

# ON INFORMATION DESIGN IN GAMES

Laurent Mathevet  
*New York University*

Jacopo Perego  
*Columbia University*

Ina Taneva  
*University of Edinburgh*

February 6, 2020

## Abstract

Information provision in games influences behavior by affecting agents' beliefs about the state, as well as their higher-order beliefs. We first characterize the extent to which a designer can manipulate agents' beliefs by disclosing information. We then describe the structure of optimal belief distributions, including a concave-envelope representation that subsumes the single-agent result of [Kamenica and Gentzkow \(2011\)](#). This result holds under various solution concepts and outcome selection rules. Finally, we use our approach to compute an optimal information structure in an investment game under adversarial equilibrium selection.

*Keywords:* information design, belief manipulation, belief distributions, extremal decomposition, concavification.

*JEL Classification:* C72, D82, D83.

---

We are grateful to the editor, Emir Kamenica, and to three anonymous referees for their comments and suggestions, which significantly improved the paper. Special thanks are due to Anne-Katrin Roesler for her insightful comments as a discussant at the 2016 Cowles Foundation conference, and to Philippe Jehiel, Elliot Lipnowski, Stephen Morris, David Pearce, József Sákovics, and Siyang Xiong for their helpful comments and conversations. We are also grateful to Andrew Clausen, Olivier Compte, Jon Eguia, Willemien Kets, Matt Jackson, Qingmin Liu, Efe Ok, Ennio Stacchetti, and Max Stinchcombe, as well as to seminar audiences at Columbia University, Institut d'Anàlisi Econòmica, New York University, Michigan State University, Paris School of Economics, Stony Brook University, University of Cambridge, University of Edinburgh, UT Austin, the SIRE BIC Workshop, the 2016 Decentralization conference, the 2016 Canadian Economic Theory conference, the 2016 Cowles Foundation conference, the 2016 North American Summer Meeting of the Econometric Society, and SAET 2017. Taneva acknowledges financial support from ESRC grant ES/N00776X/1.

E-mail: [lmath@nyu.edu](mailto:lmath@nyu.edu) and [jacopo.perego@columbia.edu](mailto:jacopo.perego@columbia.edu) and [ina.taneva@ed.ac.uk](mailto:ina.taneva@ed.ac.uk)

# 1. Introduction

Monetary incentives, such as taxes, fines, wages, and insurance are ways of manipulating agents' payoffs to incentivize a range of behaviors, from exerting effort to risk-taking. In incomplete information environments, strategic transmission of information may also be used as a tool to affect agents' behavior, in this case by manipulating their beliefs. Information design analyzes the latter, in a setting where a designer commits to disclosing information to a group of interacting agents.

Under incomplete information, agents' behavior depends, in part, on their beliefs about the uncertain state of the world. For example, an investor's belief about the quality of a new technology influences his decision whether to invest or not in the startup that launches it. However, his decision also depends on how likely he thinks other investors are to fund the startup, which in turn depends on their own beliefs about the state and other investors' decisions. Thus, an agent's beliefs about the other agents' beliefs about the state also affect his decision, as do his beliefs about their beliefs about his beliefs about the state, and so on. These higher-order beliefs are absent from the single-agent environment, but they are an important part of the information design problem with multiple interacting agents.

This paper contributes to the foundations of information design in three ways. First, we characterize the feasible distributions of agents' beliefs that a designer can induce through the choice of information structure. Information design is ultimately an exercise in belief manipulation, whether it is explicitly modeled as such or solved by way of incentive compatible distributions over actions and states. However, an information designer cannot induce just any belief distribution she wishes to. In the single-agent case, for example, the designer is constrained to distributions over beliefs about the state that on average equal the prior, a condition known as Bayes plausibility (Kamenica and Gentzkow (2011)). In the multi-agent case, an additional requirement emerges: agents' beliefs should be consistent with one another. We further establish an equivalence between (Bayes plausible and consistent) distributions of agents' beliefs and distributions of the *designer's* beliefs, which is particularly useful in applications.

Second, we represent the designer's problem in a way that exploits the structure of consistent belief distributions. We show that every consistent belief distribution can be represented as a (convex) combination of more "basic" elements, themselves belief distributions. Therefore, the problem of choosing an optimal information structure is equivalent to choosing an optimal combination of such basic elements, subject to Bayes plausibility. From this follows a two-step approach to the information design problem: the first step optimizes among the basic elements only, and

the second step optimally combines the elements selected in the first step subject to Bayes plausibility. The latter step corresponds to the concavification of a value function derived in the first step. Therefore, this representation subsumes the single-agent result of [Kamenica and Gentzkow \(2011\)](#). This two-step decomposition can be interpreted as optimizing over private information in the first step and then adding an optimal public signal in the second step.

Third, the above results apply to a variety of solution concepts and equilibrium selection rules. The choice of solution concept can address many problems in information design, in much the same way that it does in mechanism design. For example, a designer may be concerned that agents only have bounded depths of reasoning, that they can deviate in coalitions, or that they can communicate. A designer may also want to hedge against the possibility that, when there are multiple outcomes (consistent with the solution concept), agents might coordinate on the outcome most unfavorable to her. This can be achieved by choosing a robust information structure, which maximizes the designer’s payoff under an adversarial selection rule. Current methods have focused on Bayes Nash equilibrium (BNE) as a solution concept and designer-preferred equilibrium selection.

We apply our approach to an investment game where the coordination motive is a source of multiple equilibria under incomplete information, and the designer would like to maximize investment. In this problem, as in other similar ones, the possibility that agents coordinate on the equilibrium least preferred by the designer is a serious issue. In response, the designer may want to choose the information structure that maximizes investment probabilities in the worst equilibrium. Information design under such adversarial equilibrium selection is outside the scope of existing methods. We use our approach to compute the optimal information structure which takes on a simple form: every agent either receives a private message that makes investing uniquely optimal due to a combination of first- and higher-order beliefs, or receives a public message that makes it common knowledge that the state is low, hence not investing is uniquely optimal. The private messages create contagion à la [Rubinstein \(1989\)](#), which we refer to as the bandwagon effect: one message induces a first-order belief that is high enough to incentivize investment by itself, while all other messages aim to cause investment by an induction argument that uses beliefs of incrementally higher order.

The single-agent problem has been a rich subject of study since the influential work of [Kamenica and Gentzkow \(2011\)](#). The standard problem is mathematically analogous to [Aumann and Maschler \(1995\)](#), which studies a repeated game with an informed player who exploits his knowledge against an uninformed opponent.<sup>1</sup>

---

<sup>1</sup>Other early contributions to single-agent information design include [Brocas and Carrillo \(2007\)](#)

By contrast, the theory of information design in games is not as well understood. Bergemann and Morris (2016) and Taneva (2019) formulate the Myersonian approach to Bayes Nash information design. This approach is based on a notion of correlated equilibrium under incomplete information, Bayes correlated equilibrium (BCE), which characterizes all possible BNE outcomes that could arise under all information structures. The BCE approach elegantly avoids the explicit modeling of belief hierarchies, and proves useful for solving information design problems by means of linear programming. However, it fundamentally relies on BNE as a solution concept and on selecting the designer-preferred equilibrium in case of multiplicity. In contrast, our results develop the belief-based approach to information design, which can be viewed as the multi-agent analogue to Kamenica and Gentzkow (2011)’s single-agent formulation. Earlier works have studied the effects of public and private information on equilibrium behavior, efficiency, and welfare (e.g., Vives (1988), Morris and Shin (2002) and Angeletos and Pavan (2007)). More recent papers study the optimal design of information in voting games (Alonso and Câmara (2016), Chan et al. (2019)); dynamic bank runs (Ely (2017)); stress testing (Inostroza and Pavan (2017)); auctions (Bergemann, Brooks, and Morris (2017)); contests (Zhang and Zhou (2016)); or focus on public information in games (Laclau and Renou (2016)).

## 2. The Information Design Problem

A set  $N = \{1, \dots, n\}$  of agents interact in an uncertain environment. The variable  $\theta \in \Theta$  describes the uncertain state of the world, where the set  $\Theta$  is finite. Every agent  $i \in N$  has a finite action set  $A_i$  and utility function  $u_i : A \times \Theta \rightarrow \mathbb{R}$ , where  $A = \prod_i A_i$ . A priori, the agents only know that  $\theta$  is distributed according to  $\mu_0 \in \Delta\Theta$ , which is common knowledge. We refer to  $G = (\Theta, \mu_0, N, \{A_i\}, \{u_i\})$  as the **base game**.

A designer commits to disclosing information to the agents about the payoff relevant state  $\theta$ . This is modeled by an **information structure**  $(S, \pi)$ , where  $S_i \subseteq \mathcal{S}$  is the finite set of messages that agent  $i$  can receive,  $S = \prod_i S_i$  is the set of message profiles, and  $\pi : \Theta \rightarrow \Delta(S)$  is the information map.<sup>2</sup> In any state  $\theta$ , the message profile  $s = (s_i)$  is drawn according to  $\pi(s|\theta)$  and agent  $i$  privately observes  $s_i$ . An information structure can be thought of as an experiment concerning the state, such as an audit, a stress test or a medical test. As is standard in information

---

and Rayo and Segal (2010) proposed models of single-agent information design, with binary state space and sequential binary signals for the former, and optimal advertising about product attributes for the latter.

<sup>2</sup>We restrict attention to finite message spaces, because it guarantees existence of a BNE or all information structures. Any infinite set  $\mathcal{S}$  would serve the purpose.

design, this model assumes that the designer commits to the information structure at a time when she does not know the realized state, but only knows its distribution  $\mu_0$ .<sup>3</sup> The information designer's preferences are summarized by the payoff function  $v : A \times \Theta \rightarrow \mathbb{R}$ . Her objective is to maximize her expected payoff through the choice of information structure.

The combination of information structure and base game,  $\mathcal{G} = \langle G, (S, \pi) \rangle$ , constitutes a **Bayesian game** in which agents play according to a **solution concept**  $\Sigma(\mathcal{G}) \subseteq \{\sigma = (\sigma_i) \mid \sigma_i : S_i \rightarrow \Delta A_i \text{ for all } i\}$ . The resulting outcomes are distributions over action profiles and states, represented by

$$O_\Sigma(\mathcal{G}) = \{\gamma \in \Delta(A \times \Theta) : \text{there exists } \sigma \in \Sigma(\mathcal{G}) \text{ such that} \\ \gamma(a, \theta) = \sum_s \sigma(a|s)\pi(s|\theta)\mu_0(\theta) \text{ for all } (a, \theta)\}.$$

Assume that  $O_\Sigma$  is non-empty and compact-valued. Given that  $G$  is finite, this holds when  $\Sigma$  is BNE, for example. For a fixed base game, we just write  $O_\Sigma(S, \pi)$ . When  $O_\Sigma$  contains multiple outcomes, the designer expects that *one* of them will happen, which is described by a **selection rule**  $g : D \subseteq \Delta(\Theta \times A) \mapsto g(D) \in D$ . The worst and the best outcomes are natural selection criteria. A pessimistic designer, or one interested in robust information design, expects the worst outcome:

$$g(D) \in \operatorname{argmin}_{\gamma \in D} \sum_{a, \theta} \gamma(a, \theta) v(a, \theta) \quad (1)$$

for all compact  $D \subseteq \Delta(\Theta \times A)$ . An optimistic designer would instead expect the best equilibrium, with  $\operatorname{argmax}$  instead of  $\operatorname{argmin}$  in (1). Other criteria, such as random choice rules, could also be considered. Letting  $g^{(S, \pi)} := g(O_\Sigma(S, \pi))$ , the designer's ex-ante expected payoff is given by

$$V(S, \pi) := \sum_{a, \theta} g^{(S, \pi)}(a, \theta) v(a, \theta). \quad (2)$$

Finally, the information design problem is  $\sup_{(S, \pi)} V(S, \pi)$ .

---

<sup>3</sup>While commitment can be a strong assumption in some situations, it holds implicitly in repeated environments wherein a sender makes announcements periodically and wants to be trusted in the future (Best and Quigley (2018) and Mathevet, Pearce, and Stacchetti (2018)).

### 3. Information Design as Belief Manipulation

We reformulate information design into a belief manipulation problem, analogously to [Kamenica and Gentzkow \(2011\)](#)'s approach to the single-agent case. Choosing an information structure is equivalent to choosing a distribution over belief hierarchies. However, only a special class of belief (hierarchy) distributions can be induced by information structures. [Kamenica and Gentzkow \(2011\)](#) established that choosing an information structure is equivalent to choosing a *Bayes plausible* distribution over first-order beliefs. We show that a similar equivalence holds in games, provided that agents' beliefs are, in addition, consistent with each other.

#### 3.1. Belief Distributions

A **belief hierarchy**  $t_i$  for agent  $i$  is an infinite sequence  $(t_i^1, t_i^2, \dots)$  whose components are coherent<sup>4</sup> beliefs of all orders:  $t_i^1 \in \Delta\Theta$  is  $i$ 's first-order belief;  $t_i^2 \in \Delta(\Theta \times (\Delta\Theta)^{n-1})$  is  $i$ 's second-order belief (i.e., a belief about  $\theta$  and every  $j$ 's first-order beliefs); and so on. Even if the belief hierarchies are coherent, they may assign positive probability to other agents' belief hierarchies that are not coherent. [Brandenburger and Dekel \(1993\)](#) show that we can construct a set of coherent belief hierarchies  $T_i$  for every  $i$  such that there exists a homeomorphism  $\beta_i^* : T_i \rightarrow \Delta(\Theta \times T_{-i})$  for all  $i$ .<sup>5</sup> This map describes  $i$ 's beliefs about  $(\theta, t_{-i})$  given  $t_i$ , and shows that there are sets of coherent belief hierarchies for all agents that only put positive probabilities on each other, making coherency common knowledge. Let  $T := \prod_i T_i$  be the space of hierarchy profiles with common knowledge of coherency.

Given  $\mu_0$  and an information structure  $(S, \pi)$ , and upon receiving a message  $s_i$ , agent  $i$  formulates beliefs  $\mu_i(s_i) \in \Delta(\Theta \times S_{-i})$  about the state and other agents' messages in  $S_{-i} := \prod_{j \neq i} S_j$  by using Bayes' rule. In particular,  $\mu_i^1(s_i) := \text{marg}_{\Theta} \mu_i(s_i)$  describes  $i$ 's belief about the state given  $s_i$ , called *first-order belief*. Since every  $j$  has a first-order belief  $\mu_j^1(s_j)$  for every message  $s_j$ ,  $i$ 's belief about  $s_j$  (given  $s_i$ ) gives  $i$  a belief about  $\mu_j^1(s_j)$ . This belief about  $j$ 's belief about the state is  $i$ 's second-order belief  $\mu_i^2(s_i)$  given  $s_i$ .<sup>6</sup> Since every  $j$  has a second-order belief  $\mu_j^2(s_j)$  for every  $s_j$ ,  $i$  can formulate a third-order belief given  $s_i$  and so on. By induction, every  $s_i$  induces a belief hierarchy  $h_i(s_i) \in T_i$  for agent  $i$ , and every message profile  $s$  induces a profile of belief hierarchies  $h(s) := (h_i(s_i))_{i \in N}$ .

