Robust information design

V Bhaskar*

September 12, 2025

Abstract

Abstract

We study information design when agents play a supermodular game and the designer prefers higher agent actions. We assume that the designer is slightly uncertain about the exact preferences of the agents. Specifically, we perturb the payoffs of each agent via a continuous payoff shock, as in Harsanyi (IJGT 1973), and focus on *purifiable* equilibria. These are limits of some sequence of equilibria of perturbed games as the shocks vanish. When the designer has no private information, the standard solution is generically purifiable. However, this is no longer true when designer has some private information about the state, and can design experiments that can be arbitrarily correlated with her private information. With binary states or with a perfectly informed designer, in any purifiable equilibrium, the agent learns the private information of the designer, thereby lowering the designer's ex ante payoff. We also introduce the notion of the *experiment-proofness*: an experiment is experiment-proof if, after observing its outcome, the designer does not want to conduct further experiments. With multiple agents, any experiment-proof purifiable equilibrium must be public, so that all agents share the same information. Finally, if the designer is able to conduct private experiments before interacting with the agent, this results in complete unravelling – purifiable experiments are perfectly informative.

*University of Texas at Austin. Thanks to Frederic Koessler, Stephen Morris, Vasiliki Sreta, Max Stichcombe, Tristan Tomala and seminar participants at UT Austin and HEC Paris for very useful comments.

1 The model

Let $\Omega = \{\omega_0, \omega_1, .., \omega_m\}$ denote a finite set of states. Nature chooses a state from Ω according to a probability distribution $\pi := (\pi_0, \pi_1, .., \pi_m)$. The set of agents is $I = \{1, 2, .., |I|\}$. Each agent $i \in I$ has a finite set of actions A_i with n_i elements. We assume that both states and actions are totally ordered by the relation $>$, so that $\omega_{i+1} > \omega_i$ for $i < m$ and $a^i_{j+1} > a^i_j$ if $j < n_i$. Let $A = A_1 \times_2 \times ... A_{|I|}$. We assume that the agent's utility function $u : A \times \Omega \to \mathbb{R}$ is strictly supermodular in (a_i, ω), and weakly supermodular in (a_i, a_{-i}) The principal's payoff function is $\tilde{v} : A \times \Omega \to \mathbb{R}$, and is assumed to be strictly increasing in each a_i at each ω.

As usual, we extend the agent's payoff function u to $A \times \Delta(\Omega)$ by taking expectations, so that $u(a, \mu)$ denotes the agent's payoff from a at belief μ. If the agent has belief μ when he chooses his action, then he chooses an optimal action, breaking any ties in favor of the higher action. This defines the principal's payoff at any final belief, $v(\mu)$.

Given the ordering $>$ on states, we can define the partial ordering $\succ$ on $\Delta(\Omega)$ so that the set $\Delta(\Omega)$ is partially ordered by first-order stochastic dominance. We write $\mu \succ \mu'$ if μ first-order stochastically dominates μ'. Similarly, the set $\Delta(A)$ is partially ordered by first-order stochastic dominance, and for two of its elements α α', we write $\alpha \succ \alpha'$ if α first-order stochastically dominates α'.

The following lemma extends supermodularity to $A \times \Delta(\Omega)$.

Lemma 1 *Suppose that $u : A \times \Omega \to \mathbb{R}$ is strictly supermodular. Then $u : A \times \Delta(\Omega) \to \mathbb{R}$ is strictly supermodular when $\Delta(\Omega)$ is partially ordered by $\succ$, the relation of first order stochastic dominance.*

Suppose that μ' first order stochastically dominates μ, and let $a' > a$. Then, since $u(a', \omega) - u(a, \omega)$ is increasing in ω, and it follows that the expectation of $u(a', \omega) - u(a, \omega)$ with respect to μ' is strictly greater than its expecation with respect to μ. In other words, the expected utility function of the agent, $u(a, \mu)$, is strictly supermodular.

