamenica and Gentzkow (2011), p. 2591.

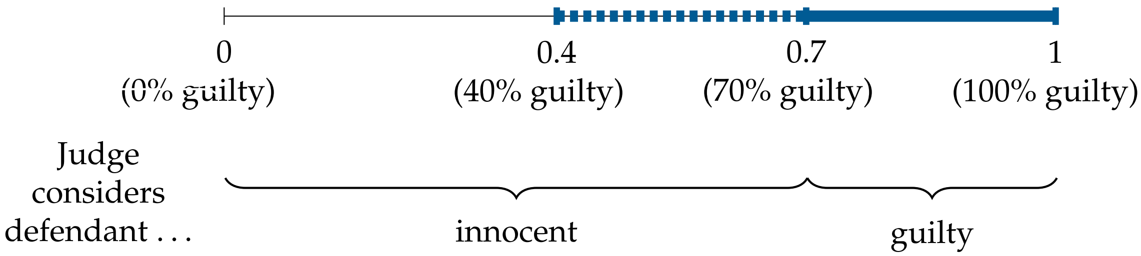

Figure 1. *The sender-preferred equilibrium of the verifiable-information game with uniform prior. The prosecutor sends the message* $[0.4, 1]$ *(blue) when the defendant is at least 40% guilty, and* $[0, 0.4]$ *otherwise. The judge convicts after the former message and acquits after the latter. The judge convicts 60% of all defendants even though her prior is that 70% are innocent.*

the largest possible payoff. When the defendant is less than 40% guilty, the prosecutor cannot credibly claim otherwise.

Notice that the judge convicts all the defendants who are at least 70% guilty, as well as $3/7$ of the defendants she considers innocent (those less than 70% guilty). In other words, the verifiable-information game leads to the same prosecutor-preferred threshold of conviction at 40% as the Bayesian persuasion game. In Bayesian persuasion, this threshold comes from calculating the optimal signal that persuades the judge to convict. In the verifiable-information game, it comes from calculating the prosecutor-preferred message that convinces the judge to convict. Regardless of the setup, the constraints are the same and boil down to the judge's interpretation of the *signal realization* (in Bayesian persuasion) or the *message* (in the verifiable disclosure game) as a recommendation to convict. The richness of the message space allows for exactly the same solution.

Related Literature

I assume that the sender communicates with the receiver using verifiable messages. This communication protocol was introduced by Milgrom (1981) and Grossman (1981). Other communication protocols include cheap talk (Crawford and Sobel, 1982) and Bayesian persuasion (Kamenica and Gentzkow, 2011). Relative to these other models of communication, Bayesian persuasion makes the sender better off because it endows him with ex-ante commitment power. Lipnowski and Ravid (2020) find that the sender's maximal equilibrium payoff from cheap talk is generally strictly lower than his payoff under commitment. Consequently, a cheap-talk sender values commitment.[2] In contrast to the result of Lipnowski and Ravid (2020), I show that the sender does not necessarily benefit from commitment if his messages are verifiable.

There is an extensive applied Bayesian persuasion literature. It includes settings in which pharmaceutical companies persuade the Food and Drug Administration to approve their drug (Kolotilin, 2015); schools persuade employers to hire their graduates (Ostrovsky and Schwarz, 2010; Boleslavsky and Cotton, 2015); matching platforms persuade sellers to match with buyers (Romanyuk and Smolin, 2019); politicians persuade voters (Alonso and Câmara, 2016; Bardhi and Guo, 2018); and governments persuade citizens (Gehlbach and Sonin, 2014; Egorov and Sonin, 2019). My contribution is to show that in all these applications, one can replace the assumption that the sender has commitment power with the assumption that the sender's messages are verifiable.

2. MODEL

There is a state space $\Omega := [0,1]$, one sender (he/him), and one receiver (she/her). The game begins with the sender observing the state of the world $\omega \in \Omega$, which is drawn from an atomless common prior distribution $p > 0$ over Ω.[3] Having observed the state of the world, the sender chooses a message m from the set M of all (Borel) subsets of Ω. His message must be verifiable: he cannot send m if $\omega \notin m$.[4] The sender's preferences depend only on the receiver's action: his payoff u_s is 1 if the receiver approves and 0 if she rejects.

[2] Lipnowski (2020) also notes that the sender reaches the commitment outcome with cheap talk if his value function is continuous in the receiver's posterior belief. That assumption is very restrictive: when the receiver is choosing between two options and the sender's preferences are state-independent, the sender's value function must be constant, meaning that no communication takes place under cheap talk, verifiable information, and Bayesian persuasion. I thank Elliot Lipnowski for this insight.

[3] For a compact metrizable space S, ΔS denotes the set of all Borel probability measures over S, endowed with a weak* topology. For $q \in \Delta\Omega$ and any Borel subset of the state space $W \subseteq \Omega$, $Q(W) = \int_W q(\omega)d\omega$ is the measure of W, and $q(\cdot \mid \cdot)$ is the conditional probability distribution: $q(\omega \mid W) = 1$ if $W = \{\omega\}$ and $q(\omega \mid W) = \frac{q(\omega)\cdot\mathbb{1}(\omega\in W)}{Q(W)}$ if $Q(W) > 0$.

[4] I borrow from Milgrom and Roberts (1986) the definition of a verifiable message as a subset of the state space that includes the true realization. This method satisfies normality of evidence (Bull and Watson, 2004), which means that it is consistent with both major ways of modeling hard evidence in the literature.