**Definition 1.** An information structure  $(S, \pi)$  induces a distribution  $\tau \in \Delta T$  over

<sup>4</sup>A hierarchy  $t$  is coherent if any belief  $t_i^k$  coincides with all beliefs of lower order,  $\{t_i^n\}_{n=1}^{k-1}$ , on lower order events:  $\text{marg}_{X_{k-1}} t_i^k = t_i^{k-1}$  for all  $k \geq 1$  where  $X_{k-1} = \text{supp } t_i^{k-1}$ .

<sup>5</sup>We often write  $\beta_i^*(t_{-i}|t_i)$  and  $\beta_i^*(\theta|t_i)$  to refer to the corresponding marginals.

<sup>6</sup>Technically speaking, a second-order belief also includes a first-order belief.

$\pi(\cdot 0)$	$s_1$	$s_2$	$\pi(\cdot 1)$	$s_1$	$s_2$
$s_1$	1	0	$s_1$	$\frac{1}{2}$	0
$s_2$	0	0	$s_2$	0	$\frac{1}{2}$

TABLE 1: A (Public) Information Structure

profiles of belief hierarchies, called a **belief(-hierarchy) distribution**, if

$$\tau(t) = \sum_{\theta} \pi(\{s : h(s) = t\}|\theta) \mu_0(\theta) \quad (3)$$

for all  $t$ .

For example, the information structure in Table 1 induces  $\tau = \frac{3}{4}t_{1/3} + \frac{1}{4}t_1$  when  $\mu_0 := \mu_0(\theta = 1) = \frac{1}{2}$ , where  $t_\mu$  is the hierarchy profile in which  $\mu := \mu(\theta = 1)$  is commonly believed.<sup>7</sup>

We categorize belief distributions into public and private. This distinction is closely linked to the nature of information that induces those distributions.

**Definition 2.** A belief distribution  $\tau$  is **public** if  $t_i^1 = t_j^1$  and  $\text{marg}_{T_{-i}} \beta_i^*(\cdot|t_i) = \delta_{t_{-i}}$  (where  $\delta$  is the Dirac measure) for all  $t \in \text{supp } \tau$  and  $|\text{supp } \tau| \geq 2$ . A belief distribution  $\tau$  is **private** if it is not public.

The first part says that agents share the same first-order beliefs and this is commonly believed among them. This is the natural translation in terms of beliefs of the standard notion of public information. Notice also that we categorize the degenerate case  $|\text{supp } \tau| = 1$  as private. When the support is a singleton this distinction is indeed mostly a matter of semantics; yet the current choice makes our characterization later on more transparent.

## 3.2. Manipulation

Consider an environment with two agents and two equally likely states  $\theta \in \{0, 1\}$ . For each agent  $i = 1, 2$ , consider two belief hierarchies,  $t_i$  and  $t'_i$ , such that

$$\beta_i^*(t_j|t_i) = \beta_i^*(t'_j|t'_i) = 1 \quad \forall i, j \neq i, \quad (4)$$

$\beta_1^*(\theta = 1|t_1) = \beta_1^*(\theta = 1|t'_2) = 0.8$  and  $\beta_1^*(\theta = 1|t'_1) = \beta_1^*(\theta = 1|t_2) = 0.2$ . In words, at  $(t_1, t_2)$  and  $(t'_1, t'_2)$ , the agents believe that  $\theta = 1$  with different probabilities, 0.8 or

<sup>7</sup>To see why, note that  $\Pr(s_1, s_1) = \frac{3}{4}$ ,  $\Pr(s_2, s_2) = \frac{1}{4}$ , and an agent  $i$  receiving message  $s_\ell$  has beliefs  $(2\ell - 1)/3$  that  $\theta = 1$  and is certain that  $j$  also received  $s_\ell$ .



0.2, and this disagreement is commonly known. Can the belief-hierarchy distribution  $\tau(t) = \tau(t') = 1/2$  be induced by some information structure? More generally, can the designer ever get agents to agree to disagree? Since [Aumann \(1976\)](#), we have known that Bayesian agents cannot agree to disagree if they have a common prior. Say that  $p \in \Delta(\Theta \times T)$  is a **common prior** if

$$p(\theta, t) = \beta_i^*(\theta, t_{-i}|t_i)p(t_i) \quad (5)$$

for all  $\theta, t$  and  $i$ . That is, all agents  $i$  obtain their belief map  $\beta_i^*$  by Bayesian updating of the same distribution  $p$ . Denote by  $\Delta^f$  the probability measures with finite support. Define

$$\mathcal{C} := \left\{ \tau \in \Delta^f T : \exists \text{ a common prior } p \text{ s.t. } \tau = \text{marg}_T p \right\} \quad (6)$$

to be the space of **consistent** (belief-hierarchy) distributions. In a consistent distribution, all agents' beliefs arise from a common prior that draws every  $t$  with the same probability as  $\tau$ , i.e.,  $\tau = \text{marg}_T p$ . Let  $p_\tau$  be the unique distribution  $p$  in (6) (uniqueness follows from [Mertens and Zamir \(1985, Proposition 4.5\)](#)).

Note that consistency does not require that  $\text{marg}_\Theta p = \mu_0$ , which is a conceptually different point. A distribution  $\tau \in \Delta^f T$  is **Bayes plausible for agent  $i$**  if

$$\sum_{t_i} \text{marg}_\Theta \beta_i^*(\cdot|t_i)\tau_i(t_i) = \mu_0,$$

that is, if agent  $i$ 's expected first-order belief equals the state distribution.

**Proposition 1.** *There exists an information structure that induces  $\tau \in \Delta^f T$ , if and only if,  $\tau$  is consistent and Bayes plausible for some agent  $i$ .*

This characterization, which builds upon [Mertens and Zamir \(1985\)](#), disciplines the designer's freedom in shaping agents' beliefs. In the one-agent case, information disclosure is equivalent to choosing a Bayes plausible distribution over first-order beliefs. In the multi-agent case, it is equivalent to choosing a *consistent* distribution over belief hierarchies, which is Bayes plausible for some agent. Importantly, it does not matter which agent  $i$  satisfies Bayes plausibility, because by consistency, if it is true for one agent, then it will hold for all.

Notice, however, that merely ensuring that Bayes plausibility holds for all agents does not guarantee consistency. In the simple example above,  $\tau$  is Bayes plausible for both agents and yet fails consistency, because  $\beta_1^*(\theta = 1, t'_2|t'_1) \cdot \frac{1}{2} = 0.2 \cdot \frac{1}{2} \neq \beta_2^*(\theta = 1, t'_1|t'_2) \cdot \frac{1}{2} = 0.8 \cdot \frac{1}{2}$  violates (5).

From an operational viewpoint, there are two distinct ways of designing a con-



sistent belief-hierarchy distribution. The first one is to design the distributions of agents' beliefs individually and, then, couple them so as to ensure consistency (see Ely (2017) for a related procedure). A different approach, formulated in the next proposition, is to design the distribution of *designer's* beliefs and then to derive from it the resulting consistent distribution of agents' beliefs.

**Proposition 2.**  $\tau \in \Delta^f(T)$  is consistent and Bayes plausible if and only if there exists  $\nu : \text{supp } \tau \rightarrow \Delta\Theta$  such that, for all  $t, \theta$  and  $i$ ,

$$\sum_t \tau(t) \nu(\theta|t) = \mu_0(\theta), \quad (7)$$

$$\nu(\theta|t_i, t_{-i}) = \beta_i^*(\theta|t_i, t_{-i}) := \frac{\beta_i^*(\theta, t_{-i}|t_i)}{\beta_i^*(t_{-i}|t_i)} \quad (8)$$

$$\tau(t_{-i}|t_i) = \beta_i^*(t_{-i}|t_i). \quad (9)$$

Whereas each agent  $i$  only observes  $t_i$ , the designer observes the entire hierarchy profile  $t = (t_1, \dots, t_n)$ , and hence is better informed than every single agent. In the above,  $\nu(t)$  is interpreted as the designer's beliefs about the state, obtained by conditioning on  $t$ . The result demonstrates that any consistent and Bayes plausible belief-hierarchy distribution corresponds to a specific distribution of designer's beliefs, and vice-versa—a relationship governed by three conditions. The first one states that the average of the designer's beliefs is equal to the distribution of states. This is a form of Bayes plausibility at the designer's level. The second requires that the designer's beliefs about the state be the same as what *any* agent would believe if he also knew the information of all the other agents. The last condition requires that  $i$ 's conditional beliefs about the other agents' hierarchies can be derived from  $\tau$  by conditioning on  $i$ 's belief hierarchy. This result suggests a different way of approaching the design of (consistent and Bayes plausible) belief-hierarchy distributions: it is equivalent to designing the distribution of the designer's beliefs subject to Bayes plausibility in (7) and deriving the agents' hierarchies via (8) and (9). We demonstrate how this can be a useful way of implementing consistency in Section 5.

### 3.3. Outcomes from Belief Distributions

To complete the formulation of information design in the space of beliefs, the equivalence between information structures and belief distributions should be more than epistemic, it should be about outcomes. For any consistent distribution  $\tau$ , let a solution concept be a collection  $\Sigma^B(\tau)$  of sets  $\Lambda \subseteq \{\sigma : \text{supp } \tau \rightarrow \Delta A\}$ .<sup>8</sup> This describes

---

<sup>8</sup>We thank Daniel Clark for pointing out an error which led to a substantial improvement of this section.

agents' behavior in the Bayesian game  $\langle G, p_\tau \rangle$ , where  $p_\tau$  is the unique common prior  $p$  such that  $\text{marg}_T p = \tau$ . The different  $\Lambda$ 's capture possibly correlated behaviors, enabled by the addition of different (belief-preserving) correlating devices to  $\tau$  (see Liu (2009)). Given any  $\Lambda \in \Sigma^B(\tau)$ , the resulting outcomes are

$$O_\Lambda(\tau) := \{\gamma \in \Delta(A \times \Theta) : \text{there exists } \sigma \in \Lambda \text{ such that} \\ \gamma(a, \theta) = \sum_t \sigma(a|t) p_\tau(t, \theta) \text{ for all } (a, \theta)\},$$

from which it follows that the designer's ex-ante expected payoff from a consistent distribution  $\tau$  is

$$w(\tau) := \sup_{\Lambda \in \Sigma^B(\tau)} \sum_{\theta, a} g(O_\Lambda(\tau))(a, \theta) v(a, \theta). \quad (10)$$

Note that  $\Sigma^B$  is useful only insofar as it captures the appropriate outcomes from solution concept  $\Sigma$ . Hence, we assume that  $\Lambda \in \Sigma^B(\tau)$  if and only if  $\Lambda$  is such that  $O_\Lambda(\tau) = O_\Sigma(S, \pi)$  for some  $(S, \pi)$  inducing  $\tau$ . In general, it is well-known that a given solution concept may not yield the same set of outcomes when applied to an information structure as when applied to the corresponding belief distribution (e.g., Ely and Peski (2006) and Liu (2009)). Indeed, even if  $(S, \pi)$  induces  $\tau$ , there may be multiple message profiles  $s$  inducing the same hierarchy profile  $t$ , thus creating opportunities for redundant correlations of agents' behavior that are not possible in  $\tau$ . In some cases, the literature makes clear which  $\Sigma^B$  should be chosen given  $\Sigma$ . For example, if  $\Sigma = \text{BNE}$ , then said correlations can be recovered (without affecting the beliefs in  $\tau$ ) by taking  $\Sigma^B$  to be Liu (2015)'s belief-preserving correlated equilibrium. Alternatively, if  $\Sigma$  is interim correlated rationalizability (Dekel, Fudenberg, and Morris (2007)), then  $\Sigma^B$  should also be interim correlated rationalizability.

## 4. Representation of Optimal Solutions

In this section, we prove that optimal solutions to information design problems in games can be seen as a combination of special distributions. As a consequence, all optimal solutions can be decomposed into optimal purely private and optimal public components, where the latter come from concavification.

### 4.1. Assumptions

Our approach can handle various selection rules and solution concepts, provided the following assumptions hold:

**(Linear Selection).** For all  $D', D'' \subseteq \Delta(\Theta \times A)$  and  $0 \leq \alpha \leq 1$ ,  $g(\alpha D' + (1 -$

$$\alpha)D'') = \alpha g(D') + (1 - \alpha)g(D'').$$

**(Invariant Solution).** Fix  $\tau, \tau', \tau'' \in \mathcal{C}$  such that  $\text{supp } \tau = \text{supp } \tau' \cup \text{supp } \tau''$  and  $\text{supp } \tau' \cap \text{supp } \tau'' = \emptyset$ .  $\Lambda \in \Sigma^{\text{B}}(\tau)$  if and only if there exist  $\Lambda' \in \Sigma^{\text{B}}(\tau')$  and  $\Lambda'' \in \Sigma^{\text{B}}(\tau'')$  such that  $\Lambda = \left\{ \sigma : \text{supp } \tau \rightarrow \Delta A \mid \sigma|_{\text{supp } \tau'} \in \Lambda' \text{ and } \sigma|_{\text{supp } \tau''} \in \Lambda'' \right\}$ .

Linearity of  $g$  is a natural assumption that requires the selection criterion to be independent of the subsets of outcomes to which it is applied. The best and the worst outcomes, defined in (1), are linear selection criteria. However, selecting the best outcome within one subset and the worst in another breaks linearity, unless the outcome is always unique.

Invariance says that play at a profile of belief hierarchies  $t$  under  $\Sigma^{\text{B}}$  is independent of the ambient distribution from which  $t$  is drawn. For instance, Liu (2015)'s correlated equilibrium satisfies invariance. And when  $\Sigma = \text{ICR}$  (that is, Interim Correlated Rationalizability by Dekel, Fudenberg, and Morris (2007)), it is well-known that  $\Sigma^{\text{B}} = \Sigma$ , which also satisfies invariance. Invariance is important because it allows us to recover the outcomes from any consistent distribution through appropriate randomizations over distributions with smaller supports and, thus, ensures linearity of the ex-ante payoff (Proposition 4 and Lemma 3 in the Appendix).

## 4.2. Representations

Information design exhibits a convex structure when seen as belief manipulation. From any consistent  $\tau'$  and  $\tau''$ , the designer can build a third distribution,  $\tau = \alpha\tau' + (1 - \alpha)\tau''$ , which can be induced by an information structure provided that it is Bayes plausible. In particular, this is true even if  $\tau'$  and  $\tau''$  are themselves *not* Bayes plausible. In technical terms,  $\mathcal{C}$  is convex and, moreover, admits extreme points.<sup>9</sup> In the tradition of extremal representation theorems,<sup>10</sup> the designer generates a consistent and Bayes plausible distribution by randomizing over extreme points, and *any* consistent and Bayes plausible distribution can be generated in this way. By Proposition 1, any information structure can thus be interpreted as a convex combination of extreme points. Importantly, these extreme points have a useful characterization: they are the minimal consistent distributions (see Lemma 2 in Appendix A.3). A consistent distribution  $\tau \in \mathcal{C}$  is **minimal** if there is no  $\tau' \in \mathcal{C}$  such that  $\text{supp } \tau' \subsetneq \text{supp } \tau$ . Let  $\mathcal{C}^{\text{M}}$  denote the set of all minimal distributions,<sup>11</sup>

<sup>9</sup>An extreme point of  $\mathcal{C}$  is an element  $\tau \in \mathcal{C}$  with the property that if  $\tau = \alpha\tau' + (1 - \alpha)\tau''$ , given  $\tau', \tau'' \in \mathcal{C}$  and  $\alpha \in [0, 1]$ , then  $\tau' = \tau$  or  $\tau'' = \tau$ .