The base model, where the principal has no private information, and where the agent's preferences are not subject to payoff shocks, is that of **?**. The principal has access to a rich set of experiments, that allow her to generate any Bayes plausible distribution of posteriors $\tau \in \Delta(\Delta(\Omega))$. If $\mu = (\mu_i)_{i=0}^m$ is a posterior belief in the support of an experiment τ, the agent chooses action a to maximize $\sum_{i=0}^m u(a, \omega_i)\mu_i$, breaking in ties in favor of the higher action, i.e. by choosing the principal's preferred action. Denote this action by $\hat{a}(\mu)$. This defines $v(\mu) := \tilde{v}(a(\mu)$, the principal's payoff at any induced belief μ. The principal's value function in the game, $V(\pi)$, is given by

the concavification of $v(\mu)$.

We now turn to the perturbed information design problem Γ^k, following **?**. We will need to consider a sequence of random variables, indexed by k, where each term in the sequence satisfies the following properties. Let $\tilde{z}^k$ be a (n)-dimensional random variable, with an atomless distribution F^k, with support on set $Z^k \subset \mathbb{R}^n$. Let $\kappa > 1$ and define $\bar{z} := \kappa \max\{u(a_0, \omega_0) - u(a_n, \omega_0), u(a_n, \omega_m) - u(a_0, \omega_m)\}$. The interpretation is that in the perturbed game, Γ^k, if the agent chooses action a_i, then his payoff his augmented by z_i^k, i-th component of the realization of the random variable.

We make the following assumptions on the payoff shocks, so that for every k, the distribution F^k satisfies:

- Lebesgue measure on the interval $[0, \bar{z}]^n$ is absolutely continuous with respect to F^k.
- F^k is absolutely continuous with respect to Lebesgue measure.[1]
- F^k converges weakly to the point mass on 0, the $n-$dimensional null vector, as $k \to \infty$.

Fix a perturbed game with, Γ^k, where the distribution of shocks is F^k. If the agent has belief μ and shock realization z, then action a_j is better than action a_i if

$$z_j - z_i \geq u(a_j, \mu) - u(a_i, \mu).$$

Thus action a_j is optimal at μ for z-values satisfying:

$$\cap_{k \neq j} \{z : z_j - z_k \geq u(a_j, \mu) - u(a_i, \mu)\}.$$

Let $\alpha(\mu, F_n)$ denote the distribution over actions in Γ^n with shock distribution F^n induced by belief μ – this is unique since there are no mass points in F^n.

Lemma 2 *If $\mu \succ \mu'$, then $\alpha(\mu, F_n) \succ \alpha(\mu', F_n)$.*

Proof. ■

Lemma 3 *Suppose that action a^* is strictly optimal at μ. Then the probability that a^* is played under $\alpha(\mu, F^n)$ converges to one as $n \to \infty$.*

We will present our results in the context of simple example, with binary states and actions. We conjecture that these results extend, with suitable modification, to the more general framework set out at the beginning, but this needs further work.

[1] This assumption does not appear to be essential for our arguments, but simplifies exposition.

2 Uninformed Principal

Let nature choose $\omega \in \{G, B\}$ where the prior probability of G is π. The agent's action set is binary, $\{Y, N\}$. Her payoffs in the base game are 0 if she chooses N. If she chooses Y, her payoff is 1 if $\omega = G$, and $-\ell$ if $\omega = B$. The principal's payoff function is state-independent, and equals 1 if Y is chosen and 0 if N is chosen. Since the agent chooses Y if and only if her belief μ exceeds the threshold μ^*, it follows that the payoff of the principal, v, at any posterior belief μ is given by $v(\mu) = 1$ if $\mu \geq \mu^*$ and $v(\mu) = 0$ if $\mu < \mu^*$. Following **?**, the value function at any prior belief π following her optimal experiment, denoted by $V(\pi)$, equals 1 if $\mu \geq \mu^*$ and $V(\pi) = \frac{\pi}{\mu^*}$ if $\pi < \mu^*$.

At any belief μ, let $A^*(\mu)$ denote the set of payoff maximizing actions for the agent. We assume that the agent chooses the maximal element from A^*, and denote this by $\hat{a}(\mu)$.