The receiver chooses between approval (action 1) and rejection (action 0). Her preferences are described by a utility function $u : \{0,1\} \times \Omega \to \mathbb{R}$. The receiver approves (the proposal in) state ω if her *net payoff from approval* $\delta(\omega) := u(1,\omega) - u(0,\omega)$ is non-negative.[5] I define the receiver's complete-information *approval set* $A := \{\omega \in \Omega \mid \delta(\omega) \geq 0\}$ to include all the states of the world she wishes to approve under *complete information*. I assume that the receiver rejects the proposal under her prior belief, i.e. $\mathbb{E}_p[\delta(\omega)] < 0$.

EXAMPLE 1 (THE PROSECUTOR AND THE JUDGE). Here I introduce the continuous version of the seminal example from Kamenica and Gentzkow (2011). The sender is the prosecutor and the receiver is the judge. The state of the world reflects how guilty the defendant is, with values ranging from 0% guilty (0) to 100% guilty (1). The prosecutor observes the state of the world drawn from the common uniform prior.

The judge wishes to convict the defendants who are "sufficiently guilty". Specifically, the judge has a (known) threshold for conviction $v \in \Omega$; her net payoff from approval is $\delta(\omega) = 1$ if $\omega \geq v$ and $\delta(\omega) = -1$ if $\omega < v$, and her complete-information approval set is $A = \{\omega \in \Omega \mid \omega \geq v\}$. Figure 2 illustrates the preferences of the judge with $v = 0.7$.

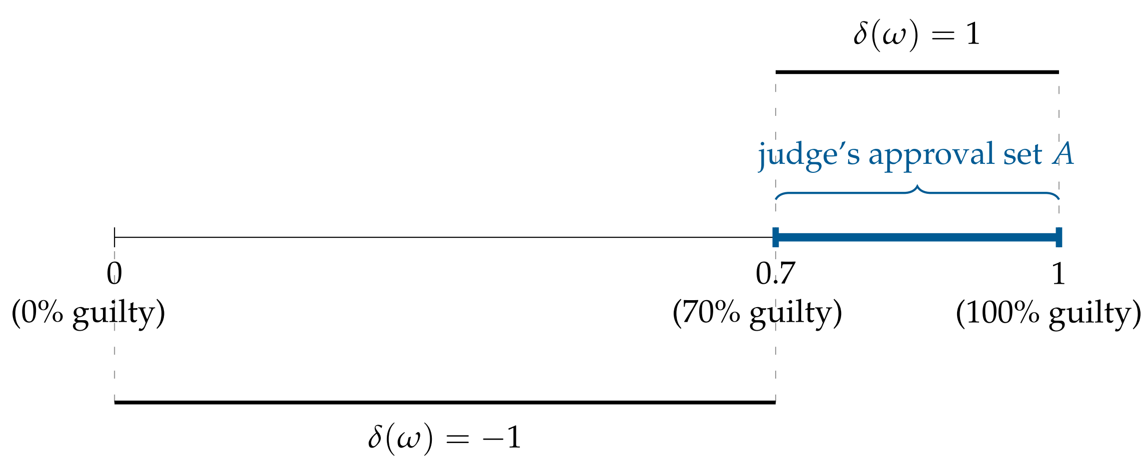

Figure 2. *The state space Ω, with the judge's threshold for conviction at $v = 0.7$. Her net payoff from approval $\delta(\omega)$ equals 1 if the defendant is at least 70% guilty, and -1 otherwise. She prefers to convict whenever her net payoff from approval is positive.*

[5] I assume that the receiver breaks ties in favor of approval when $\delta(\omega) = 0$. This tiebreaker is necessary for the existence of the sender-preferred equilibrium, and is inconsequential in all other equilibria.

EQUILIBRIUM OUTCOMES

I consider perfect Bayesian equilibria of this game. An *equilibrium* consists of three measurable maps—a sender's strategy $\sigma : \Omega \to \Delta M$, a receiver's approval strategy $a : M \to \{0, 1\}$, and a receiver's belief system $q : M \to \Delta\Omega$—satisfying the conditions below.

DEFINITION 1. *A triple* (σ, a, q) *is an* equilibrium *if*

(i) $\forall \omega \in \Omega$, $\sigma(\cdot \mid \omega)$ *is supported on* $\arg\max_{m \in M} u_s\big(a(m)\big)$, *subject to* $\omega \in m$*;*

(ii) $\forall m \in M$, $a(m) = \mathbb{1}\left(\mathbb{E}_{q(\cdot \mid m)}[\delta(\omega)] \geq 0\right)$*;*

(iii) $\forall m \in M$ *such that* $\int_\Omega \sigma(m \mid \omega) dP(\omega) > 0$, $q(\omega \mid m) = \frac{\sigma(m \mid \omega) \cdot p(\omega)}{\int_\Omega \sigma(m \mid \omega') dP(\omega')}$*;*

(iv) $\forall m \in M$, $\operatorname{supp} q(\cdot \mid m) \subseteq m$.

In words, *(i)* states that the sender sends a message with positive probability only if it maximizes his payoff; *(ii)* states that the receiver approves the proposal whenever her expected net payoff from approval is non-negative under her posterior belief; *(iii)* states that the receiver's posterior beliefs are Bayes-rational on the equilibrium path; *(iv)* states that the receiver's posterior beliefs on and off the path are concentrated on the states in which the sender can verify the message.

An outcome of the game is a description of what action the receiver takes for each realization of the state of the world.

DEFINITION 2.