<sup>10</sup>E.g., Minkowski–Caratheodory theorem, Krein–Milman theorem, and Choquet's integral representation theorem.

<sup>11</sup>Minimal belief subspaces appeared in contexts other than information design in Heifetz and Neeman (2006), Barelli (2009), and Yildiz (2015).

which is nonempty by basic inclusion arguments. From this definition follows a nice interpretation in terms of information. The minimal distributions correspond to *purely private* information. By definition, any non-minimal distribution  $\tau$  contains two consistent components with support  $\text{supp } \tau'$  and  $\text{supp } \tau \setminus \text{supp } \tau'$ . A public signal makes it common knowledge among the agents which of these two components their beliefs are in. Since a minimal belief distribution has only one component, it contains no such public signal. As such, it is purely private information (possibly degenerate).

Owing to their mathematical status as extreme points, the minimal consistent distributions correspond to the basic elements from which *all* other consistent distributions can be constructed. In the single-agent case, the minimal distributions are the agent's first-order beliefs. The results below formalize their special role in information design.<sup>12</sup>

**Theorem 1** (Representation Theorem). *The designer's maximization problem can be represented as*

$$\begin{aligned} \sup_{(S, \pi)} V(S, \pi) = & \sup_{\lambda \in \Delta^f(\mathcal{C}^M)} \sum_{e \in \text{supp } \lambda} w(e) \lambda(e) \\ \text{subject to} & \sum_{e \in \text{supp } \lambda} \text{marg}_{\Theta} p_e \lambda(e) = \mu_0. \end{aligned} \quad (11)$$

**Corollary 1** (Within–Between Maximizations). *For any  $\mu \in \Delta\Theta$ , let*

$$w^*(\mu) := \sup_{e \in \mathcal{C}^M: \text{marg}_{\Theta} p_e = \mu} w(e). \quad (12)$$

*Then, the designer's maximization problem can be represented as*

$$\begin{aligned} \sup_{(S, \pi)} V(S, \pi) = & \sup_{\lambda \in \Delta^f \Delta\Theta} \sum_{\text{supp } \lambda} w^*(\mu) \lambda(\mu) \\ \text{subject to} & \sum_{\text{supp } \lambda} \mu \lambda(\mu) = \mu_0. \end{aligned} \quad (13)$$

From the representation theorem, the designer maximizes her expected payoff *as if* she were optimally randomizing over minimal consistent distributions, subject to posterior beliefs averaging to  $\mu_0$  across those distributions. Every minimal distribution  $e$  induces a Bayesian game and leads to an outcome for which the designer receives expected payoff  $w(e)$ . Every minimal distribution has a (marginal) distribution on states,  $\text{marg}_{\Theta} p_e = \mu$ , and the “further” that is from  $\mu_0$ , the “costlier” it

---

<sup>12</sup>We further illustrate the notion of minimal distribution in the Supplementary Appendix by characterizing minimal distributions for public and conditionally independent information.

is for the designer to use it. In this sense, the constraint in (11) can be seen as a form of budget constraint.

The corollary decomposes the representation theorem into two steps. First, there is a **maximization within**—given by (12)—that takes place among all the minimal distributions with  $\text{marg}_{\Theta} p_e = \mu$  and for all  $\mu$ . All minimal distributions with the same  $\mu$  contribute equally toward the Bayes plausibility constraint; hence, the designer should choose the best one among them, i.e., the one that gives the highest value of  $w(e)$ . Interestingly, maximization within delivers the optimal value of private information, which takes the form of a value function  $\mu \mapsto w^*(\mu)$ . The possibility to identify the optimal value of private information comes from the fact that all minimal distributions represent purely private information.

Second, there is a **maximization between** that concavifies the value function, thereby optimally randomizing between the minimal distributions that maximize within. This step is akin to a public signal  $\lambda$  that “sends” all agents to different minimal distributions  $e$ , thus making  $e$  common knowledge. From standard arguments (Rockafellar (1970, p.36)), the rhs of (13) is a characterization of the **concave envelope** of  $w^*$ , defined as  $(\text{cav } w^*)(\mu) = \inf\{g(\mu) : g \text{ concave and } g \geq w^*\}$ . Hence, the corollary delivers a concave-envelope characterization of optimal design. In the one-agent case,  $\{e \in \mathcal{C}^M \text{ s.t. } \text{marg}_{\Theta} p_e = \mu\} = \{\mu\}$ , hence  $w^* = w$  in (12) and the theorem reduces to maximization between.

The above decomposition is most useful when maximization within is performed over a restricted set of minimal belief distributions. Such restrictions are made in the context of *constrained* information design. For example, in some applications, it may be appropriate to limit the maximization within to conditionally independent private information. Alternatively, there are environments in which, imposing restrictions can be done without loss of generality. This is shown in the application of the next section, where consistency, Bayes plausibility and the payoff externalities impose enough structure that the unconstrained optimal solution can be computed by restricting maximization within to a small subset of minimal belief distributions. Whether these restrictions constrain the optimal solution or not, the restricted set of minimal belief distributions forms the smallest class of distributions from which that optimal solution can be derived by means of concavification.

## 5. Application: Fostering Investment under Adversarial Selection

Monetary incentives have been used to stimulate investment and technology adoption (e.g., tax incentives by governments), to stabilize banks and currencies (e.g., financial interventions by central banks), and to increase efforts in organizations (e.g., compensation schemes by companies). In such situations, often characterized by coordination motives, the strategic provision of information by a third party constitutes a different way of leveraging the underlying complementarities.

We consider the problem of fostering investment in an interaction where two agents are choosing whether or not to invest,  $\{I, N\}$ , given an uncertain state  $\theta \in \{-1, 2\}$ . The payoffs of the interaction are summarized in Table 2. Let  $\mu_0 := \text{Prob}(\theta = 2) > 0$  denote the probability of the high state.

$(u_1, u_2)$	$I$	$N$
$I$	$\theta, \theta$	$\theta - 1, 0$
$N$	$0, \theta - 1$	$0, 0$

TABLE 2: Investment Game

Under complete information, each agent has a dominant strategy to invest when  $\theta = 2$  and not to invest when  $\theta = -1$ . Under incomplete information, however, a coordination problem arises. An agent with a “moderate” belief that  $\theta = 2$  will invest if and only if he believes that the other agent is likely enough to invest as well. This gives rise to multiple equilibria.<sup>13</sup>

Consider an information design problem with the following features:

- (a) The designer wants to stimulate investment and values the complementarities between agents. This is modeled by a symmetric and monotone payoff function:  $v(I, I) > v(I, N) = v(N, I) > v(N, N) = 0$ , such that  $v(I, I) \geq \frac{3}{2}v(I, N)$ . The latter condition includes all supermodular<sup>14</sup> designers.

---

<sup>13</sup>This game differs in an important way from Carlsson and van Damme (1993) and Morris and Shin (2003), because the coordination problem arises only under incomplete information. It also differs from Rubinstein (1989), Kajii and Morris (1997), and Hoshino (2018), because no equilibrium of the complete information game is robust to the introduction of incomplete information. For example,  $(N, N)$  is a dominant equilibrium in the low state  $\theta = -1$ , which makes  $(I, I)$  sensitive to the introduction of incomplete information.

<sup>14</sup>Supermodular (submodular) designers are those for which  $v(I, I) + v(N, N) \geq (\leq)v(N, I) + v(I, N)$ .

(b) The solution concept is Bayes Nash Equilibrium (BNE).

(c) The *min* selection rule, defined in (1), chooses the worst BNE outcome.

Adversarial equilibrium selection, defined by the *min* in (c), corresponds to a form of robust information design: the designer aims to maximize the payoff she would obtain if the worst equilibrium for her were played. When there are multiple equilibria, it is difficult to predict which one the agents will coordinate on. Therefore, in environments where erroneous coordination can be particularly detrimental, it is especially important that information be designed to secure the highest possible payoff.

The current method available for Bayes-Nash information design, based on BCE (see Bergemann and Morris (2016) and Taneva (2019)), cannot be used to solve this problem, because it does not apply to adversarial selection. To see why, consider any  $\mu_0 \in [\frac{1}{2}, \frac{2}{3})$ . The solution to the BCE program prescribes to provide no information to the agents, namely  $\pi^*(I, I|\theta) = 1$  for all  $\theta$ . This is because the BCE method implicitly invokes the *max* selection rule (see Bergemann and Morris (2018) for further discussion). Since  $(I, I)$  and  $(N, N)$  are both BNE in the absence of any additional information, the max rule selects  $(I, I)$ . Instead, under the min rule,  $(N, N)$  will be selected, which results in the smallest possible payoff for the designer. We use our approach to compute an optimal information structure and show that, under adversarial selection, the designer can achieve a higher expected payoff for  $\mu_0 \in [\frac{1}{2}, \frac{2}{3})$  by revealing some information privately to one of the agents while leaving the other agent uninformed.

## 5.1. Worst-Equilibrium Characterization

We begin with a characterization of the worst BNE for the designer. When an agent believes that  $\theta = 2$  with probability larger than  $2/3$ , investing is uniquely optimal<sup>15</sup> for him, irrespective of his belief about the other agent's action. Investing can also be uniquely optimal even when an agent's belief that  $\theta = 2$  is less than  $2/3$ , if that agent believes the other agent will invest with large enough probability.

Using the concepts from Section 3, let  $\rho_i^k$  be the set of hierarchies defined inductively as follows:

$$\rho_i^1 = \left\{ t_i : \beta_i^*(\{\theta = 2\} \times T_j | t_i) > \frac{2}{3} \right\}$$

$$\rho_i^k = \left\{ t_i : \beta_i^*(\{\theta = 2\} \times T_j | t_i) + \frac{1}{3} \beta_i^*(\Theta \times \rho_j^{k-1} | t_i) > \frac{2}{3} \right\}.$$

---

<sup>15</sup>Formally, uniquely rationalizable.



If agent  $i$ 's hierarchy  $t_i$  is in  $\rho_i^1$ , then he believes with probability greater than  $2/3$  that  $\theta = 2$ , and his unique optimal response is to play  $I$ . If  $t_i \in \rho_i^2$ , then agent  $i$  assigns high enough probability either to  $\theta = 2$  or to agent  $j$  playing  $I$  (due to  $t_j \in \rho_j^1$ ), so that  $I$  is again uniquely optimal. By induction, the same conclusion holds for hierarchies in  $\rho_i^k$  for any  $k \geq 1$ . Letting  $\rho_i := \cup_{k \geq 1} \rho_i^k$ , the unique optimal action for an agent with belief in  $\rho_i$  is  $I$ . This implies that, in *all* BNEs, agent  $i$ 's equilibrium strategy must choose  $I$  with certainty when his hierarchy is in  $\rho_i$ .

Given a belief distribution  $\tau \in \mathcal{C}$ , the worst equilibrium for our designer is such that all agents play  $I$  *only* when their beliefs belong to  $\rho_i$  (that is, *only* when  $I$  is uniquely rationalizable) and play  $N$  otherwise: for all  $i$  and  $t_i$ ,

$$\sigma_i^{\text{MIN}}(I|t_i) = \begin{cases} 1 & \text{if } t_i \in \rho_i \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

## 5.2. Solution

Following Theorem 1 and Corollary 1 we solve the information design problem in two steps: maximization within and maximization between.

### 5.2.1. Maximization Within

We first characterize the state distributions for which the maximization within induces joint investment (i.e., investment by both agents) with certainty in the worst equilibrium. For  $\mu > 2/3$ , each agent invests regardless of the other's decision. Therefore, it is optimal to provide no information. For  $\mu < 2/3$ , instead, information can be used, at least in some cases, to induce joint investment with certainty by exploiting the equilibrium structure. In these cases, the designer induces some agent (say agent 1) to invest at a type  $t_1$  only based on his first-order belief being greater than  $2/3$ , i.e.,  $t_1 \in \rho_1^1$ . Leveraging the presence of  $t_1$ , the designer can induce a type  $t_2$  of agent 2 to invest, with a first-order belief lower than  $2/3$ . This type finds investing uniquely optimal because of the sufficiently high probability it assigns to  $t_1$ , i.e.,  $t_2 \in \rho_2^2$ . This allows for a type  $t'_1$  of agent 1 to invest, with even lower first-order beliefs, based on the probability it assigns to  $t_2$ , i.e.,  $t'_1 \in \rho_1^3$ . This chain of reasoning continues for as many messages as the designer sends, each message corresponding to a different hierarchy.

We combine this insight with the conditions of Proposition 2 to obtain a system of inequalities and one equation ((29)-(32) in the Appendix), which characterizes the minimal distributions that induce joint investment with certainty at  $\mu$ . Simple addition of the inequalities shows that  $\mu$  must be greater than  $1/2$ , as stated in the

following result.

**Claim 1.** *Joint investment with certainty can be achieved if and only if  $\mu > 1/2$ .*

In the maximization within, revealing no information for  $\mu \in (1/2, 2/3)$  yields no investment in the worst equilibrium. Therefore, the designer has to reveal some (private) information to induce joint investment with certainty for this range of  $\mu$ . In fact, for all  $\mu \in (1/2, 2/3)$  this can be achieved with the simple minimal distributions presented in Table 3, where  $\varepsilon$  is a small positive number. From Claim 1 it follows

$e_\mu^*$	$t_2 : \mu_2 = \mu$
$t_1 : \mu_1 = 2/3 + \varepsilon$	$3(\mu - \varepsilon) - 1$
$t'_1 : \mu'_1 = 1/3 + \varepsilon$	$2 - 3(\mu - \varepsilon)$

TABLE 3: Optimal Minimal Distributions for  $\mu \in (1/2, 2/3)$

that the designer can no longer generate joint investment with certainty for  $\mu \leq 1/2$ . But can she still generate enough investment that her value from the maximization within is relevant for the concavification stage? As we show in the next section, for many monotone designers the answer is no; hence maximization within for  $\mu \leq 1/2$  can be dispensed with.

### 5.2.2. Maximization Between

Assuming the designer does not maximize within for  $\mu < 1/2$ , we obtain a value function  $\tilde{w}^*(\mu)$  equal to  $v(I, I)$  if  $\mu > 1/2$  and 0 otherwise. This function is plotted (dashed lines) in Figure 1 for a designer with  $v(I, I) = 2$  and  $v(N, I) = v(I, N) = 1$ . The concave envelope of  $\tilde{w}^*$ ,

$$(\text{cav } \tilde{w}^*)(\mu) = \begin{cases} v(I, I) & \text{if } \mu > 1/2 \\ 2\mu v(I, I) & \text{if } \mu \leq 1/2, \end{cases}$$

is also depicted (solid blue lines) in Figure 1. Our next result states that this concave envelope represents the optimal solution. We prove that the value of maximization within for  $\mu \leq 1/2$  cannot be relevant for the concavification, and therefore its computation within this range can be omitted without loss of generality.