The base model, where the principal has no private information, and where the agent's preferences are not subject to payoff shocks, is that of **?**. The principal has access to a rich set of experiments, that allow her to generate any Bayes plausible distribution of posteriors $\tau \in \Delta(\Delta(\Omega))$. If $\mu = (\mu_i)_{i=0}^m$ is a posterior belief in the support of an experiment τ, the agent chooses action a to maximize $\sum_{i=0}^m u(a, \omega_i)\mu_i$, breaking in ties in favor of the higher action, i.e. by choosing the principal's preferred action. Denote this action by $\hat{a}(\mu)$. This defines $v(\mu) := \tilde{v}(a(\mu)$, the principal's payoff at any induced belief μ. The principal's value function in the game evaluated at any prior belief π, $V(\pi)$, is given by the concavification of $v(\mu)$.

Let $A^* \subset A = \{a \in A : \exists \mu \in \Delta(\Omega) : U(a, \mu) \geq U(a', \mu) \forall a' \in A\}$. That is, A^* is the set of actions which are optimal for the agent at some belief.

Assumption 4 *For any action $a \in A^*$ that is optimal at some belief $\hat{\mu}$, there exists an open set O in $\Delta(\Omega)$, with $\hat{\mu}$ on the boundary of O, such that a is strictly optimal at any belief in O.*

The agent's payoff function u is an element in $\mathbb{R}^{mn}$. At any belief μ, the set $\{u \in \mathbb{R}^{mn} : U(a_i, \mu) \geq U(a_j, \mu)\}$ defines a half-space in $\mathbb{R}^{mn}$. Thus a_i is optimal at μ if u belongs to the intersection of these $n-1$ half-spaces. The assumption states that either this intersection is null, or it contains an open set.

Let us now perturb the payoffs of the agent, so her payoff from choosing Y is augmented by η. We consider a sequence of distribution $< F_n >$ with the following properties. For each n, F_n is atomless, and the interval $[-1, 1]$ is absolutely continuous with respect to F_n. Further, as $n \to \infty$, F_n converges to the Dirac distribution on 0. We call the information design problem given distribution F_n the game $\Gamma^n(F_n)$. If we

fix the sequence F_n, we can call this game Γ^n.

Observe that in any perturbed game Γ^n, our assumptions on F_n ensure that at any belief μ, the agent chooses every action with strictly positive probability. Let $\alpha_n(\mu) \in \Delta(A)$ denote the distribution over actions induced in Γ^n at posterior belief μ – since F_n is atomless, this is unique, and also a continuous function of μ. Let $v_n(\mu)$ denote the expected utility of the principal in Γ^n when the agent's posterior belief equals μ – this equals the expectation $\tilde{v}(a)$ when a has distribution α_n, and is also continuous as a function of μ. Consequently, the principal's payoff is continuous, as function of the induced belief distribution, $\tau \in \Delta(\Delta(\Omega))$. Since the set of feasible belief distributions that the principal can induce is compact, there exists an optimal choice for the principal. In other words, an equilibrium exists in the game Γ^n. Furthermore, every equilibrium must induce the same value for the principal, since otherwise, it would not be optimal for the principal. Hence we can talk in terms of the principal's value function. Let $V_n(\mu; F_n)$ denote the value function of the principal in Γ^n.

Lemma 5 *If $\mu \succ \mu'$, then $V_n(\mu; F_n) > V_n(\mu'; F_n)$.*

The underlying expected utility function in the unperturbed game, $U(a, \mu)$ is strictly supermodular. Consequently, if $\mu \succ \mu'$, the action distribution at μ FOSD the action distribution at μ'.

Note that the perturbed game is an information design problem with a privately-informed receiver. This is, in general, considerably more complex, and harder to characterize, than the case without private information for the receiver – see **?**. However, our focus is on the limit case, where the private information is vanishingly small. This proves to be more tractable.