- *An* outcome $\alpha : \Omega \to [0, 1]$ *specifies* $\forall \omega \in \Omega$ *the probability* $\alpha(\omega)$ *that the receiver approves the proposal in state* ω.
- *An outcome* α *is an* equilibrium outcome *if it corresponds to some equilibrium.*[6]

[6] Specifically, α is an equilibrium outcome if there exists an equilibrium (σ, a, q) such that $\forall \omega \in \Omega$, $\alpha(\omega) = \sum_{m \in \mathcal{M}} \sigma(m \mid \omega)$, where $\mathcal{M} := \{m \subseteq \Omega \mid a(m) = 1\}$ is the set of messages that convince the receiver to approve.

- *An outcome* α^c *is a* commitment outcome *if it solves*[7]

$$\max_{\alpha} \int_{\Omega} \alpha dP, \quad \textit{subject to} \quad \begin{aligned} &\forall \omega \in \Omega,\, 0 \le \alpha(\omega) \le 1, \\ &\int_{\Omega} \alpha \delta dP \ge 0. \end{aligned} \tag{1}$$

Some outcomes are deterministic, meaning that in every state ω the receiver either approves or rejects the proposal with certainty. For such outcomes, we can partition Ω into the states that the receiver approves and those that she rejects.

DEFINITION 3.

- *An outcome* α *is* deterministic *if* $\alpha(\omega) \in \{0,1\}$ *for every* $\omega \in \Omega$.
- *The* set of approved states W *in a deterministic outcome* α *is* $W := \{\omega \in \Omega \mid \alpha(\omega) = 1\}$.

3. ANALYSIS

DIRECT IMPLEMENTATION

Consider a deterministic equilibrium outcome with a set of approved states W. Suppose that the sender learns that $\omega \in A$. One message that is available to the sender in this state (and unavailable in every other state) is $\{\omega\}$. Since that message is verifiable, upon receiving it, the receiver learns with certainty that the state is ω, and, since ω is in the receiver's complete-information approval set, she approves the proposal. Thus, in every equilibrium, the receiver should be approving every $\omega \in A$; otherwise, the sender has a profitable deviation toward full disclosure. This observation gives rise to the sender's incentive-compatibility (IC) constraint

$$A \subseteq W. \tag{IC}$$

[7] Under commitment, the model in this paper is a version of the model of Alonso and Câmara (2016) with one receiver and continuous state space. According to Kamenica and Gentzkow (2011), the optimal straightforward experiment is supported on the set $\{s^+, s^-\}$, where s^+ induces the posterior q^+ and leads to a recommendation of approval, while s^- induces the posterior q^- and leads to a recommendation of rejection. The outcome then takes the form $\alpha^c(\omega) = \text{prob}(s^+ \mid \omega)$.

Next, if the receiver approves every state in W, then she expects that on average, her net payoff from approval is non-negative. Thus we obtain the receiver's obedience constraint

$$\mathbb{E}_p[\delta(\omega) \mid W] \geq 0. \tag{obedience}$$

The first result of this paper allows us to restrict attention to deterministic outcomes with sets of approved states $W \subseteq \Omega$ that satisfy these two constraints.

THEOREM 1. *Every equilibrium outcome is deterministic. Furthermore, $W \subseteq \Omega$ is an equilibrium set of approved states if and only if it satisfies the sender's* (IC) *constraint and the receiver's* (obedience) *constraint.*

The proofs of Theorem 1 and the other results are in the appendix. Here I describe the intuition behind Theorem 1. The first part of the theorem says that every equilibrium outcome is deterministic: the receiver either approves or rejects the proposal with probability one in every state of the world. To verify this, suppose instead that in some state, the receiver mixes between approval and rejection. Since the receiver sometimes approves, the sender has access to at least one message that convinces the receiver to approve. But then the sender can deviate and send that message with certainty so that the receiver approves with probability one. Hence, every equilibrium outcome is deterministic.

Next, if W is an equilibrium set of approved states, it satisfies the sender's (IC) constraint; otherwise the sender could deviate to full disclosure. To see why W also satisfies the receiver's (obedience) constraint, we implement this set of approved states directly. Specifically, let the sender send the message W if $\omega \in W$ and the message $\Omega \setminus W$ if $\omega \notin W$. Intuitively, the (obedience) constraint states that the receiver interprets the message W as a recommendation to approve. If the sender induces approval in every state in W in the original equilibrium, he also induces approval with the pooling message W.

Finally, suppose that $W \subseteq \Omega$ satisfies (IC) and (obedience). Then we can construct an equilibrium that directly implements the set of approved states W. Let the sender send the message W for every state within W and the message $\Omega \setminus W$ for every state outside of W. Then, by the (obedience) constraint, the receiver interprets the message W as a recommendation to approve. Off the equilibrium path, let the receiver be "skeptical" and assume that any unexpected message comes from the worst possible state. Then the sender has no profitable deviations: if $\omega \in W$, he is getting the highest possible payoff; if $\omega \notin W$, then he cannot replicate the message W, and the receiver rejects after every other message.

Note that Theorem 1 is a version of the communication revelation principle for games with verifiable information. According to Myerson (1986) and Forges (1986), any equilibrium outcome of a mediated sender–receiver game may be implemented truthfully and obediently. In the present context, this means that *(i)* the sender truthfully reveals the state of the world to the mediator, *(ii)* the mediator translates this report into an action recommendation for the receiver, and *(iii)* the receiver obediently follows her recommendation. Which equilibrium outcome is to be implemented is decided by the mediator at Step *(ii)*. Conveniently, Theorem 1 also provides necessary and sufficient conditions for a set of approved states to be implementable in equilibrium.

Equilibrium Range and Value of Commitment

For the purposes of characterizing equilibrium outcomes, Theorem 1 allows us to restrict attention to sets $W \subseteq \Omega$ satisfying (IC) and (obedience). I rank equilibria in terms of the sender's ex-ante utility, which is the same as his ex-ante odds of approval and equals $P(W)$, the prior measure of the set of approved states.