**Claim 2.**  *$(\text{cav } \tilde{w}^*)(\mu_0)$  is the designer's optimal expected value at  $\mu_0$ .*

The optimal belief-hierarchy distribution is given by:

- for  $\mu_0 > \frac{2}{3}$ , reveal no information

- for  $\mu_0 \in (\frac{1}{2}, \frac{2}{3}]$ , use  $e_\mu^*$  with probability 1.
- for  $\mu_0 \leq \frac{1}{2}$ , use a public signal that either draws  $e_{1/2+\varepsilon}^*$  with probability  $2\mu_0/(1+2\varepsilon)$  or makes it common knowledge that  $\theta = -1$  with the remaining probability.<sup>16</sup>

The corresponding optimal information structure for  $\mu_0 \leq \frac{1}{2}$  is presented in Table 4.

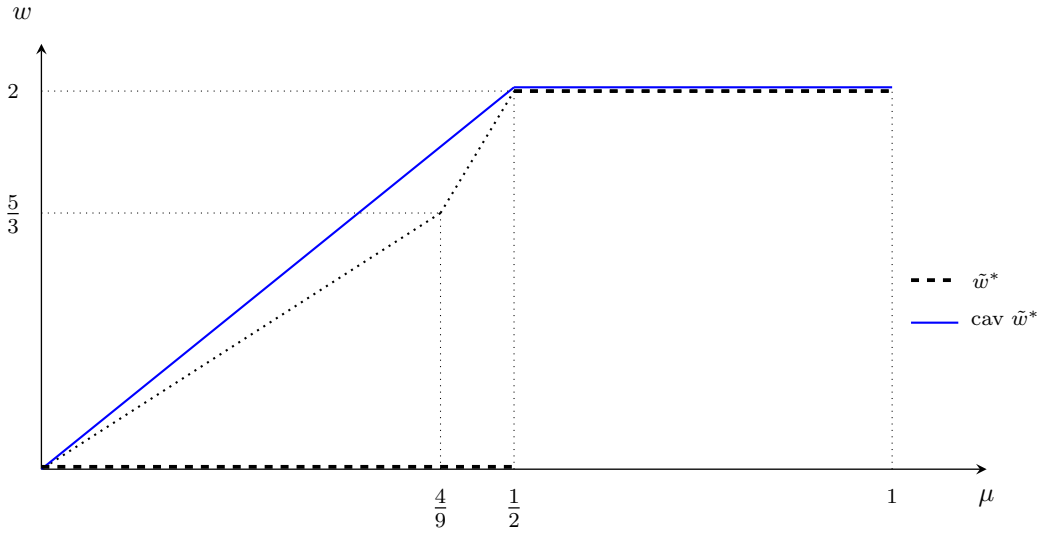


FIGURE 1: Optimal Value.

$\pi^*(\cdot   \theta = -1)$	$s$	$q$	$\pi^*(\cdot   \theta = 2)$	$s$	$q$
$s, s'_1$	$\frac{\mu_0}{1-\mu_0} \frac{1/3-\varepsilon}{1+2\varepsilon}$	0	$s, s'_1$	$\frac{2/3+\varepsilon}{1+2\varepsilon}$	0
$s, s''_1$	$\frac{\mu_0}{1-\mu_0} \frac{2/3-\varepsilon}{1+2\varepsilon}$	0	$s, s''_1$	$\frac{1/3+\varepsilon}{1+2\varepsilon}$	0
$q$	0	$\frac{1-2\mu_0+2\varepsilon}{(1-\mu_0)(1+2\varepsilon)}$	$q$	0	0

TABLE 4: Optimal information structure for  $\mu_0 \leq 1/2$ .

<sup>16</sup>Under adversarial selection, an optimal solution may not exist for all  $\mu_0$  (for instance at  $\mu_0 = 1/2$  in our example). However, the supremum of Theorem 1 can always be approached with arbitrary precision for all  $\mu_0$ . In our example,  $(\text{cav } \tilde{w}^*)(\mu_0)$  gives the exact value of the supremum for all  $\mu_0$ . It is obtained by concavifying the function  $\tilde{w}_\varepsilon^*$ , equal to 0 for all  $\mu < 1/2 + \varepsilon$  and 2 otherwise, and taking the pointwise limit as  $\varepsilon \downarrow 0$ . The function  $\tilde{w}_\varepsilon^*$  corresponds to using  $e_\mu^*$  from Table 3 (which depends on  $\varepsilon$ ) for  $\mu \geq 1/2 + \varepsilon$  and giving no information for lower  $\mu$ .

### 5.3. Extension

When the designer values unilateral investment enough that Claim 2 no longer holds, maximization within may become relevant for  $\mu \leq 1/2$ , because its value may dominate  $(\text{cav } \tilde{w}^*)(\mu)$ . For such designers and state distributions, the optimal minimal distribution will necessarily put mass on at least one hierarchy at which  $N$  is also rationalizable and will thus be played in the worst equilibrium. For these hierarchies, Lemma 5 in the Appendix establishes that it is optimal to set the first-order beliefs to zero, because, for a fixed  $\mu$ , this allows the designer to put more weight on other hierarchies that do invest. Therefore, an optimal minimal distribution can be found by solving our system ((29) – (32)), with as few first-order beliefs equal to zero as possible to find a solution at each  $\mu$ .<sup>17</sup>

### 5.4. Discussion

Private information plays a central role in the optimal design of *robust* incentives under adversarial selection. In our application, private information fosters robust investment by making agents uncertain about each other’s beliefs about the state, while, at the same time, making them certain about each other’s behavior. By giving some information to one of the agents while leaving the other one uninformed, the designer makes the latter uncertain about the beliefs of the former in a way that induces joint investment with certainty. It is sufficient for the uninformed agent to believe with high enough probability that investing is dominant for the other agent, in order to make investment uniquely optimal for him as well. In contrast, it is public information that plays the central role under favorable equilibrium selection. Notice that public information can also provide robust incentives to invest under adversarial selection, but only if it makes both agents extremely optimistic (with first-order beliefs greater than  $2/3$ ). This use of information, however, is suboptimal, because joint investment can be achieved through private information, even when agents are not as optimistic. How this is possible is explained next.

Optimal information disclosure under adversarial selection brings out two important principles underpinning the role of private information: *bandwagon effect* and *extreme pessimism*. The bandwagon effect is the force that provides robust incentives to invest without relying on agents’ (mutual) extreme optimism about the state. There needs to be only one type of one agent (agent 1 of type  $t_1$ ) with first-order beliefs about the state high enough to make investment strictly dominant

---

<sup>17</sup>We do not need to compute the maximization within in our example for  $\mu < 1/2$ , due to Claim 2. However, we used our system to compute its value for  $R = L = 2$  (that is, two messages per player) and have plotted the result (the dotted line) in Figure 1 for completeness.

at that type. Then, investing becomes uniquely optimal for the other agent (agent 2 of type  $t_2$ ) under a more pessimistic belief about the state, if he puts sufficiently high probability on agent 1 being extremely optimistic (i.e., on agent 1 being of type  $t_1$ ). In turn, investing becomes uniquely optimal for the second type of agent 1 (type  $t'_1$ ), with an even lower first-order belief, because he is certain that player 2 will always invest. This contagion process trades off pessimism about the state for optimism about the other agent's investment by taking advantage of the two kinds of payoff complementarities. This trade off can only be exploited by informative private signals.

Just as investment should be induced efficiently, so should be non-investment. In our application, the designer causes an extreme form of pessimism, according to which an agent is certain that the state is bad whenever he chooses not to invest. If, instead, the designer were to induce a moderate belief about the state, at which both actions could be optimal depending on coordination, then not investing would be played in the worst equilibrium. Therefore, the designer should make him as pessimistic as possible, since this does not change his behavior in the worst equilibrium, while allowing the designer to increase the probability of robust investment. This step also takes advantage of the complementarities: extremely pessimistic beliefs incentivize non-investment because the complementarities between an agent's action and the state overpower the strategic complementarities. This allows the designer to put more probability on higher beliefs at which the bandwagon effect induces robust investment. Extreme pessimism also emerges in the solution to the single-agent problem of [Kamenica and Gentzkow \(2011\)](#), in which the agent is certain about the state when taking the undesirable (to the designer) action. As in our application, this in turn allows the designer to induce the desirable action with maximal probability subject to Bayes plausibility.

## 6. Conclusion

This paper contributes to the foundations of information design with multiple interacting agents. Our representation results formulate the belief-based approach to the problem, and decompose it into maximization within and between, where the latter is concavification. This approach accommodates various equilibrium selection rules and solution concepts, which can be used to analyze diverse topics, such as robustness, bounded rationality, collusion, or communication. We provide an economic application based on a two-agent investment game, and apply our approach to solve the information design problem under adversarial equilibrium selection. An obvious avenue for future research is to generalize this robust information design to a class of

games with strategic complementarities, for which our results from the investment game provide the fundamental logic. Examining the implications of heterogeneous prior distributions among the agents is another interesting extension of the current framework.

# Appendix

## A. Proofs of Main Results

### A.1. Proof of Proposition 1

Let  $\tau$  be induced by some  $(S, \pi)$ , so that

$$\tau(t) = \sum_{\theta} \pi(\{s : h(s) = t\} | \theta) \mu_0(\theta) \quad (15)$$

for all  $t \in \text{supp } \tau$ . Define  $p \in \Delta(\Theta \times T)$  as

$$p(\theta, t) = \pi(\{s : h(s) = t\} | \theta) \mu_0(\theta) \quad (16)$$

for all  $\theta$  and  $t \in \text{supp } \tau$ . It is immediate from (15) and (16) that  $\text{marg}_T p = \tau$  and so  $\text{marg}_{T_i} p = \tau_i$  for all  $i$ . Further, when any agent  $i$  forms his beliefs under  $(S, \pi)$ , he computes  $\mu_i : S_i \rightarrow \Delta(\Theta \times S_{-i})$  by conditioning  $\pi(s | \theta) \mu_0(\theta)$  on  $s_i$ , so that  $\beta_i^* : T_i \rightarrow \Delta(\Theta \times T_{-i})$  is given by the conditional of  $p$  given  $t_i$ . That is,

$$p(\theta, t) = \beta_i^*(\theta, t_{-i} | t_i) \text{marg}_{T_i} p(t_i)$$

for all  $i, \theta$ , and  $t \in \text{supp } \tau$ . This shows  $\tau \in \mathcal{C}$ . Finally,

$$\sum_{t_i \in \text{supp } \tau_i} \beta_i^*(\theta | t_i) \tau_i(t_i) := \sum_{t \in \text{supp } \tau} p(\theta, t) = \sum_t \pi(\{s : h(s) = t\} | \theta) \mu_0(\theta) = \mu_0(\theta)$$

for all  $\theta$ , which proves Bayes plausibility.

Suppose now that  $\tau \in \mathcal{C}$  and satisfies Bayes plausibility. Let us show that these conditions are sufficient for  $\tau$  to be induced by some  $(S, \pi)$ . Define information structure  $(\text{supp } \tau, \pi_\tau)$  where

$$\pi_\tau(t | \cdot) : \theta \mapsto \frac{1}{\mu_0(\theta)} \beta_i^*(\theta, t_{-i} | t_i) \tau_i(t_i) \quad (17)$$

for all  $t \in \text{supp } \tau$ , which is defined independently of the choice of  $i$  because  $\tau \in \mathcal{C}$ . First, let us verify that  $\pi_\tau$  is a valid information structure. Bayes plausibility says

$$\sum_{t_i \in \text{supp } \tau_i} \beta_i^*(\theta | t_i) \tau_i(t_i) = \mu_0(\theta),$$



which guarantees that

$$\sum_{t \in \text{supp } \tau} \pi_\tau(t|\theta) = \frac{1}{\mu_0(\theta)} \sum_{t \in \text{supp } \tau} \beta_i^*(\theta, t_{-i}|t_i) \tau_i(t_i) = 1,$$

(i.e.,  $\pi(\cdot|\theta)$  is a probability distribution for every  $\theta$ ). By construction, this information structure is such that, when any agent  $j$  receives  $t_j$ , his beliefs are  $\mu_j(\cdot|t_j) = \beta_j^*(\cdot|t_j)$ , also because  $\tau \in \mathcal{C}$ . To prove that  $\pi_\tau$  generates  $\tau$ , we also need to check that

$$\tau(t) = \sum_{\theta} \pi_\tau(t|\theta) \mu_0(\theta) \quad (18)$$

for all  $t \in \text{supp } \tau$ . By (17), the rhs of (18) is equal to  $\beta_i^*(t_{-i}|t_i) \tau_i(t_i)$ , which equals  $\tau(t)$  because  $\tau \in \mathcal{C}$  (in particular, because  $\text{marg}_{\Theta} p = \tau$ ).  $\square$

## A.2. Proof of Proposition 2

*Proof.* Suppose  $\tau$  is consistent and Bayes plausible. By definition, there exists a common prior  $p \in \Delta(\Theta \times T)$  such that, for all  $t, \theta$  and  $i$ ,

$$p(\theta, t) = \beta_i^*(\theta, t_{-i}|t_i) p(t_i)$$

and  $\text{marg}_T p = \tau$ . Moreover, by Bayes' plausibility,  $\text{marg}_{\Theta} p = \mu_0$ . Define  $\nu(\theta|t) := p(\theta|t)$  for all  $t$  and  $\theta$ . We show that conditions (7), (8) and (9) are satisfied. First,

$$\begin{aligned} \sum_t \tau(t) \nu(\theta|t) &= \sum_t p(t) p(\theta|t) \\ &= \sum_t p(t) \frac{p(\theta, t)}{p(t)} \\ &= \mu_0(\theta), \end{aligned}$$

Moreover, since  $p$  is a common prior,  $p(\theta, t_{-i}|t_i) = \beta_i^*(\theta, t_{-i}|t_i)$  and  $p(t_{-i}|t_i) = \beta_i^*(t_{-i}|t_i)$  for all  $i, t$  and  $\theta$ . Thus, for all  $i, t$  and  $\theta$ ,

$$\nu(\theta|t_i, t_{-i}) := p(\theta|t_i, t_{-i}) = \frac{p(\theta, t_{-i}|t_i)}{p(t_{-i}|t_i)} = \beta_i^*(\theta|t_i, t_{-i}).$$

Finally, since  $\text{marg}_T p = \tau$  and  $p$  is a common prior,

$$\tau(t_{-i}|t_i) = p(t_{-i}|t_i) = \beta_i^*(t_{-i}|t_i),$$

for all  $i$  and  $t$ .