Definition 6 *$V^*(\pi)$ is a robust value function if for each π and every sequence $< F_n >$ satisfying the above conditions, $\lim_{n\to\infty} V_n(\pi; F_n) = V^*(\pi)$.*

Proposition 7 *If $\mu^* < 1$, then $V^*(\pi) = V(\pi)$ for any π in $\Delta(\Omega)$ .e. the robust value function exists and coincides with the value function in the absence of payoff shocks at every belief.*

Proposition 8 *If assumption 4 is satisfied then $V^*(\pi) = V(\pi)$ for any π in $\Delta(\Omega)$ .e. the robust value function exists and coincides with the value function in the absence of payoff shocks at every belief.*

Proof. Fix π and Let $\psi : \Omega \to \Delta(S)$ denote an optimal experiment for the principal that satisfies the following properties. Without loss of generality, the set of signals

equals $\{a^1, a^2, .., a^K\}$, the set actions induced by the principal, that results in value $V(\pi)$. Furthermore, each induced action is taken only at a single belief. Thus we may associate the beliefs $\{\mu^1, \mu^2, .., \mu^K\}$ with each of the corresponding actions, and these beliefs arise with the probabilities $(\tau^1, .., \tau^K)$, so that $\sum_1^K \mu^i \tau^i = \pi$. Given $\delta > 0$, ee construct a new experiment $\psi(\delta)$, with support on the set of beliefs $\{\tilde{\mu}^0, \tilde{\mu}^1, \tilde{\mu}^2, .., \tilde{\mu}^K\}$, as follows. If the agent has a strict preference for a^i at μ^i, then for each ω, the probability of signal a^i is unaltered, so that $\tilde{\mu}^i = \mu^i$ and $\tilde{\tau}^i = \tau^i$. Suppose now that the agent has a weak preference for a^i at μ^i, and is indifferent between a^i and some a^j. Assumption 4 implies that the belief μ^i cannot be degenerate, and must assign positive probability to two or more states. Assumption 4 implies there exists a belief $\tilde{\mu}^i$ that is within δ distance of μ^i, where the agent has a strict preference for a^i. Furthermore, since beliefs only depend on the relative probabilities of different states, it is possible to induce $\tilde{\mu}^i$ entirely by reducing the probability of signal a^i at one (or more) of the states, and by assigning this probability to signal a^0 at these states. Let $\tilde{\mu}^0$ denote the belief induced at signal a^0.

In the game Γ^n with payoff shocks, let the principal choose the experiment $\psi(\delta).In the unperturbed game \Gamma$, for any $i > 0$, a^i is strictly optimal at $\tilde{\mu}^i$. Thus, as $F^n \to 0$, for any $i > 0$, the probability that a^i is played at $\tilde{\mu}^i$ converges to 1. As δ was arbitrary, we can let $\delta \to 0$, so that as $F^n \to 0$, the joint distribution over $A \times \Omega$ induced by the agent's actions converges to that in Γ. Thus, $V_n(\pi) \to V(\pi)$.

Finally, the convergence is uniform: in the perturbed game, the principal's maximum value function exists, since the principal's payoff is continuous in τ, and the feasible set is compact. The value function, $V_n(\pi)$ is continuous in π. Since we have a sequence of continuous functions converging to $V(.)$, convergence must be uniform.

■

First, we show that assumption 4 implies that if the agent is indifferent between a^i and some other action a' at belief μ^i, then μ^i must assign positive probability to two or more states. Suppose that agent is indifferent between a^i and a' at ω_i. Assumption 4 requires that at any interior belief μ that is close to assigning probability 1 to $\omega, U(a^i, \mu) > U(a', \mu)$ and $U(a^i, \mu) < U(a', \mu)$, a contradiction.

For each belief μ^i where the agent is indifferent between a^i and some other action, construct a new belief $\tilde{\mu}^i$ by reducing the probability of the lowest state in the support of μ^i by $\epsilon > 0$. Since the number of states is finite, ϵ can be chosen small enough such that this is feasible for each such i. For any belief μ^i where the agent strictly prefers a^i, let $\tilde{\mu}^i = \mu^i$. Construct a new belief $\tilde{\mu}^0$ by pooling all the probabilities from the ϵ reductions.