In the *sender-worst equilibrium*, the set of approved states $\underline{W}$ minimizes the sender's ex-ante utility across all equilibria. Thus, the (IC) constraint binds and $\underline{W} = A$. In this equilibrium, the receiver approves the proposal if and only if she approves it under complete information. Hence, the sender-worst equilibrium is outcome-equivalent to *full disclosure* (also known as *full unraveling*), which is salient in the verifiable-information literature.[8]

In the *sender-preferred equilibrium*, the set of approved states $\overline{W}$ maximizes the sender's ex-ante utility across all equilibria. Mathematically,

$$\overline{W} = \arg\max_{W \subseteq \Omega} P(W), \quad \text{subject to} \quad \begin{aligned} A &\subseteq W, \\ \mathbb{E}_p[\delta(\omega) \mid W] &\geq 0. \end{aligned}$$

[8]See, e.g., Milgrom (1981), Grossman (1981), Milgrom and Roberts (1986), and the review by Milgrom (2008).

To find the sender-preferred equilibrium, we would increase the ex-ante measure of the set of approved states W so long as the receiver, when approving, expects that her net payoff from approval is non-negative, on average. Because the state space is continuous, $\overline{W}$ makes the receiver exactly indifferent between approval and rejection, which means that her (obedience) constraint is binding.

THEOREM 2. *The sender-preferred set of approved states $\overline{W}$ is characterized by a cutoff value $c^* > 0$ such that*

- *the receiver almost surely approves the proposal if $\delta(\omega) > -c^*$ and rejects it if $\delta(\omega) < -c^*$;*[9]
- *whenever the receiver approves the proposal, her expected net payoff from approval is zero: $\mathbb{E}_p\left[\delta(\omega) \mid \overline{W}\right] = 0$.*

Furthermore, the sender-preferred equilibrium outcome is a commitment outcome.

First, notice that the receiver's (obedience) constraint binds; otherwise we could increase the value of the objective while still satisfying that constraint. I prove the first part of Theorem 2 by contradiction. Suppose that the sender-preferred set of approved states $\overline{W}$ is not characterized by a cutoff value for the receiver's net payoff from approval. Then there exist two sets $X, Y \subseteq \Omega$, of positive and equal measure, such that $\overline{W}$ includes X, $\overline{W}$ does not include Y, yet the receiver's net payoff from approval is higher for any state in Y than for any state in X. Consider an alternative set of approved states W^* formed by replacing X with Y, i.e. $W^* = (\overline{W} \setminus X) \cup Y$. The sender has the same ex-ante payoff at W^* and $\overline{W}$, because the sets X and Y have the same measure. But the (obedience) constraint for W^* is loose, while for $\overline{W}$ it is binding. This is because every state in Y is "cheaper" in terms of the constraint than each state in X. Thus, we can improve upon both $\overline{W}$ and W^*, which is a contradiction.

Next, let us compare the problems of *(i)* finding the sender-preferred equilibrium outcome and *(ii)* finding the commitment outcome. In *(i)*, we maximize the ex-ante measure of the set of approved states subject to the (IC) and (obedience) constraints. In *(ii)*, the sender maximizes his ex-ante utility subject to an obedience-like constraint on the re-

[9] Almost surely with respect to the prior distribution p.

ceiver. Crucially, under commitment, the sender does not face an IC constraint. Also, a commitment outcome may not be deterministic.

A commitment outcome is characterized by a cutoff value for the receiver's net payoff of approval, for the same reason $\overline{W}$ is.[10] That is, the receiver certainly approves (rejects) the states with a net payoff from approval above (below) some threshold. Furthermore, that threshold is negative, since the receiver certainly approves every state in her complete-information approval set. Hence, any commitment outcome satisfies the sender's IC constraint.

In a non-deterministic commitment outcome, there is some non-empty set $\mathcal{D} \subseteq \Omega$ on which the sender induces both actions of the receiver with positive probability. Since any commitment outcome is characterized by a cutoff value, the receiver's net payoff from approval must be the same for every state in $\mathcal{D}$. We can therefore transform the non-deterministic outcome into a deterministic outcome by partitioning $\mathcal{D}$ into two subsets and letting the sender recommend one action on each subset with certainty. Thanks to the continuity of the state space, such partitioning does not affect the objective function or the (obedience) constraint of the receiver. Consequently, this is a valid deterministic commitment outcome. Since this commitment outcome satisfies the sender's IC constraint, it is an equilibrium outcome.

EXAMPLE 2 (THE PROSECUTOR AND THE JUDGE). Consider the setting from Example 1, and suppose that the judge's threshold for conviction is $v = 0.7$.

In the sender-worst equilibrium, the set of approved states is $\underline{W} = [0.7, 1]$, which coincides with the judge's complete-information approval set. That is, the judge convicts if and only if the defendant is at least 70% guilty, which is what she would do under complete information.

In the sender-preferred equilibrium, the prosecutor maximizes the odds of conviction subject to the judge's (obedience) constraint, which states that, given the message $\overline{W}$, the judge's average net payoff from approval is non-negative. Recall that the judge's net payoff from approval equals 1 if the defendant is guilty (i.e., if $\omega \in [0.7, 1]$) and -1 if

[10] Alonso and Câmara (2016) prove that if the state space is finite, then the solution under commitment features a cutoff state.

the defendant is innocent (i.e., if $\omega \in [0, 0.7)$). Thus, to maximize the odds of conviction, the prosecutor pools the guilty defendants with as many innocent ones as possible, while making sure that on average across this pool, the judge still wants to convict. Mathematically, the prosecutor selects $\beta \in [0, 0.7)$ to solve $\int_{0.7}^{1} dP + \int_{\beta}^{0.7} dP = 0$. For the uniform prior, we get $\beta = 0.4$. Figure 3 illustrates the sender-preferred equilibrium.