Now, consider  $\tau \in \Delta^f(T)$  and  $\nu : \text{supp } \tau \rightarrow \Delta\Theta$  such that (7), (8) and (9) hold. Define  $p \in \Delta(\Theta \times T)$  as  $p(\theta, t) := \tau(t) \nu(\theta|t)$  for all  $\theta$  and  $t$ . We want to show that

$p$  is a common prior for  $\tau$ . First, note that, by definition of  $p$ ,  $p(t) = \sum_{\theta} p(\theta, t) = \tau(t) \sum_{\theta} \nu(\theta|t) = \tau(t)$ , for all  $t$ . Therefore,  $\text{marg}_T p = \tau$ . Second, for all  $i, t$  and  $\theta$ ,

$$\begin{aligned} p(\theta, t_{-i}|t_i) &= \frac{p(\theta, t)}{p(t_i)} \\ &= \frac{\tau(t)\nu(\theta|t)}{\tau(t_i)} \\ &= \tau(t_{-i}|t_i)\nu(\theta|t) \\ &= \beta_i^*(t_{-i}|t_i)\beta_i^*(\theta|t) = \beta_i^*(\theta, t_{-i}|t_i). \end{aligned}$$

The first equality comes from Bayes' rule and the second equality comes from the definition of  $p$ . In the fourth equality, we used conditions (8) and (9). Finally, in the last equality, we used the definition of  $\beta_i^*(\theta|t_i, t_{-i})$ . This implies that  $p(\theta, t) = \beta_i^*(\theta, t_{-i}|t_i)p(t_i)$  for all  $\theta, t$  and  $i$ . Therefore,  $p$  is a common prior for  $\tau$  or, equivalently,  $\tau \in \mathcal{C}$ . Finally,  $\tau$  is Bayes plausible because, given (7),

$$p(\theta) = \sum_t p(\theta, t) = \sum_t \tau(t)\nu(\theta|t) = \mu_0(\theta),$$

for all  $\theta$ . □

### A.3. Proof of Theorem 1

**Lemma 1.**  $\mathcal{C}$  is convex.

*Proof.* Take  $\alpha \in [0, 1]$  and  $\tau', \tau'' \in \mathcal{C}$ . By definition of  $\mathcal{C}$ , there are  $p_{\tau'}$  and  $p_{\tau''}$  such that  $\text{marg}_T p_{\tau'} = \tau'$  and  $\text{marg}_T p_{\tau''} = \tau''$  and

$$\begin{aligned} p_{\tau'}(\theta, t) &= \beta_i^*(\theta, t_{-i}|t_i)\tau'_i(t_i), \\ p_{\tau''}(\theta, t) &= \beta_i^*(\theta, t_{-i}|t_i)\tau''_i(t_i), \end{aligned} \tag{19}$$

for all  $\theta, i$  and  $t$ . Define  $\tau := \alpha\tau' + (1 - \alpha)\tau''$  and note that  $\tau_i = \alpha\tau'_i + (1 - \alpha)\tau''_i$ , by the linearity of Lebesgue integral. Define

$$p_{\tau}(\theta, t) := \beta_i^*(\theta, t_{-i}|t_i)\tau_i(t_i)$$

for all  $i, \theta$ , and  $t \in \text{supp } \tau$ . Notice that  $p_{\tau}$  is well-defined, because of (19). Thus,

$$\text{marg}_T p_{\tau} = \alpha\text{marg}_T p_{\tau'} + (1 - \alpha)\text{marg}_T p_{\tau''} = \alpha\tau' + (1 - \alpha)\tau'' = \tau$$

and we conclude that  $\tau \in \mathcal{C}$ . □

Although  $\mathcal{C}$  is convex, it is not closed because we can build sequences in  $\mathcal{C}$  with growing supports, only converging to a belief-hierarchy distribution with an infinite

support. Still, the next lemma proves that minimal (consistent) distributions are the extreme points of the set of consistent distributions.

**Lemma 2.**  $\mathcal{E} = \mathcal{C}^M$ .

*Proof.* First, we show that  $\mathcal{C}^M \subseteq \mathcal{E}$ . Take  $\tau \in \mathcal{C}^M$ . If  $\tau \notin \mathcal{E}$ , then there exist  $\tau', \tau'' \in \mathcal{C}$  with  $\tau' \neq \tau''$  and  $\alpha \in (0, 1)$  such that  $\tau := \alpha\tau' + (1 - \alpha)\tau''$ . Since  $\tau \in \mathcal{C}^M$ ,  $\tau \in \mathcal{C}$ . Moreover,  $\text{supp } \tau' \cup \text{supp } \tau'' \subseteq \text{supp } \tau$ . For  $\lambda \in \mathbb{R}_+$ , let  $\tau_\lambda := \tau + \lambda(\tau - \tau'')$ , which is a linear combination between  $\tau'$  and  $\tau''$ . Indeed, by construction  $\tau - \tau'' = \alpha(\tau' - \tau'')$  and, therefore, we can rewrite  $\tau_\lambda = \alpha(1 + \lambda)\tau' + (1 - \alpha(1 + \lambda))\tau''$ . Clearly,  $\sum_{t \in \text{supp } \tau} \tau_\lambda(t) = 1$  for all  $\lambda \in \mathbb{R}_+$ . Define  $\Lambda := \{\lambda \geq 0 : \forall t \in \text{supp } \tau, 0 \leq \tau_\lambda(t) \leq 1\}$  so that, by construction,  $\tau_\lambda \in \mathcal{C}$  for all  $\lambda \in \Lambda$ . Next, we establish a number of simple properties of  $\Lambda$ . The set  $\Lambda$  is non-empty since both  $\lambda = 0$  and  $\lambda = \frac{1-\alpha}{\alpha}$  belong to it, which can be verified by substitution. Moreover, it is easy to check that  $\Lambda$  is convex. The set  $\Lambda$  is closed. To see this, it is enough to consider any increasing sequence  $(\lambda_m) \subset \Lambda^\infty$  such that  $\lambda_m \nearrow \lambda$ . We want to show  $\lambda \in \Lambda$ . For all  $t \in \text{supp } \tau$ , by definition of  $\tau$ , the sequence  $\tau_{\lambda_m}(t)$  is monotone (either non-decreasing or non-increasing) and, by definition of  $\Lambda$ , it lives in the compact interval  $[0, 1]$ . Therefore, it converges in  $\lim_m \tau_{\lambda_m}(t) \in [0, 1]$ . Hence,  $\lambda \in \Lambda$ . Therefore, we can write  $\Lambda = [0, \tilde{\lambda}]$  where  $\tilde{\lambda} := \max \Lambda \geq \frac{1-\alpha}{\alpha} > 0$ .

We want to show that  $\text{supp } \tau_{\tilde{\lambda}} \subsetneq \text{supp } \tau$ . To see this, let  $\tilde{t} \in \text{supp } \tau_{\tilde{\lambda}}$  and suppose  $\tilde{t} \notin \text{supp } \tau$ . Then, by definition,  $\tau_{\tilde{\lambda}}(\tilde{t}) = -\tilde{\lambda}\tau''(\tilde{t}) \leq 0$ , which is impossible. Moreover, there must exist  $t \in \text{supp } \tau$  such that  $\tau_{\tilde{\lambda}}(t) = 0$ . To see this, suppose not, i.e., suppose  $\text{supp } \tau = \text{supp } \tau_{\tilde{\lambda}}$ . Then, for all  $t \in \text{supp } \tau$ , we would have that  $\tau_{\tilde{\lambda}}(t) > 0$ . Since  $\tau_{\tilde{\lambda}} \in \mathcal{C}$  by construction, and since  $|\text{supp } \tau| > 1$  (otherwise  $\tau' = \tau''$ ), we also have that  $\tau_{\tilde{\lambda}}(t) < 1$  for all  $t \in \text{supp } \tau$ . Let  $T^- := \{t \in \text{supp } \tau : \tau(t) - \tau''(t) < 0\}$  and  $T^+ := \{t \in \text{supp } \tau : \tau(t) - \tau''(t) > 0\}$ . These sets are non-empty by assumption ( $\tau' \neq \tau''$ ). For  $t \in T^-$ , let  $\lambda(t) := \frac{-\tau(t)}{\tau(t) - \tau''(t)}$  and notice that  $0 = \tau(t) + \lambda(t)(\tau(t) - \tau''(t)) < \tau_{\tilde{\lambda}}(t)$ , implying  $\lambda(t) > \tilde{\lambda}$ . Similarly, for  $t \in T^+$  let  $\lambda(t) := \frac{1 - \tau(t)}{\tau(t) - \tau''(t)}$  and notice that  $1 = \tau(t) + \lambda(t)(\tau(t) - \tau''(t)) > \tau_{\tilde{\lambda}}(t)$ , implying  $\lambda(t) > \tilde{\lambda}$ . Now define  $\lambda' := \min\{\lambda(t) : t \in T^+ \cup T^-\}$ , which is well-defined since  $T^+ \cup T^-$  is finite. By construction,  $\tau_{\lambda'} \in \mathcal{C}$  and  $\lambda' > \tilde{\lambda}$ , a contradiction to the fact that  $\tilde{\lambda}$  is the max. Therefore, we conclude that  $\text{supp } \tau_{\tilde{\lambda}} \subsetneq \text{supp } \tau$  and thus  $\tau \notin \mathcal{C}^M$ .

We now show the converse,  $\mathcal{C}^M \supseteq \mathcal{E}$ . Suppose  $\tau \in \mathcal{C}$  is not minimal, i.e., there is a  $\tilde{\tau} \in \mathcal{C}$  such that  $\text{supp } \tilde{\tau} \subsetneq \text{supp } \tau$ . Define  $\tau', \tau'' \in \Delta T$  as  $\tau'(\cdot) := \tau(\cdot | \text{supp } \tilde{\tau})$  and  $\tau''(\cdot) := \tau(\cdot | \text{supp } \tau \setminus \text{supp } \tilde{\tau})$ , the conditional distributions of  $\tau$  on  $\text{supp } \tilde{\tau}$  and  $\text{supp } \tau \setminus \text{supp } \tilde{\tau}$ . Clearly,

$$\tau = \alpha\tau' + (1 - \alpha)\tau'' \quad (20)$$

where  $\alpha = \tau(\text{supp } \tilde{\tau}) \in (0, 1)$ . Since  $\text{supp } \tilde{\tau}$  is belief-closed, so is  $\text{supp } \tau \setminus \text{supp } \tilde{\tau}$ . To see why, note that for any  $t \in \text{supp } \tau \setminus \text{supp } \tilde{\tau}$ , if there were  $i, \tilde{t} \in \text{supp } \tilde{\tau}$  and  $\theta \in \Theta$  such that  $p_\tau(\theta, \tilde{t}_{-i}|t_i) > 0$ , then this would imply  $p_\tau(\theta, t_i, \tilde{t}_{-i}) > 0$  and, thus,  $p_\tau(\theta, t_i, \tilde{t}_{-(ij)}|\tilde{t}_j) > 0$  (where  $\tilde{t}_{-(ij)}$  is the hierarchy profiles for agents other than  $i$  and  $j$ ). This implies that at  $\tilde{t}_j$  — a hierarchy that agent  $j$  can have in  $\tilde{\tau}$  — agent  $j$  assigns strictly positive probability to a hierarchy of agent 1 that is not in  $\text{supp } \tilde{\tau}$ . This contradicts the fact that  $\text{supp } \tilde{\tau}$  is belief-closed. Since  $\tau'$  and  $\tau''$  are derived from a consistent  $\tau$  and are supported on a belief-closed subspace,  $\tau'$  and  $\tau''$  are consistent. Given that  $\tau'' \neq \tau'$ , (20) implies that  $\tau$  is not an extreme point.  $\square$

**Proposition 3.** *For any  $\tau \in \mathcal{C}$ , there exist unique  $\{e_\ell\}_{\ell=1}^L \subseteq \mathcal{C}^M$  and weakly positive numbers  $\{\alpha_\ell\}_{\ell=1}^L$  such that  $\sum_{\ell=1}^L \alpha_\ell = 1$  and  $\tau = \sum_{\ell=1}^L \alpha_\ell e_\ell$ .*

*Proof.* Take any  $\tau \in \mathcal{C}$ . Either  $\tau$  is minimal, in which case we are done, or it is not, in which case there is  $\tau' \in \mathcal{C}$  such that  $\text{supp } \tau' \subsetneq \text{supp } \tau$ . Similarly, either  $\tau'$  is minimal, in which case we conclude that there exists a minimal  $e^1 := \tau'$  such that  $\text{supp } e^1 \subsetneq \text{supp } \tau$ , or there is  $\tau'' \in \mathcal{C}$  such that  $\text{supp } \tau'' \subsetneq \text{supp } \tau'$ . Given that  $\tau$  has finite support, finitely many steps of this procedure deliver a minimal consistent belief-hierarchy distribution  $e^1$ ,  $\text{supp } e^1 \subsetneq \text{supp } \tau$ . Since  $\tau$  and  $e^1$  are both consistent and hence, their supports are belief-closed,  $\text{supp } \tau \setminus \text{supp } e^1$  must be belief-closed. Given that  $\text{supp } \tau \setminus \text{supp } e^1$  is a belief-closed subset of  $\text{supp } \tau$  and  $\tau$  is consistent, define a new distribution  $\tau^2$  as

$$p_{\tau^2}(\theta, t) := \frac{p_\tau(\theta, t)}{\tau(\text{supp } \tau \setminus \text{supp } e^1)}$$

for all  $\theta \in \Theta$  and  $t \in \text{supp } \tau \setminus \text{supp } e^1$ . By construction,  $\text{supp } \tau^2 = \text{supp } \tau \setminus \text{supp } e^1$ . Moreover, since  $\tau \in \mathcal{C}$ ,  $p_\tau(\theta, t) = \beta_i^*(\theta, t_{-i}|t_i)\tau(t_i)$  for all  $\theta \in \Theta$ ,  $t \in \text{supp } \tau$ , and  $i$ . Hence,

$$p_{\tau^2}(\theta, t) = \frac{p_\tau(\theta, t)}{\tau(\text{supp } \tau^2)} = \frac{\beta_i^*(\theta, t_{-i}|t_i)\tau(t_i)}{\tau(\text{supp } \tau^2)} = \beta_i^*(\theta, t_{-i}|t_i)\tau^2(t_i)$$

for all  $\theta \in \Theta$ ,  $t \in \text{supp } \tau^2$ , and  $i$ . In addition,

$$\text{marg}_T p_{\tau^2}(\theta, t) = \frac{\text{marg}_T p_\tau(\theta, t)}{\tau(\text{supp } \tau^2)} = \frac{\tau(t)}{\tau(\text{supp } \tau^2)} = \tau^2(t)$$

for all  $\theta \in \Theta$  and  $t \in \text{supp } \tau^2$ . Hence,  $\tau^2 \in \mathcal{C}$ . Therefore, we can repeat the procedure again for distribution  $\tau^2 \in \mathcal{C}$ . Since  $\tau$  has finite support, there exists  $L \in \mathbb{N}$  such that, after  $L$  steps of this procedure, we obtain consistent a belief-hierarchy distribution  $\tau^L$  that is also minimal. We denote  $e^L := \tau^L$  and our procedure terminates. By construction, we have that for each  $t \in \text{supp } \tau$ , there exists a unique  $\ell \in \{1, \dots, L\}$

such that  $\tau = e^\ell \tau(\text{supp } e^\ell)$ . Therefore,  $\tau = \sum_{\ell=1}^L \alpha_\ell e^\ell$  where  $\alpha_\ell := \tau(\text{supp } e^\ell) > 0$  and  $\sum_{\ell=1}^L \alpha_\ell = \sum_{\ell=1}^L \tau(\text{supp } e^\ell) = 1$ .