Suppose that $\pi > \mu^*$, and suppose that the principal provides no additional infor-

mation in game $\Gamma^n(F_n)$. The agent's payoff from choosing Y equals $\pi-(1-\pi)\ell+\eta$, while her payoff from choosing N is zero. Thus the agent chooses N if $\eta < (1+\ell)(\mu^* - \pi) < 0$ (since $\pi > \mu^*$). Thus $F_n((1+\ell)(\mu^* - \pi)) \to 0$ as $n \to \infty$, and $V_n(\pi) \to 1$ as $n \to \infty$.

Suppose that $\pi \leq \mu^*$. Pick $\tilde{\mu} = \mu^* + \epsilon$, where ϵ is a small positive number. Let the principal conduct a binary experiment, which induces beliefs 0 and $\tilde{\mu}$. By the same argument as in the previous paragraph, the principal's payoff, conditional on inducing belief $\tilde{\mu}$, converges to 1 as $n \to \infty$. Thus the principal's payoff, converges to $\frac{\pi}{\tilde{\mu}}$. By choosing ϵ small enough, we can approximate $\frac{\pi}{\mu^*}$ arbitrarily closely.

The preceding arguments show that $V^*(\pi) \geq V(\pi)$ for every π. Since 1 is an upper bound on the value function in any game, this establishes that $V^*(\pi) = V(\pi)$ for $\pi > \mu^*$. If $\pi \leq \mu^*$, observe that at any belief $\mu < \mu^*$, the principal's payoff converges to zero as $n \to \infty$. Let θ_n denote the probability of the event $\{\mu \geq \mu^*\}$ under optimal information design in the game $\Gamma^n(F_n)$. Bayesian consistency implies that $\theta_n \leq \frac{\pi}{\mu^*}$ for any n. Thus $V^*(\pi) \leq \frac{\pi}{\mu^*}$.

The following example shows why assumption 4 is required. Let $\Omega = \{0,1\}, A = \{0, \frac{1}{2}, 1\}$. $u(0,0,) = u(1,1) = 1, u(1,0) = u(0,1) = 0$. $u(\frac{1}{2},0) = u(\frac{1}{2},1) = 0.5$.. The principal payoff depends only on a and equals 1 if $a = 0.5$, and 0 otherwise. Let μ denote the (posterior) probability of state 1. For $\mu \in [0.5) \cup (0.5,1], v(\mu) = 0$, while $v(0.5) = 1$. Thus $V(\pi) = 2\pi$ for $\pi \leq 0.5$ and $V(\pi)$ =2(1-π) if $\pi > 0.5$. With payoff shocks, $V(\pi)$ cannot be approximated – for example, if the payoff shock for each action is iid, then the probability of action 0.5 at any belief is bounded above by $\frac{1}{3}$.

assumption that $\mu^* < 1$ is satisfied for generic payoffs of the agent. Indeed, $\mu^* = 1$ can only arise if the action Y is weakly dominated for the agent, and only optimal at state G. It is also necessary. If $\mu^* = 1$, then $V(\pi) = \pi$, and the optimal experiment is fully revealing. However, with payoff shocks, the probability that the agent chooses Y when she knows that the state is G equals $1 - F_n(0)$. Hence, if $\lim_{n\to\infty} F_n(0)$ need not, in general, exist. If it exists and equals $K > 0$, then $V^*(\pi) = \pi(1-K)$, which is strictly less than π.