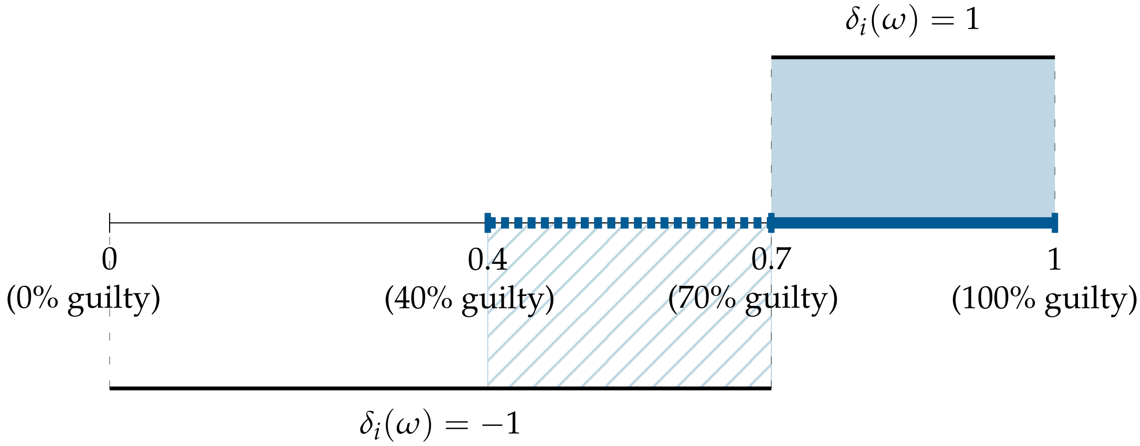

Figure 3. *The sender-preferred set of approved states $\overline{W} = [0.4, 1]$ when the prior is $U[0, 1]$. The receiver's expected net payoff from approval after receiving the message $[0.4, 1]$ is zero, because the area under $\delta(\omega)$ taken over the judge's complete-information approval set (solid blue) is equal to the area above $\delta(\omega)$ outside of the judge's complete-information approval set (dashed blue).*

As Kamenica and Gentzkow (2011) remark of their example, "This leads the judge to convict with probability 60 percent. Note that the judge knows 70 percent of defendants are innocent, yet she convicts 60 percent of them!"[11] In the setup of this paper, the prosecutor reaches the same outcome without having to commit to an experiment. He does so by using an equilibrium strategy of saying "The defendant is guilty" whenever the defendant is at least 40% guilty, and "The defendant is innocent" otherwise. All he needs is sufficient evidence of his claim that the defendant is indeed at least 40% guilty.

4. Robustness

Here I discuss how the two main results of this paper, the *recommendation principle* of Theorem 1 and the *no benefit from commitment* result of Theorem 2, generalize to more

[11] In the sender-preferred equilibrium, every defendant in $[0.4, 1]$ is convicted, which is 60% of all defendants. Since the judge considers the defendants in $[0, 0.7]$ innocent, 70% of the defendants are innocent ex ante.

complex environments. Specifically, I extend the model in two directions. Firstly, I show that both results hold in the model with many (independent) receivers. Secondly, I show that if a receiver is choosing among three or more actions, the recommendation principle still holds, but the sender-preferred equilibrium outcome may or may not reach the commitment payoff.

Many Receivers

Suppose there is a set $I = \{1, \dots, n\}$, $n \geq 2$, of receivers. Each receiver's payoff depends only on the state ω and her own action. More specifically, receiver $i \in I$ has some net payoff from approval $\delta_i(\omega)$ that defines her complete-information approval set A_i. The sender's preferences are state-independent and weakly monotone in each receiver's action. (For example, the sender could be a prosecutor and the receivers jurors, with the prosecutor needing to convince at least half of the jurors to reach the desired verdict.) Assume that the sender communicates with each receiver in private: he chooses $m_i \in M$ such that $\omega \in m_i$ for every $i \in I$, and receiver i observes only her own message m_i.

As in the one-receiver case, we can define an *outcome* as a set of approved states $W_i \subseteq \Omega$ for each receiver $i \in I$. Since the receivers are independent, we can claim that *in equilibrium*, receiver $i \in I$ approves W_i if and only if it satisfies her obedience constraint. At the same time, since the sender's preferences are monotone in every receiver's action, W_i should also satisfy the sender's IC constraint for each $i \in I$. We thus obtain a generalization of the recommendation principle that accommodates multiple receivers. Using this, we can characterize the set of equilibrium payoffs of the sender. In his least preferred equilibrium, the sender reveals the state of the world to each receiver, i.e. $\underline{W}_i = A_i$ for every $i \in I$. In his most preferred equilibrium, the sender chooses $(\overline{W}_1, \dots, \overline{W}_n)$ so as to maximize his objective subject to each receiver's obedience constraint. In that equilibrium, we see by the same argument as in the one-receiver case that the sender reaches the full-commitment payoff.

One Receiver Choosing among 3+ Actions

Suppose now that there is one receiver, who chooses her action from a set $J = \{0, 1, \dots, m\}$ with $m \geq 2$. Define the receiver's *complete-information approval set for action j* as the set A_j

consisting of all states of the world in which she prefers to take action j when she is fully informed. Also suppose that the sender's preferences are state-independent, and his payoff is increasing in the receiver's action.