Finally, we prove uniqueness. By way of contradiction, suppose that  $\tau$  admits two minimal representations, that is,

$$\tau = \sum_{\ell} \alpha_\ell e^\ell = \sum_k \xi_k \hat{e}^k$$

such that  $e^\ell \neq \hat{e}^k$  for some  $\ell, k$ . This implies that for some  $t \in \text{supp } \tau$  and some  $\ell, k$ , it holds that  $t \in \text{supp } e^\ell \cap \text{supp } \hat{e}^k$  with  $e^\ell \neq \hat{e}^k$ . Two cases are possible:

(i)  $\text{supp } e^\ell \neq \text{supp } \hat{e}^k$

Since  $e^\ell, \hat{e}^k \in \mathcal{C}$ ,  $\text{supp } e^\ell$  and  $\text{supp } \hat{e}^k$  are belief-closed, which in turn implies  $T^{\ell,k} := \text{supp } e^\ell \cap \text{supp } \hat{e}^k$  (nonempty by assumption) is also belief-closed. Therefore, there exists a distribution  $e^{\ell,k}$  supported on  $T^{\ell,k}$  and described by

$$p_{e^{\ell,k}}(\theta, t) := \frac{p_\tau(\theta, t)}{\tau(T^{\ell,k})}$$

for all  $\theta \in \Theta$  and  $t \in T^{\ell,k}$ , which is consistent and  $\text{supp } e^{\ell,k} \subsetneq \text{supp } e^\ell$ . This contradicts the minimality of  $e^\ell$ .

(ii)  $\text{supp } e^\ell = \text{supp } \hat{e}^k$

Since  $e^\ell, \hat{e}^k \in \mathcal{C}$ , there exist common priors  $p$  and  $\hat{p}$  in  $\Delta(\Theta \times T)$  such that  $\text{marg}_T p = e^\ell$  and  $\text{marg}_T \hat{p} = \hat{e}^k$ . Thus,  $\text{supp } \text{marg}_T p = \text{supp } \text{marg}_T \hat{p}$ . This implies  $\text{supp } p = \text{supp } \hat{p}$  (If not, without loss there would exist some  $i, \tilde{t} \in \text{supp } \text{marg}_T p$  and  $\tilde{\theta} \in \Theta$ , such that  $\beta_i(\tilde{\theta}, \tilde{t}_{-i} | \tilde{t}_i) > 0$ , while for all  $i$  and  $t \in \text{supp } \text{marg}_T \hat{p}$ ,  $\beta_i(\tilde{\theta}, t_{-i} | t_i) = 0$ . This would contradict  $\text{supp } \text{marg}_T p = \text{supp } \text{marg}_T \hat{p}$ ). By Propositions 4.4 and 4.5 in Mertens and Zamir (1985), there can be only one common prior with a given finite support in  $\Delta(\Theta \times T)$ , hence  $p = \hat{p}$ . In turn,  $e^\ell = \hat{e}^k$ , which contradicts that  $e^\ell \neq \hat{e}^k$ .  $\square$

Now, we prove linearity of  $w$ . The point is to show that the set of outcomes of a mixture of subspaces of the universal type space can be written as a similar mixture of the sets of outcomes of these respective subspaces.

For any  $\tau', \tau'' \in \mathcal{C}$ , let

$$\begin{aligned} \alpha O_{\Lambda'}(\tau') + (1 - \alpha) O_{\Lambda''}(\tau'') := \\ \{ \alpha \gamma' + (1 - \alpha) \gamma'' : \gamma' \in O_{\Lambda'}(\tau') \text{ and } \gamma'' \in O_{\Lambda''}(\tau'') \}. \end{aligned} \quad (21)$$

**Proposition 4.** *If  $\Sigma^B$  is invariant, then for all  $\tau', \tau'' \in \mathcal{C}$  and  $\tau = \alpha\tau' + (1 - \alpha)\tau''$  with  $\alpha \in [0, 1]$ , and for all  $\Lambda \in \Sigma^B(\tau)$ , there are  $\Lambda' \in \Sigma^B(\tau')$  and  $\Lambda'' \in \Sigma^B(\tau'')$  such that*

$$O_\Lambda(\tau) = \alpha O_{\Lambda'}(\tau') + (1 - \alpha) O_{\Lambda''}(\tau''). \quad (22)$$

*Proof.* Take any  $\tau', \tau'' \in \mathcal{C}$  and  $\alpha \in [0, 1]$  and let  $\tau = \alpha\tau' + (1 - \alpha)\tau''$ . Take any  $\Lambda \in \Sigma^B(\tau)$  and  $\sigma \in \Lambda$ . Define

$$\gamma_\sigma(a, \theta) := \sum_{t \in \text{supp } \tau} \sigma(a|t) p_\tau(t, \theta) \quad \forall (a, \theta), \quad (23)$$

so that  $O_\Lambda(\tau) = \{\gamma_\sigma : \sigma \in \Lambda\}$ . It follows from invariance that  $\Lambda' := \{\sigma|_{\text{supp } \tau'} : \sigma \in \Lambda\}$  is in  $\Sigma^B(\tau')$  and  $\Lambda'' := \{\sigma|_{\text{supp } \tau''} : \sigma \in \Lambda\}$  is in  $\Sigma^B(\tau'')$ . Moreover, notice that

$$\begin{aligned} \{\sigma|_{\text{supp } \tau' \cap \text{supp } \tau''} : \sigma \in \Lambda\} &= \{\sigma'|_{\text{supp } \tau' \cap \text{supp } \tau''} : \sigma' \in \Lambda'\} \\ &= \{\sigma''|_{\text{supp } \tau' \cap \text{supp } \tau''} : \sigma'' \in \Lambda''\}. \end{aligned}$$

Now define

$$\begin{aligned} \gamma'_\sigma(a, \theta) &:= \sum_{t \in \text{supp } \tau'} \sigma|_{\text{supp } \tau'}(a|t) p_{\tau'}(\theta, t) \quad \forall (a, \theta) \\ \gamma''_\sigma(a, \theta) &:= \sum_{t \in \text{supp } \tau''} \sigma|_{\text{supp } \tau''}(a|t) p_{\tau''}(\theta, t) \quad \forall (a, \theta), \end{aligned}$$

so that  $O_{\Lambda'}(\tau') = \{\gamma'_\sigma : \sigma \in \Lambda\}$  and  $O_{\Lambda''}(\tau'') = \{\gamma''_\sigma : \sigma \in \Lambda\}$ . Since  $p_\tau = \alpha p_{\tau'} + (1 - \alpha) p_{\tau''}$ , (23) implies that for all  $\sigma \in \Lambda$ ,  $a$  and  $\theta$

$$\begin{aligned} \gamma(a, \theta) &= \sum_{t \in \text{supp } \tau} \sigma(a|t) (\alpha p_{\tau'}(\theta, t) + (1 - \alpha) p_{\tau''}(\theta, t)) \\ &= \alpha \sum_{t \in \text{supp } \tau'} \sigma|_{\text{supp } \tau'}(a|t) p_{\tau'}(\theta, t) + (1 - \alpha) \sum_{t \in \text{supp } \tau''} \sigma|_{\text{supp } \tau''}(a|t) p_{\tau''}(\theta, t) \\ &= \alpha \gamma'_\sigma(a, \theta) + (1 - \alpha) \gamma''_\sigma(a, \theta). \end{aligned}$$

Hence,  $O_\Lambda(\tau) = \alpha O_{\Lambda'}(\tau') + (1 - \alpha) O_{\Lambda''}(\tau'')$ . □

**Lemma 3.** *The function  $w$  is linear over  $\mathcal{C}$ .*

*Proof.* Take any  $\tau', \tau'' \in \mathcal{C}$  and  $\alpha \in [0, 1]$  and let  $\tau = \alpha\tau' + (1 - \alpha)\tau''$ . By Proposition 4, we know that for all sequences  $(\Lambda_n)$  in  $\Sigma^B(\tau)$ , there exist sequences  $(\Lambda'_n)$  in  $\Sigma^B(\tau')$  and  $(\Lambda''_n)$  in  $\Sigma^B(\tau'')$  such that

$$O_{\Lambda_n}(\tau) = \alpha O_{\Lambda'_n}(\tau') + (1 - \alpha) O_{\Lambda''_n}(\tau'') \quad \forall n.$$

Since  $g$  is linear,

$$\begin{aligned} \sum_{\theta,a} g(O_{\Lambda_n}(\tau))(a, \theta) v(a, \theta) &= \sum_{\theta,a} g(\alpha O_{\Lambda'_n}(\tau') + (1 - \alpha) O_{\Lambda''_n}(\tau''))(a, \theta) v(a, \theta) \\ &= \alpha \sum_{\theta,a} g(O_{\Lambda'_n}(\tau')) v(a, \theta) + (1 - \alpha) \sum_{\theta,a} g(O_{\Lambda''_n}(\tau''))(\theta, a) v(a, \theta). \end{aligned} \quad (24)$$

Choose  $(\bar{\Lambda}_n)$  in  $\Sigma^{\text{B}}(\tau)$  such that  $w(\tau) = \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\bar{\Lambda}_n}(\tau))(a, \theta) v(a, \theta)$ ,  $(\bar{\Lambda}'_n)$  in  $\Sigma^{\text{B}}(\tau')$  such that  $w(\tau') = \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\bar{\Lambda}'_n}(\tau'))(a, \theta) v(a, \theta)$ , and  $(\bar{\Lambda}''_n)$  in  $\Sigma^{\text{B}}(\tau'')$  such that  $w(\tau'') = \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\bar{\Lambda}''_n}(\tau''))(a, \theta) v(a, \theta)$ . By (24) it must be that

$$\begin{aligned} w(\tau) &= \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\bar{\Lambda}_n}(\tau))(a, \theta) v(a, \theta) \\ &\leq \alpha \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\bar{\Lambda}'_n}(\tau')) v(a, \theta) + (1 - \alpha) \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\bar{\Lambda}''_n}(\tau''))(\theta, a) v(a, \theta) \\ &= \alpha w(\tau') + (1 - \alpha) w(\tau''). \end{aligned}$$

Next, choose  $(\Lambda'_n)$  in  $\Sigma^{\text{B}}(\tau')$  and  $(\Lambda''_n)$  in  $\Sigma^{\text{B}}(\tau'')$  such that

$$\begin{aligned} \alpha w(\tau') + (1 - \alpha) w(\tau'') &= \\ \alpha \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\Lambda'_n}(\tau')) v(a, \theta) + (1 - \alpha) \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\Lambda''_n}(\tau''))(\theta, a) v(a, \theta) \end{aligned} \quad (25)$$

and such that

$$\{\sigma' |_{\text{supp } \tau' \cap \text{supp } \tau''} : \sigma' \in \Lambda'_n\} = \{\sigma'' |_{\text{supp } \tau' \cap \text{supp } \tau''} : \sigma'' \in \Lambda''_n\} =: \Lambda'''_n.$$

Note that the above restriction to sequences describing the same behavior over  $\text{supp } \tau' \cap \text{supp } \tau''$  is without loss, because a maximizing designer must be indifferent between their selected outcomes. Therefore, the rhs of (25) can be written as

$$\begin{aligned} \alpha \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\tilde{\Lambda}'_n}(\tilde{\tau}')) v(a, \theta) + (1 - \alpha) \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\tilde{\Lambda}''_n}(\tilde{\tau}''))(\theta, a) v(a, \theta) \\ + \lim_{n \rightarrow \infty} \sum_{\theta,a} g(O_{\Lambda'''_n}(\tau'''))(\theta, a) v(a, \theta) \end{aligned} \quad (26)$$

for some sequences  $(\tilde{\Lambda}'_n)$  in  $\Sigma^{\text{B}}(\tilde{\tau}')$ ,  $(\tilde{\Lambda}''_n)$  in  $\Sigma^{\text{B}}(\tilde{\tau}'')$ , and  $(\Lambda'''_n)$  in  $\Sigma^{\text{B}}(\tau''')$ , where  $\tilde{\tau}'$  is the consistent distribution with support  $\text{supp } \tau' \setminus \text{supp } \tau''$ ,  $\tilde{\tau}''$  is the consistent distribution with support  $\text{supp } \tau'' \setminus \text{supp } \tau'$ , and  $\tau'''$  is the consistent distribution with



support  $\text{supp } \tau'' \cap \text{supp } \tau'$ . By applying invariance (twice),

$$\tilde{\Lambda}_n := \left\{ \sigma : \text{supp } \tau \rightarrow \Delta A \mid \sigma|_{\text{supp } \tau'} \in \tilde{\Lambda}'_n, \sigma|_{\text{supp } \tau''} \in \tilde{\Lambda}''_n \text{ and } \sigma|_{\text{supp } \tau'''} \in \Lambda'''_n \right\}$$

must be in  $\Sigma^{\text{B}}(\tau)$  which, together with linearity of  $g$ , ensures that equation (26) is equal to

$$\lim_{n \rightarrow \infty} \sum_{\theta, a} g(O_{\tilde{\Lambda}_n}(\tau)) v(a, \theta) = w(\tau).$$

Hence, we obtain  $w(\tau) = \alpha w(\tau') + (1 - \alpha)w(\tau'')$ .

□

**Proof of Theorem 1.** Fix a prior  $\mu_0 \in \Delta(\Theta)$  and take any information structure  $(S, \pi)$ . From Proposition 1, it follows that  $(S, \pi)$  induces a consistent belief-hierarchy distribution  $\tau \in \mathcal{C}$  such that  $\text{marg}_{\Theta} p_{\tau} = \mu_0$ . By definition of  $\Sigma^{\text{B}}$  and  $w$ , we have  $V(S, \pi) \leq w(\tau)$  and, thus,  $\sup_{(S, \pi)} V(S, \pi) \leq \sup\{w(\tau) \mid \tau \in \mathcal{C} \text{ and } \text{marg}_{\Theta} p_{\tau} = \mu_0\}$ . Moreover, Proposition 1 also implies that, for  $\tau \in \mathcal{C}$  such that  $\text{marg}_{\Theta} p_{\tau} = \mu_0$ , there exists an information structure  $(S, \pi)$  that induces  $\tau$  and such that  $V(S, \pi) = w(\tau)$ . Therefore,  $\sup_{(S, \pi)} V(S, \pi) \geq \sup\{w(\tau) \mid \tau \in \mathcal{C} \text{ and } \text{marg}_{\Theta} p_{\tau} = \mu_0\}$ . We conclude that

$$\sup_{(S, \pi)} V(S, \pi) = \sup_{\substack{\tau \in \mathcal{C} \\ \text{marg}_{\Theta} p_{\tau} = \mu_0}} w(\tau). \quad (27)$$

By Proposition 3, there exists a unique  $\lambda \in \Delta^f(\mathcal{C}^{\text{M}})$  such that  $\tau = \sum_{e \in \text{supp } \lambda} \lambda(e)e$ . Since  $p$  and  $\text{marg}$  are linear,

$$\text{marg}_{\Theta} p_{\tau} = \text{marg}_{\Theta} p_{\sum_e \lambda(e)e} = \sum_{e \in \text{supp } \lambda} \lambda(e) \text{marg}_{\Theta} p_e.$$

Then, by Lemma 3 and (27), we have

$$\begin{aligned} \sup_{(S, \pi)} V(S, \pi) &= \sup_{\lambda \in \Delta^f(\mathcal{C}^{\text{M}})} \sum_e w(e) \lambda(e) \\ &\text{subject to } \sum_e \text{marg}_{\Theta} p_e \lambda(e) = \mu_0, \end{aligned} \quad (28)$$

which concludes the proof.