3 Informed Principal

Now consider the informed principal. To model the principal's information, and her set of available experiments, we will use the framework suggested by **?** and **?**. An experiment ξ is a Lebesgue measurable mapping, $\xi : \Omega \times [0,1] \to S$, where S is a finite space of signal realizations.. The interpretation is as follows. At the outset, before any experiment is chosen, nature determines the realization $x \in [0,1]$ of a single uniformly

distributed random variable. If ξ is conducted, and state ω is realized, the realization x of the uniform random variable determines the signal realization $\xi(\omega, x)$. Note that x is not observed by any decision maker. However, the signal realization $\xi(\omega, x)$ is observer by every observer of the experiment. The probability that signal realization $s \in S$ is observed when the state is ω equals $\lambda\{x : \xi(\omega, x) = s\}$, where λ denotes Lebesgue measure on the line. Consequently, Bayesian updating implies that the beliefs of any observer of the experiment are

$$\mu(\omega|s) = \frac{\pi(\omega)\lambda\{x : \xi(\omega, x) = s\}}{\sum_{\omega' \in \Omega} \pi(\omega')\lambda\{x : \xi(\omega', x) = s\}}. \tag{1}$$

To proceed, we shall make the following assumption.

Assumption 9 *It is common knowledge that the principal gets her information via a known experiment $\hat{\xi}$, which has a finite set of signal realizations.*

One immediate consequence of assumption 9 is that the principal can verifiably disclose her private information, by conducting publicly the experiment $\hat{\xi}$. That is, if realization of the random variable happens to be some x that determines the signal s that the principal observes, then since x is unchanged, the same signal s will be realized when the experiment is repeated publicly. Let $\nu(s|\omega) := \lambda\{x : \xi(\omega, x) = s\}$. Let $\nu(s) = \sum_{\omega \in \Omega} \nu(s|\omega)$.

To summarize, the extensive form game Γ is as follows:

- Nature chooses $\omega \in \Omega$ according to π.
- The experiment $\hat{\xi}$ is conducted, and the principal observes the signal realized from this experiment. The agent only knows that $\hat{\xi}$ has been conducted.
- The principal chooses a public experiment ξ. In particular, ξ may be a compound experiment (ξ_1, ξ_2), where ξ is a public experiment with signal realization space S_1, and ξ_2 is a mapping from S_1 to the set of possible experiments. The outcome of experiment ξ is publicly observed.
- The agent chooses an action in A.

Let $\mathcal{M} = \{\mu_1, \mu_2, ..., \mu_K\}$ denote the set of private principal beliefs induced by the experiment ξ. For our main results, we will make the following assumption:

Assumption 10
1. *The beliefs in $\mathcal{M}$ are totally ordered by $\succ$, the relation of first order stochastic dominance.*
2. *The principal's ordinal preference ordering over the agent's actions coincides with the agent's ordering of actions, i.e. ϕ is the identity permutation.*

3.1 The pooling solution

We may, without loss of generality, assume that the experiment $\hat{\xi}$ has the following structure. Under the experiment $\hat{\xi}$, the principal observes an informative signal $s \in \{g, b\}$, where g occurs with probability $\hat{p}$ when $\omega = G$ and with probability $\hat{q} < \hat{p}$ when $\omega = B$. Without loss of generality, we can assume that $\hat{\xi}(G, x) = g$ if $x \leq \hat{p}$ and $\hat{\xi}(B, x) = g$ if $x \leq \hat{q}$.

We now show that there is a pooling equilibrium, as in **?** that is Pareto-efficient from the point of view of the different types of principal, and uniquely so. Our main result in this subsection is the following proposition. Recall that $V(\mu)$ denotes the principal's value in the symmetric information information design problem when the common prior belief equals μ.

Proposition 11 *There exists a pooling equilibrium where both types of principal offer the same experiment. Under this equilibrium:*

- *The payoff of type g of principal equals $V(\mu_g)$.*
- *The payoff of type b of principal is (weakly) greater than $V(\mu_b)$.*
- *The ex ante value of the principal, i.e. before experiment $\hat{\xi}$ is conducted, equals $V(\pi)$, the value that the principal can earn in the absence of private information.*
- *The equilibrium is the unique Pareto efficient equilibrium from the point of the principals.*

We now describe the equilibrium and prove the above proposition.