An outcome is now a partition of the state space into $m+1$ subsets, $(W_0, W_1, \ldots, W_m)$, where $W_j \subseteq \Omega$ consists of the states in which the receiver takes action $j \in J$. The receiver's obedience constraint states that she plays action j in state $\omega \in W_j$, because that action maximizes her utility given her posterior belief $p(\cdot \mid W_j)$. Next, the sender's IC constraint states that he cannot profitably deviate to full disclosure; that is, if $\omega \in A_j$, then $\omega \in W_j \cup \cdots \cup W_m$. As a result, in environments in which the receiver has multiple actions, the recommendation principle generalizes as follows: a partition $(W_0, W_1, \ldots, W_m)$ of the state space is an equilibrium outcome if and only if, for every action $j \in J$, W_j satisfies the IC and obedience constraints.

As it turns out, when $m \geq 2$, the sender may or may not reach the commitment payoff in his most preferred equilibrium. Intuitively, in Bayesian persuasion, the sender can commit to a signal that recommends an intermediate action most of the time. In equilibrium, such a solution may violate the sender's IC constraint for the highest action. The determination of necessary and sufficient conditions for the equivalence of payoffs remains an open problem which I leave for further research. However, once a commitment outcome is known, it is easy to check whether it is implementable in equilibrium. The answer is affirmative if the outcome satisfies the sender's IC constraint for every action.

Consider the following example from Gentzkow and Kamenica (2016), in which the receiver has three actions available. For each action $a \in \{0, 1, 2\}$, the receiver's utility is $u(a, \omega) = 3a\omega - a^2$; her complete-information approval sets are $A_0 = [0, \frac{1}{3})$, $A_1 = [\frac{1}{3}, \frac{2}{3})$, $A_2 = [\frac{2}{3}, 1]$. Given the belief $q \in \Delta\Omega$, the receiver prefers to take action 2 if $\mathbb{E}_q[x] \geq \frac{2}{3}$, action 1 if $\mathbb{E}_q[\omega] \in \left[\frac{1}{3}, \frac{2}{3}\right)$, and action 0 otherwise. The sender's payoff is given by $u_s(0) = 0$, $u_s(1) = 1$, and $u_s(2) = 3$. The prior is uniform. The authors find that one way to reach the commitment payoff is by inducing action 0 on $[0, \frac{8}{48}]$, action 1 on $\left[\frac{11}{48}, \frac{21}{48}\right]$, and action 2 on $\left[\frac{8}{48}, \frac{11}{48}\right] \cup \left[\frac{21}{48}, 1\right]$. Clearly, in this commitment outcome the sender does not have profitable deviations toward full disclosure: if $\omega \in A_2$, then the receiver is playing a_2; if $\omega \in A_1$, then the receiver is playing a_1 or a_2. Thus, we can conclude that this commitment outcome is also an equilibrium outcome.

5. CONCLUSION

This paper studies how an informed sender with state-independent preferences can use verifiable information to persuade a receiver to approve his proposal. I find that the equilibrium outcomes can be characterized as the sets of approved states that satisfy the receiver's obedience constraint and the sender's incentive-compatibility constraint. In the sender-worst equilibrium, information unravels, and the receiver acts as if fully informed. The sender-preferred equilibrium is the commitment outcome of the Bayesian persuasion game. Consequently, when the sender's preferences are state-independent and the receiver is choosing between two actions, the sender reaches the full-commitment payoff using only evidence, with no need for commitment power, as long as the message space is sufficiently rich.

REFERENCES

ALONSO, RICARDO and ODILON CÂMARA (2016), "Persuading Voters", *American Economic Review*, 106, 11 (Nov. 2016), pp. 3590-3605. (pp. 6, 9, 13.)

BARDHI, ARJADA and YINGNI GUO (2018), "Modes of Persuasion toward Unanimous Consent", *Theoretical Economics*, 13, 3, pp. 1111-1149. (p. 6.)

BOLESLAVSKY, RAPHAEL and CHRISTOPHER COTTON (2015), "Grading Standards and Education Quality", *American Economic Journal: Microeconomics*, 7, 2 (May 2015), pp. 248-279. (p. 6.)

BULL, JESSE and JOEL WATSON (2004), "Evidence Disclosure and Verifiability", *Journal of Economic Theory*, 118, 1 (Sept. 2004), pp. 1-31. (p. 6.)

CRAWFORD, VINCENT P. and JOEL SOBEL (1982), "Strategic Information Transmission", *Econometrica*, 50, 6 (Nov. 1982), p. 1431. (p. 5.)

EGOROV, GEORGY and KONSTANTIN SONIN (2019), "Persuasion on Networks", *SSRN Electronic Journal*. (p. 6.)

FORGES, FRANÇOISE (1986), "An Approach to Communication Equilibria", *Econometrica*, 54, 6 (Nov. 1986), p. 1375. (p. 11.)

GEHLBACH, SCOTT and KONSTANTIN SONIN (2014), "Government Control of the Media", *Journal of Public Economics*, 118 (Oct. 2014), pp. 163-171. (p. 6.)

GENTZKOW, MATTHEW and EMIR KAMENICA (2016), "A Rothschild-Stiglitz Approach to Bayesian Persuasion", *American Economic Review*, 106, 5 (May 2016), pp. 597-601. (p. 16.)

GROSSMAN, SANFORD J. (1981), "The Informational Role of Warranties and Private Disclosure about Product Quality", *The Journal of Law and Economics*, 24, 3 (Dec. 1981), pp. 461-483. (pp. 3, 5, 11.)

KAMENICA, EMIR and MATTHEW GENTZKOW (2011), "Bayesian Persuasion", *American Economic Review*, 101, 6 (Oct. 2011), pp. 2590-2615. (pp. 3-5, 7, 9, 14.)