□

## B. Appendix of The Application

### B.1. Bandwagon Effect and Extreme Pessimism

We first show that, given the structure of the worst equilibrium, in any optimal distribution with  $m$  hierarchies in total, an agent invests at hierarchy  $t_i$  in the worst equilibrium if and only if  $t_i$  rationalizes  $I$  uniquely based on beliefs of order  $m$  or lower.

#### Lemma 4. (Bandwagon Effect)

Suppose  $\tau^*$  is an optimal belief-hierarchy distribution. Let  $m = \sum_{i=1,2} |\text{supp } \tau_i^*|$ . Then, for all  $i$  and  $t_i \in \text{supp } \tau_i^*$ ,  $\sigma_i^{\text{MIN}}(I|t_i) = 1$  if and only if  $t_i \in \cup_{k=1}^m \rho_i^k$ .

*Proof.* The “if” part follows from (14) and  $\cup_{k=1}^m \rho_i^k \subseteq \cup_{k \geq 1} \rho_i^k$ . The “only if” part we prove by contradiction. Suppose  $\hat{t}_i \in \text{supp } \tau_i^*$  and  $\sigma_i^{\text{MIN}}(I|\hat{t}_i) = 1$ , but  $\hat{t}_i \notin \cup_{k=1}^m \rho_i^k$ . Then, by (14) it must be that  $\hat{t}_i \in \cup_{k \geq m+1} \rho_i^k$ . Notice that  $\rho_i^k \subseteq \rho_i^{k+1}$  for all  $k \geq 1$ . Thus, for any  $t_i \in \rho_i$ , there exists a number  $k^*(t_i) \in \mathbb{N}^+$  such that  $t_i \in \rho_i^{k^*(t_i)}$  and  $t_i \notin \rho_i^{k^*(t_i)-1}$ , where  $\rho_i^0 = \emptyset$ . That is,  $k^*(t_i)$  is the smallest  $k$  such that  $t_i \in \rho_i^k$ . Let  $n \geq 1$  be such that  $k^*(\hat{t}_i) = m + n$ . That is,  $\hat{t}_i \in \rho_i^{m+n}$  and by definition

$$\beta_i^*(\{\theta = 2\} \times T_j|\hat{t}_i) + \frac{1}{3}\beta_i^*(\Theta \times \rho_j^{m+n-1}|\hat{t}_i) > \frac{2}{3},$$

while  $\hat{t}_i \notin \rho_i^{m+n-1}$  and thus

$$\beta_i^*(\{\theta = 2\} \times T_j|\hat{t}_i) + \frac{1}{3}\beta_i^*(\Theta \times \rho_j^{m+n-2}|\hat{t}_i) \leq \frac{2}{3}.$$

This implies  $\beta_i^*(\Theta \times \rho_j^{m+n-1}|\hat{t}_i) > \beta_i^*(\Theta \times \rho_j^{m+n-2}|\hat{t}_i)$ , hence  $\rho_j^{m+n-2} \subsetneq \rho_j^{m+n-1}$ . Therefore, there exists  $\tilde{t}_j \in \rho_j^{m+n-1}$  such that  $\tilde{t}_j \notin \rho_j^{m+n-2}$ , hence  $k^*(\tilde{t}_j) = m + n - 1$ . By the same argument, there exists  $\tilde{t}_i$  such that  $k^*(\tilde{t}_i) = m + n - 2$ , and so on. This process continues for  $m + n - 1$  steps in total, i.e. there needs to be  $\bar{t}_j$  such that  $k^*(\bar{t}_j) = 1$  if  $m+n$  is even or  $\bar{t}_i$  such that  $k^*(\bar{t}_i) = 1$  if  $m+n$  is odd. Hence, there need to be at least  $m + n$  different hierarchies, which contradicts  $m = \sum_{i=1,2} |\text{supp } \tau_i^*|$ .  $\square$

An implication of Lemma 4 is that there is an optimal  $\tau^*$  such that for every  $k = 2, \dots, m$  it holds that either  $\text{supp } \tau_i^* \cap \rho_i^k = \emptyset$  for  $i = 1, 2$  or  $\text{supp } \tau_i^* \cap \rho_i^k = t_i^k$ ,  $\text{supp } \tau_i^* \cap \rho_i^{k-1} = \emptyset$ , and  $\text{supp } \tau_j^* \cap \rho_j^{k-1} = t_j^{k-1}$  for  $j \neq i$ ,  $i = 1, 2$ . Next, we show that it is never optimal to induce a belief hierarchy at which an agent has multiple rationalizable actions. In particular, the hierarchies which rationalize action  $N$  are optimally set to have first-order beliefs of 0, i.e., extreme pessimism.

**Lemma 5. (Extreme Pessimism)**

If  $\tau^*$  is an optimal belief-hierarchy distribution and if there is a BNE  $\sigma$  such that  $\sigma_i(N|t_i) > 0$  for some  $i$  and  $t_i \in \text{supp } \tau_i^*$ , then  $\beta_i^*(\theta = 2|t_i) = 0$ .

*Proof.* Consider a consistent and Bayes plausible optimal minimal distribution  $\tau$  centered at  $\mu > 0$ . First, we argue that, since  $\tau$  is optimal, it must be that for both  $i$  there exists  $t_i \in \text{supp } \tau_i$  such that  $t_i \in \rho_i$ . If not,  $\tau$  would be strictly dominated by  $\hat{\tau} = x \cdot t_{\frac{2}{3}+\varepsilon} + (1-x) \cdot t_0$  with  $x = \min \left\{ \frac{\mu}{\frac{2}{3}+\varepsilon}, 1 \right\}$ , where  $t_{\bar{\mu}}$  is a hierarchy profile at which  $\beta_i^*(\theta = 2|t_i) = \bar{\mu}$  for all  $i$  is common knowledge. Indeed, the designer's expected payoff under  $\hat{\tau}$  in the worst equilibrium is  $x \cdot v(I, I) + (1-x) \cdot v(N, N)$ . Instead, if for some  $i$ ,  $t_i \notin \rho_i$  for all  $t_i \in \text{supp } \tau_i$ , the designer's expected payoff is bounded above by  $x \cdot v(I, N) + (1-x) \cdot v(N, N)$ , which is strictly smaller than the payoff under  $\hat{\tau}$ .

By way of contradiction, suppose that for some agent  $i$  we have  $\tilde{t}_i \in \text{supp } \tau_i$  such that  $\tilde{t}_i \notin \rho_i$  and  $\beta_i^*(\theta = 2|\tilde{t}_i) > 0$ . Therefore, in the worst equilibrium, agent  $i$  will play  $N$  at  $\tilde{t}_i$ , i.e.,  $\sigma_i^{\text{MIN}}(I|\tilde{t}_i) = 0$ .

By Lemma 4, we know that any investing  $t_i$  is in  $\cup_{k=1}^m \rho_i^k$  and there is a unique  $t_i^k \in (\rho_1^k \cup \rho_2^k) \cap \text{supp } \tau_i$ . Consider the distribution  $p'$  obtained from  $p_\tau$  by transferring all the probability mass  $\sum_{t_j} p_\tau(\tilde{t}_i, t_j, \theta = 2)$  as described below. This operation changes the hierarchies in  $\text{supp } p_\tau$  according to  $f_i : \text{supp } \text{marg}_{T_i} p_\tau \rightarrow \text{supp } \text{marg}_{T_i} p'$  for all  $i$ .

- (A) – If  $\text{supp } \tau_i \cap \rho_i^1 \neq \emptyset$ , then  $p'(f_i(t_i^1), f_j(t_j), \theta = 2) = p_\tau(t_i^1, t_j, \theta = 2) + p_\tau(\tilde{t}_i, t_j, \theta = 2)$  for all  $t_j$ ;
- If  $\text{supp } \tau_i \cap \rho_i^1 = \emptyset$  and  $m$  is even, then  $p'(f_i(t_i^{2k}), f_j(t_j^{2k-1}), \theta = 2) = p_\tau(t_i^{2k}, t_j^{2k-1}, \theta = 2) + p_\tau(\tilde{t}_i, t_j^{2k-1}, \theta = 2)$  for all  $k = 1, \dots, m/2$ ;
- If  $\text{supp } \tau_i \cap \rho_i^1 = \emptyset$  and  $m$  is odd, then  $p'(f_i(t_i^{2k}), f_j(t_j^{2k-1}), \theta = 2) = p_\tau(t_i^{2k}, t_j^{2k-1}, \theta = 2) + p_\tau(\tilde{t}_i, t_j^{2k-1}, \theta = 2)$  for all  $k = 1, \dots, (m+1)/2$ , where  $t_i^{m+1}$  is a new hierarchy with  $p'(t_i^{m+1}, f_j(t_j), \theta = -1) = 0$  for all  $t_j$ .
- (B)  $p'(f_i(\tilde{t}_i), f_j(t_j), \theta = 2) = 0$  for all  $t_j$ , and
- (C)  $p'(f_i(t_i), f_j(t_j), \theta) = p_\tau(t_i, t_j, \theta)$  otherwise.

Now, consider the belief-hierarchy distribution  $\tau'$  induced by  $p'$ . By construction of  $p'$ , (A)-(C) ensure that the beliefs of all (relevant) orders have weakly increased for both agents in all hierarchies, except for  $\tilde{t}_i$ . Therefore, if agent  $i$  with hierarchy  $t_i \in T_i \setminus \{\tilde{t}_i\}$  has uniquely rationalizable action  $I$ , and hence plays  $I$  in any equilibrium under  $\tau$ , this also holds for agent  $i$  with hierarchy  $f_i(t_i)$  under  $\tau'$ . On the other hand, agent  $i$  with hierarchy  $\tilde{t}_i$  played  $N$  in the worst equilibrium under  $\tau$ , and

this continues to hold for agent  $i$  with hierarchy  $f_i(\tilde{t}_i)$  under  $\tau'$ . Moreover,  $\tau'$  is necessarily Bayes plausible (since it is induced by  $p'$ ). Finally, again by construction of  $p'$ , probability mass has only been transferred from non-investing to investing types, hence a designer with symmetric and monotone payoffs must have a strictly higher expected payoff under  $\tau'$  than under  $\tau$ , a contradiction.  $\square$

## B.2. Characterization of the Optimal Minimal Distributions

The above lemmas imply that optimal design takes a special form in this example. First, if it is necessary to induce  $N$ , then by Lemma 5 the designer should do it by making the agent extremely pessimistic about the state. In this way, for a given  $\mu$ , the designer can put higher probability on hierarchies that induce  $I$ . Second, Lemma 4 implies that optimal design incentivizes joint investment through contagion, which we call the “bandwagon effect.” To uniquely rationalize joint investment, based on agents’ beliefs of *bounded* order, some hierarchy of an agent must initiate investment by having first-order beliefs greater than  $2/3$ . For any given  $\mu$ , it is never optimal to put mass on more than one such “first-order investor” hierarchy, because we can induce the other agent to invest at lower (than  $2/3$ ) first-order beliefs, if he puts a sufficiently high probability on that first-order investor. Likewise, it is optimal to have only one such “second-order investor”, as we may induce the agent with the first-order investor hierarchy to now invest at a different hierarchy with even lower first-order beliefs, if he puts a sufficiently high probability on that second-order investor hierarchy of his opponent. By proceeding in this way, we generate the lowest possible first-order beliefs overall, and thus generate joint investment with certainty at the lowest possible state distribution  $\mu$ . When joint investment can no longer be sustained, and  $N$  has to played with positive probability, we set the first-order beliefs of the highest-order investor hierarchies to 0, by virtue of Lemma 5.

For an arbitrary  $m < \infty$ , suppose the designer sends  $m$  messages in total in the maximization within. Let us index each hierarchy of agent 1 by  $l = 1, \dots, L$  and of agent 2 by  $r = 1, \dots, R$ . By Lemmas 5 and 4, the optimal minimal distributions will have either  $L = R = \frac{m}{2}$  if  $m$  is even, or  $L = \frac{m+1}{2}$  and  $R = \frac{m-1}{2}$  if  $m$  is odd.

Given a commonly known  $\mu := \text{Prob}(\theta = 2)$ , denote an optimal minimal distri-

bution  $e_\mu^*$  by

$e_\mu^*$	$t_{2,1}$	$\dots$	$t_{2,r}$	$\dots$	$t_{2,R}$
$t_{1,1}$	$A_{11}$	$\dots$	$A_{1r}$	$\dots$	$A_{1R}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{1,l}$	$A_{l1}$	$\dots$	$A_{lr}$	$\dots$	$A_{lR}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{1,L}$	$A_{L1}$	$\dots$	$A_{Lr}$	$\dots$	$A_{LR}$

where all entries  $A_{lr}$  are positive and  $\sum_{l,r} A_{lr} = 1$ . Define  $A_i^1 := \sum_r A_{lr}$  and  $A_r^2 := \sum_l A_{lr}$ . Let  $\mu_1^l := \beta_1^*(\theta = 2|t_{1,l})$  and  $\mu_2^r := \beta_2^*(\theta = 2|t_{2,r})$  denote the agents' first-order beliefs at the respective hierarchies.

For  $\mu > \frac{2}{3}$ , the optimal minimal distribution is given by Table 5, where  $\mu_i^1 = \mu$

$e_\mu^*$	$t_{2,1}$
$t_{1,1}$	1

TABLE 5: Optimal Minimal Distributions for  $\mu \geq 2/3$

for  $i = 1, 2$ . The designer simply lets agents act under common knowledge of  $\mu$ , as investment is uniquely optimal in this range.

For  $\mu \leq \frac{2}{3}$ , the following system yields the optimal minimal distributions  $e_\mu^*$  such that all agents invest at all  $t_i$ :

$$\left\{ \begin{array}{l} \frac{\sum_r A_{1r} \mu^{1r}}{A_1^1} > \frac{2}{3} \end{array} \right. \quad (29)$$

$$\left\{ \begin{array}{l} \frac{\sum_l A_{lr} \mu^{lr}}{A_r^2} + \frac{1}{3} \frac{\sum_{l=1}^r A_{lr}}{A_r^2} > \frac{2}{3} \end{array} \right. \quad \text{for } r = 1, \dots, R \quad (30)$$

$$\left\{ \begin{array}{l} \frac{\sum_r A_{lr} \mu^{lr}}{A_l^1} + \frac{1}{3} \frac{\sum_{r=1}^{l-1} A_{lr}}{A_l^1} > \frac{2}{3} \end{array} \right. \quad \text{for } l = 2, \dots, L \quad (31)$$

$$\left\{ \begin{array}{l} \sum_{l,r} A_{lr} \mu^{lr} = \mu \end{array} \right. \quad (32)$$

where  $\mu^{lr} := \beta^*(\theta = 2|t_{1,l}, t_{2,r})$  denotes the designer's belief that  $\theta = 2$  given the profile of agents' hierarchies  $(t_{1,l}, t_{2,r})$ . We have used Proposition 2 and Lemma 4 to construct this system (29) – (32). Indeed, we have incorporated conditions (7) to (9) from Proposition 2: all agents' first-order beliefs should be a  $\tau(t_{-i}|t_i)$ -weighted

average of the designer's beliefs. For example, in (29)

$$\mu_1^1 := \frac{\sum_r A_{1r} \mu^{1r}}{A_1^1}$$

represents agent 1's (first-order) belief that  $\theta = 2$  at  $t_{1,1}$ . Likewise, the first expressions on the LHS of (30) and (31) represent first-order beliefs  $\mu_2^r$  and  $\mu_1^l$ , respectively. Condition (32) captures (7) from Proposition 2.