KOLOTILIN, ANTON (2015), "Experimental Design to Persuade", *Games and Economic Behavior*, 90 (Mar. 2015), pp. 215-226. (p. 6.)

LIPNOWSKI, ELLIOT (2020), "Equivalence of Cheap Talk and Bayesian Persuasion in a Finite Continuous Model", *note*. (p. 6.)

LIPNOWSKI, ELLIOT and DORON RAVID (2020), "Cheap Talk With Transparent Motives", *Econometrica*, 88, 4, pp. 1631-1660. (pp. 5, 6.)

MILGROM, PAUL R. (1981), "Good News and Bad News: Representation Theorems and Applications", *The Bell Journal of Economics*, 12, 2, p. 380. (pp. 3, 5, 11.)

— (2008), "What the Seller Won't Tell You: Persuasion and Disclosure in Markets", *Journal of Economic Perspectives*, 22, 2 (Mar. 2008), pp. 115-131. (p. 11.)

MILGROM, PAUL R. and JOHN ROBERTS (1986), "Relying on the Information of Interested Parties", *The RAND Journal of Economics*, 17, 1, p. 18. (pp. 6, 11.)

MYERSON, ROGER B. (1986), "Multistage Games with Communication", *Econometrica*, 54, 2 (Mar. 1986), p. 323. (p. 11.)

OSTROVSKY, MICHAEL and MICHAEL SCHWARZ (2010), "Information Disclosure and Unraveling in Matching Markets", *American Economic Journal: Microeconomics*, 2, 2 (May 2010), pp. 34-63. (p. 6.)

ROMANYUK, GLEB and ALEX SMOLIN (2019), "Cream Skimming and Information Design in Matching Markets", *American Economic Journal: Microeconomics*, 11, 2 (May 2019), pp. 250-276. (p. 6.)

APPENDIX: OMITTED PROOFS

PROOF OF THEOREM 1

Part I: suppose, on the contrary, that there exists a non-deterministic equilibrium outcome α with $\alpha(\omega) \in (0,1)$ for some $\omega \in \Omega$. Then, $\alpha(\omega) > 0$ implies that there exists a convincing message m such that $\sigma(m \mid \omega) > 0$ and $\mathbb{E}_{q(\cdot \mid m)}[\delta(\omega)] \geq 0$. Then, the sender has a profitable deviation to $\widetilde{\sigma}(m \mid \omega) = 1$: his payoff in ω increases from $\alpha(\omega) < 1$ to 1.

Part II: consider equilibrium (σ, a, q) with the set of approved states W. W must satisfy the sender's (IC) constraint, or else the sender can deviate to full disclosure. Next, let $\mathcal{M} := \{m \in M \mid a(m) = 1\}$ be the set of messages that convince the receiver to approve. Notice that if the sender has access to a convincing message in state ω, then he has to convince the receiver with probability 1 in that state. Consequently, the sender must be sending only convincing messages from the states that the receiver eventually approves, or $\forall \omega \in W, \sum_{m \in \mathcal{M}} \sigma(m \mid \omega) = 1$.

Next, consider message $m \in \mathcal{M}$. By the equilibrium condition *(iv)*, $\operatorname{supp} q(\cdot \mid m) \subseteq m$. Furthermore, $m \subseteq W$ because if $\omega \in m$ such that $m \in \mathcal{M}$, then $\omega \in W$. Notice every message $m \in \mathcal{M}$ convinces the receiver to approve, i.e.

$$\int_W \delta(\omega) \frac{\sigma(m \mid \omega)}{\int_W \sigma(m \mid \omega') dP(\omega')} dP(\omega) \geq 0 \iff \int_W \delta(\omega) \sigma(m \mid \omega) dP(\omega) \geq 0.$$

Take the sum the above inequality over all $m \in \mathcal{M}$:

$$\sum_{m \in \mathcal{M}} \int_W \delta(\omega) \sigma(m \mid \omega) dP(\omega) = \int_W \delta(\omega) \underbrace{\sum_{m \in \mathcal{M}} \sigma(m \mid \omega)}_{=1,\, \forall \omega \in W} dP(\omega) \iff \mathbb{E}_p[\delta(\omega) \mid W] \geq 0,$$

meaning that W satisfies the (obedience) constraint, as well.

<u>*Part III*</u>: suppose that $W \subseteq \Omega$ satisfies (IC) and (obedience). Let the sender's strategy be $\sigma(W \mid \omega) = \mathbb{1}(\omega \in W)$ and $\sigma(W^c \mid \omega) = \mathbb{1}(\omega \in W^c)$, where $W^c := \Omega \setminus W$. On the path, receiver only hears two messages, W and W^c. By (obedience), she approves after message W because her expected net payoff from approval is non-negative. On the other hand, she rejects after message W^c because her net payoff from approval is negative for every $\omega \in W^c$. In words, the sender sends two messages and the receiver interprets them as a recommendation to approve or reject. Off-the-path, i.e. following any message $m \neq W, \Omega \setminus W$, let the receiver have "skeptical beliefs"

$$\forall m \subseteq A, \ \operatorname{supp} q(\cdot \mid m) \subseteq m, \text{so that } \mathbb{E}_q[\delta(\omega)] \geq 0,$$

$$\forall m \nsubseteq A, m \neq W, \ \operatorname{supp} q(\cdot \mid m) \subseteq m \setminus A, \text{so that } \mathbb{E}_q[\delta(\omega)] < 0$$

that assign positive probability to states within the complete-information approval set if and only if the message comprises of these states only. Then, the sender does not have profitable deviations. We have thus constructed an equilibrium with the set of approved states W.

Proof of Theorem 2

Claim 1 : $\overline{W}$ solves the relaxed problem

$$\max_{W \subseteq \Omega} P(W), \quad \text{subject to} \quad \int_W \delta dP \geq 0 \tag{2}$$

and binds the (obedience) constraint, i.e. $\int_{\overline{W}} \delta dP = 0$.

Firstly, since $\delta(\omega) \geq 0$ for every $\omega \in A$, the solution of the relaxed problem automatically satisfies the (IC) constraint. Secondly, if $\overline{W}$ does not bind the (obedience) constraint, one can increase the value of the objective function while satisfying that constraint.

Claim 2 : $\overline{W}$ is characterized by a cutoff value of δ.

If not, there exist $X, Y \subseteq \Omega$ such that *(i)* $P(X) = P(Y) > 0$; *(ii)* $\forall \omega \in X$, $\forall \omega' \in Y$, $\delta(\omega) < \delta(\omega')$; *(iii)* $X \subseteq \overline{W}$ and $Y \subseteq \Omega \setminus \overline{W}$. In words, the sender-preferred set of approved states includes the set X, does not include the set Y, yet the receiver has a higher net payoff of approving any state in Y over any state in X. Let $W^* := (\overline{W} \setminus X) \cup Y$. The value of the objective function is the same for $\overline{W}$ and W^*:

$$P(\overline{W}) = P(\overline{W} \setminus X) + P(X) = P(\overline{W} \setminus X) + P(Y) = P(W^*).$$

The obedience constraint for $\overline{W}$ is $\int_{\overline{W} \setminus X} \delta dP + \int_X \delta dP = 0$. The obedience constraint for W^* is $\int_{W^* \setminus Y} \delta dP + \int_Y \delta dP > 0$. The last inequality follows from *(1)* $W^* \setminus Y = \overline{W} \setminus X$, so the first term in both constraints is the same, and *(2)* $\int_X \delta dP < \int_Y \delta dP$, so the second term in the second constraint is strictly larger. Thus, $\overline{W}$ and W^* reach the same ex-ante payoff of the sender, but the (obedience) constraint for W^* is loose. From Claim 1, the (obedience) constraint binds at the optimum. Hence, both $\overline{W}$ and W^* can be improved upon.

Claim 3 : any commitment outcome α^c is characterized by a cutoff value $c^{BP} > 0$, such that

$$\begin{aligned} \alpha^c(\omega) &= 1 && \text{if } \delta(\omega) > -c^{BP}, \\ \alpha^c(\omega) &\in [0,1], && \text{if } \delta(\omega) = -c^{BP}, \\ \alpha^c(\omega) &= 0, && \text{if } \delta(\omega) < -c^{BP}. \end{aligned}$$

α^c is characterized by a cutoff value for the same reason why $\overline{W}$ is. If it was not, then there exist $X, Y \subseteq \Omega$ such that

- $\int\limits_X \alpha^c dP = \int\limits_Y (1-\alpha^c)dP$;
- $\forall \omega \in X$, $\forall \omega' \in Y$, $\delta(\omega) < \delta(\omega')$;
- $\forall \omega \in X$, $\alpha^c(\omega) > 0$ and $\forall \omega \in Y$, $\alpha^c(\omega) < 1$.

Then, letting $\alpha^*(\omega) = \alpha^c(\omega)$ for all $\omega \notin X \cup Y$, $\alpha^*(\omega) = 1$ if $\omega \in Y$, $\alpha^*(\omega) = 0$ if $\omega \in X$ leads to the same level of the objective function and a looser constraint.

Claim 4 : there exists a deterministic commitment outcome with set of approved states $\overline{W}$.

Consider a commitment outcome α^c, and let $\mathcal{D} := \{\omega \in \Omega \mid 0 < \alpha^c(\omega) < 1\}$ be the set of states the receiver approves and rejects with a positive probability. Since α^c is characterized by the cutoff value c^{BP}, for every $\omega \in \mathcal{D}$, $\delta(\omega) = -c^{BP}$.

Next, let $\widetilde{\alpha}(\omega) = \alpha^c(\omega)$ for all $\omega \notin \mathcal{D}$ and $\widetilde{\alpha}(\omega) = \mathbb{1}(\omega \in X)$ for all $\omega \in \mathcal{D}$, where $X \subseteq \mathcal{D}$ solves

$$\int\limits_{\mathcal{D}} \alpha^c(\omega)dP(\omega) = \int\limits_{\mathcal{D}} \widetilde{\alpha}(\omega)dP(\omega) = P(X).$$

Now compare the commitment outcome α^c and the candidate outcome $\widetilde{\alpha}$, keeping in mind that they only differ on $\mathcal{D}$. The value of the sender's objective function is the same:

$$\int\limits_{\mathcal{D}} \alpha dP = \int\limits_{\mathcal{D}} \widetilde{\alpha} dP = P(X);$$

the constraint is also the same:

$$\int\limits_{\mathcal{D}} \underbrace{\delta(\omega)}_{=-c^{BP}, \forall \omega \in \mathcal{D}} \alpha^c(\omega)dP(\omega) = -c^{BP} \cdot \int\limits_{\mathcal{D}} \widetilde{\alpha}(\omega)dP(\omega) = -c^{BP} \cdot P(X).$$

Consequently, $\widetilde{\alpha}(\omega) = \mathbb{1}(\omega \in \mathcal{D}_1 \cup X)$ is a *deterministic* commitment outcome. Finally, notice that any deterministic commitment outcome also solves the relaxed problem (2). Hence, there exists a deterministic commitment outcome with the set of approved states $\overline{W}$.