Inequalities (29) to (31) are the investment constraints inherited from the equilibrium characterization in (14) and Lemma 4. As discussed above, some agent must have a hierarchy at which he is the first-order investor. Without loss, choose  $t_{1,1}$  to play that role and require  $\mu_1^1 > \frac{2}{3}$  (this is (29)). Choose  $t_{2,1}$  to be the second-order investor, who invests based on second-order beliefs, i.e.,  $t_{2,1} \in \rho_2^2$ . This is equivalent to  $\mu_2^1 + \frac{1}{3}\beta_2^*(t_{1,1}|t_{2,1}) > \frac{2}{3}$ , where

$$\begin{aligned} \mu_2^1 &:= \frac{\sum_l A_{l1} \mu^{l1}}{A_1^2} \\ \beta_2^*(t_{1,1}|t_{2,1}) &:= \frac{A_{11}}{A_1^2}. \end{aligned} \tag{33}$$

This is precisely (30) for  $r = 1$ , where (33) is 2's second-order belief that  $t_1 \in \rho_1^1$ . Given that agent 2 invests based on second-order beliefs at  $t_{2,1}$ , agent 1 invests based on third-order beliefs if  $t_{1,2} \in \rho_1^3$ , which is equivalent to condition (31) for  $l = 2$ , and so on. In conclusion, the system (29)-(32) describes the most efficient way of inducing joint investment with probability 1. If the system has a solution, it must be  $e_\mu^*$ .

*Proof of Claim 1.* “Only if”: By way of contradiction, suppose that  $e_\mu^*$  induces joint investment with certainty at  $\mu \leq 1/2$ . Then,  $e_\mu^*$  must solve (29)-(32). Adding up inequalities (29) and (31) for  $l = 2, \dots, L$ , we obtain:

$$\sum_{l,r} A_{lr} \mu^{lr} > \frac{2}{3} \sum_l A_l^1 - \frac{1}{3} \sum_{l=2}^L \sum_{r=1}^{l-1} A_{lr}$$

which can be rewritten as

$$\mu > \frac{2}{3} - \frac{1}{3} \sum_{l=2}^L \sum_{r=1}^{l-1} A_{lr} \tag{34}$$

using equation (32). Adding up equations (30) for  $r = 1, \dots, R$ , we similarly obtain:

$$\mu > \frac{2}{3} - \frac{1}{3} \sum_r \sum_{l=1}^r A_{lr}. \quad (35)$$

Notice that  $\sum_{l=2}^L \sum_{r=1}^{l-1} A_{lr} + \sum_r \sum_{l=1}^r A_{lr} = \sum_r^R \sum_l^L A_{lr} = 1$ . Therefore, adding up equations (34) and (35) we get  $2\mu > 2 \cdot \frac{2}{3} - \frac{1}{3}$ , which implies  $\mu > \frac{1}{2}$ .

“If”: Suppose that  $\mu > 1/2$ . For  $\mu > 2/3$ ,  $e_\mu^*$  in Table 5 induces joint investment with certainty. For  $\mu \in (1/2, 2/3)$ , it can be easily checked that  $e_\mu^*$  in Table 3 is a solution to the system (29)-(32) with  $R = 1$  and  $L = 2$ , and hence induces joint investment with certainty.  $\square$

From Claim 1 we know that for  $\mu \leq 1/2$ , the designer can no longer ensure  $I$  will be played at all hierarchies. By Lemma 5, the hierarchies at which  $N$  is played must have first-order beliefs of zero. To obtain the optimal minimal distributions for  $\mu \leq 1/2$ , we set

$$\mu_2^R := \frac{\sum_l A_{lR} \mu^{lR}}{A_R^2} = 0 \quad (36)$$

and replace (30) for  $r = R$  with it. We then find the smallest  $\mu$  for which the system has a solution and repeat this procedure by setting  $\mu_1^L = 0$  and replacing (31) for  $l = L$  with it, and so on.

*Proof of Claim 2.* Consider the case when the last  $K$  types of agent 1 and the last  $K$  types of agent 2 do not invest.<sup>18</sup> This means that the system becomes

$$\left\{ \begin{array}{l} \sum_r A_{1r} \mu^{1r} > \frac{2}{3} A_1^1 \end{array} \right. \quad (37)$$

$$\left\{ \begin{array}{l} \sum_l A_{lr} \mu^{lr} > \frac{2}{3} A_r^2 - \frac{1}{3} \sum_{l=1}^r A_{lr} \end{array} \right. \quad \text{for } r = 1, \dots, R - K \quad (38)$$

$$\left\{ \begin{array}{l} \sum_r A_{lr} \mu^{lr} > \frac{2}{3} A_l^1 - \frac{1}{3} \sum_{r=1}^{l-1} A_{lr} \end{array} \right. \quad \text{for } l = 2, \dots, L - K \quad (39)$$

$$\left\{ \begin{array}{l} \sum_l A_{lr} \mu^{lr} = 0 \end{array} \right. \quad \text{for } r = R - K + 1, \dots, R \quad (40)$$

$$\left\{ \begin{array}{l} \sum_r A_{lr} \mu^{lr} = 0 \end{array} \right. \quad \text{for } l = L - K + 1, \dots, L \quad (41)$$

$$\left\{ \begin{array}{l} \sum_{l,r} A_{lr} \mu^{lr} = \mu \end{array} \right. \quad (42)$$

<sup>18</sup>Notice that the only options are that the last  $K$  types of both agents do not invest, or the last  $K$  types of agent 1 and  $K - 1$  types of agent 2 do not invest.

where the first-order beliefs of the last  $K$  non-investing types of both agents have been set equal to zero. Adding up each side of the inequalities corresponding to agent 1 ((37), (39), and (41)) and using (42), we get:

$$\mu > \frac{2}{3} \sum_{l=1}^{L-K} A_l^1 - \frac{1}{3} \sum_{l=2}^{L-K} \sum_{r=1}^{l-1} A_{lr} = \frac{2}{3} \sum_{l=1}^{L-K} \sum_{r=1}^{R-K} A_{lr} + \frac{2}{3} \sum_{l=1}^{L-K} \sum_{r=R-K+1}^R A_{lr} - \frac{1}{3} \sum_{l=2}^{L-K} \sum_{r=1}^{l-1} A_{lr} \quad (43)$$

and correspondingly for agent 2 (adding up each side of (38) and (40)):

$$\mu > \frac{2}{3} \sum_{r=1}^{R-K} A_r^2 - \frac{1}{3} \sum_{r=1}^{R-K} \sum_{l=1}^r A_{lr} = \frac{2}{3} \sum_{l=1}^{L-K} \sum_{r=1}^{R-K} A_{lr} + \frac{2}{3} \sum_{l=L-K+1}^L \sum_{r=1}^{R-K} A_{lr} - \frac{1}{3} \sum_{r=1}^{R-K} \sum_{l=1}^r A_{lr}. \quad (44)$$

Summing up the above two equations we obtain:

$$2\mu > \sum_{l=1}^{L-K} \sum_{r=1}^{R-K} A_{lr} + \frac{2}{3} \sum_{l=1}^{L-K} \sum_{r=R-K+1}^R A_{lr} + \frac{2}{3} \sum_{l=L-K+1}^L \sum_{r=1}^{R-K} A_{lr}, \quad (45)$$

where we have used that:

$$\frac{1}{3} \sum_{l=2}^{L-K} \sum_{r=1}^{l-1} A_{lr} + \frac{1}{3} \sum_{r=1}^{R-K} \sum_{l=1}^r A_{lr} = \frac{1}{3} \sum_{l=1}^{L-K} \sum_{r=1}^{R-K} A_{lr}.$$

Notice that (45) can be written as:

$$2\mu > \Pr(\text{both invest}) + \frac{2}{3} \Pr(\text{only 1 invests}) + \frac{2}{3} \Pr(\text{only 2 invests}). \quad (46)$$

Multiplying both sides by  $v(I, I) > 0$ , we obtain:

$$2\mu v(I, I) > v(I, I) \left[ \Pr(\text{both invest}) + \frac{2}{3} \Pr(\text{only 1 invests}) + \frac{2}{3} \Pr(\text{only 2 invests}) \right]. \quad (47)$$

Notice that the left-hand side of this inequality is  $(\text{cav } \tilde{w}^*)(\mu)$  for  $\mu \leq 1/2$ . Let us denote the right-hand side of (47) by  $\text{RHS}(47)$ . The designer's expected value at any optimal  $e_\mu^* \in \mathcal{C}^M$  with  $m$  messages can be written as

$$\begin{aligned} \mathbb{E}_{e_\mu^*}[v] &= v(I, I) \Pr(\text{both invest}) \\ &\quad + v(I, N) \Pr(\text{only 1 invests}) + v(N, I) \Pr(\text{only 2 invests}). \end{aligned} \quad (48)$$



Then,

$$\begin{aligned} \text{RHS(47)} - \mathbb{E}_{e_\mu^*}[v] &= \left( \frac{2}{3}v(I, I) - v(N, I) \right) \Pr(\text{only 1 invests}) \\ &\quad + \left( \frac{2}{3}v(I, I) - v(N, I) \right) \Pr(\text{only 2 invests}). \end{aligned} \quad (49)$$

Hence, if  $\frac{2}{3}v(I, I) - v(N, I) \geq 0$ ,  $\text{RHS(47)} \geq \mathbb{E}_{e_\mu^*}[v]$ . This in turn implies

$$(cav \tilde{w}^*)(\mu) > \mathbb{E}_{e_\mu^*}[v]$$

for  $\mu \leq 1/2$  and for any optimal  $e_\mu^* \in \mathcal{C}^M$ . □

## References

- Alonso, Ricardo and Odilon Câmara. 2016. “Persuading Voters.” *American Economic Review* 106 (11):3590–3605.
- Angeletos, George-Marios and Alessandro Pavan. 2007. “Efficient Use of Information and Social Value of Information.” *Econometrica* 75 (4):1103–1142.
- Aumann, Robert J. 1976. “Agreeing to Disagree.” *The Annals of Statistics* 4 (6):1236–1239.
- Aumann, Robert J. and Michael B. Maschler. 1995. *Repeated Games with Incomplete Information*. Cambridge, MA: MIT Press.
- Barelli, Paulo. 2009. “On the genericity of full surplus extraction in mechanism design.” *Journal of Economic Theory* 144 (3):1320–1332.
- Bergemann, Dirk, Benjamin Brooks, and Stephen Morris. 2017. “First-Price Auctions with General Information Structures: Implications for Bidding and Revenue.” *Econometrica* 85 (1):107–143.
- Bergemann, Dirk and Stephen Morris. 2016. “Bayes Correlated Equilibrium and the Comparison of Information Structures in Games.” *Theoretical Economics* 11:487–522.
- . 2018. “Information Design: A Unified Perspective.” *Journal of Economic Literature* .
- Best, James and Daniel Quigley. 2018. “Persuasion for the Long-Run.” Oxford University WP.
- Brandenburger, Adam and Eddie Dekel. 1993. “Hierarchies of Beliefs and Common Knowledge.” *Journal of Economic Theory* 59 (1):189–198.
- Brocas, Isabelle and Juan. D Carrillo. 2007. “Influence through ignorance.” *The RAND Journal of Economics* 38 (4):931–947.
- Carlsson, H. and E. van Damme. 1993. “Global games and equilibrium selection.” *Econometrica* 61 (5):989–1018.
- Chan, Jimmy, Seher Gupta, Fei Li, and Yun Wang. 2019. “Pivotal Persuasion.” *Journal of Economic Theory* 180:178–202.
- Dekel, Edie, Drew Fudenberg, and Stephen Morris. 2007. “Interim Correlated Rationalizability.” *Theoretical Economics* 2:15–40.

- Ely, Jeffrey. 2017. “Beeps.” *American Economic Review* 107 (1):31–53.
- Ely, Jeffrey and Marcin Peski. 2006. “Hierarchies of belief and interim rationalizability.” *Theoretical Economics* 1 (1):19–65.
- Heifetz, Aviad and Zvika Neeman. 2006. “On the Generic (Im)Possibility of Full Surplus Extraction in Mechanism Design.” *Econometrica* 74 (1):213–233.
- Hoshino, Tetsuya. 2018. “Using Strategic Uncertainty to Persuade Multiple Agents.” Pennsylvania State University, WP.
- Inostroza, Nicolas and Alessandro Pavan. 2017. “Persuasion in Global Games with Application to Stress Testing.” WP, Northwestern University.
- Kajii, Atsushi and Stephen Morris. 1997. “The Robustness of Equilibria to Incomplete Information.” *Econometrica* 65 (6):1283–1309.
- Kamenica, Emir and Matthew Gentzkow. 2011. “Bayesian Persuasion.” *American Economic Review* 101 (6):2590–2615.
- Laclau, Marie and Ludovic Renou. 2016. “Public Persuasion.” WP.
- Liu, Qingmin. 2009. “On redundant types and Bayesian formulation of incomplete information.” *Journal of Economic Theory* 144 (5):2115–2145.
- . 2015. “Correlation and common priors in games with incomplete information.” *Journal of Economic Theory* 157 (C):49–75.
- Mathevet, Laurent, David Pearce, and Ennio Stacchetti. 2018. “Reputation and Information Design.” WP, New York University.
- Mertens, Jean-Francois and Shmuel Zamir. 1985. “Formulation of Bayesian analysis for games with incomplete information.” *International Journal of Game Theory* 14 (1):1–29.
- Morris, Stephen and Hyun Song Shin. 2002. “Social Value of Public Information.” *American Economic Review* 92 (5):1521–1534.
- . 2003. “Global games: theory and applications.” In *Advances in Economics and Econometrics, the Eighth World Congress*. Cambridge University Press.
- Rayo, Luis and Ilya Segal. 2010. “Optimal Information Disclosure.” *Journal of Political Economy* 118 (5):949 – 987.
- Rockafellar, R. Tyrrell. 1970. *Convex Analysis*. Princeton, New Jersey: Princeton University Press, 2nd ed.

- Rubinstein, Ariel. 1989. "The Electronic Mail Game: Strategic Behavior under "Almost Common Knowledge"." *American Economic Review* 79 (3):385–91.
- Taneva, Ina. 2019. "Information Design." *American Economic Journal: Microeconomics*, Forthcoming.
- Vives, Xavier. 1988. "Aggregation of Information in Large Cournot Markets." *Econometrica* 56 (4):851–876.
- Yildiz, Muhamet. 2015. "Invariance to representation of information." *Games and Economic Behavior* 94 (C):142–156.
- Zhang, Jun and Junjie Zhou. 2016. "Information Disclosure in Contests: A Bayesian Persuasion Approach." *Economic Journal* 126:2197–2217.