If the prior belief π exceeds μ^*, the equilibrium has both types of principal offering the null experiment, and getting their maximum payoff, 1. Clearly, neither type of principal has an incentive to deviate. So let us focus on the case when $\pi < \mu^*$. Define the two thresholds $\tilde{q}_g$ and $\tilde{q}_b$ so that they satisfy the following equations:

$$\frac{\pi\hat{p}}{\pi\hat{p} + (1-\pi)\tilde{q}_g} = \mu^*.$$

$$\frac{\pi(1-\hat{p})}{\pi(1-\hat{p}) + (1-\pi)(1-\tilde{q}_b)} = \mu^*.$$

The experiment $\tilde{\xi}$ is defined as follows:

- The signal space is binary, $\{\gamma, \beta\}$.
- $\tilde{\xi}(G, x) = \gamma$ for all x.
- $\tilde{\xi}(B, x) = g$ if $x \in [0, \tilde{q}_g] \cup [\tilde{q}_b, 1]$.

Consider the following equilibrium. Both types of principal offer the experiment $\tilde{\xi}$. Thus, the agent's interim belief on ω, after this offer, but before the results of the experiment, equals the prior π. If the agent observes signal γ, his posterior belief on ω equals μ^*. If he observes β, his posterior belief is zero. The three claims in the proposition, regarding payoffs, are straightforward to verify.

To complete the description of the equilibrium, we need to specify the beliefs of the experiment when the principal deviates and offers some experiment $\xi \neq \tilde{\xi}$. Let us assume that signal s is realized when ξ is conducted. Suppose that for some $\omega \in \Omega$, the set $\{x : \hat{\xi}(x, \omega) = b\} \cap \{x : \tilde{\xi}(x, \omega) = s\}$ has positive Lebsegue measure. Then the agent's believes that the principal has privately observed signal b, and conducts his Bayesian update on the state ω conditional on the joint event that experiment $\hat{\xi}$ has yielded signal b and experiment $\tilde{\xi}$ has yielded signal s. On the other hand, if the set $\{x : \hat{\xi}(x, \omega) = b\} \cap \{x : \tilde{\xi}(x, \omega) = s\}$ is of zero-measure, both when $\omega = G$ and $\omega = B$, the agent conducts his Bayesian inference on ω conditional only on ξ yielding signal s.[2] In other words, the agent is suspicious of any deviation ξ, and assumes that it is the principal with a bad signal who has deviated, unless the result of the experiment ξ shows that it must have come from the principal with a good signal.

Under this assumption on beliefs, let us first verify that type b of principal has no incentive to deviate. Consider a deviation to an arbitrary experiment ξ'. Let s be an arbitrary signal that arises with positive probability when the experiment ξ' is conducted. Now, if the intersection of the sets $\{x : \exists \omega : \xi'(\omega, x) = s\}$ and $\{x : \exists \omega : \hat{\xi}(\omega, x) = b\}$ is null, then signal s cannot arise when type b chooses ξ'. Consequently, we may disregard any such signal. In other words, the only relevant signals when b conducts ξ' are such that the agent's Bayesian update is conducted on the assumption that the agent is of type b. However, in this case, the maximal payoff that type b can get is $V(\mu_b)$. However, since the payoff of type b under $\tilde{\xi}$ is greater, type b has no incentive to deviate.

Turning to the principal of type g, observe that if $\mu_g \geq \mu^*$, she gets her maximal possible payoff of 1, and hence has no incentive to deviate. So assume $\mu_g < \mu^*$, where the payoff under $\tilde{\xi}$ for this type is strictly less than one. Observe that this principal can infer, from observing g that the either $\omega = G$ and $x \leq \hat{p}$ or $\omega = B$ and $x \leq \hat{q}$. Consequently, under $\tilde{\xi}$, this principal gets a payoff of 0 only when $\omega = B$ and $x \in (q_g, \hat{q}]$. To increase her payoff, the principal conduct an experiment where the agent is induced to choose Y at some signal s that arises when $\omega = B$ and $x \in X$, where the set $X \subset [0, \hat{q}]$ and $\lambda(X) > q_g$. However, if $\lambda(X) > q_g$, then the only way in which the

[2]This is equivalent to the agent conducting his Bayesian update assuming that $\hat{\xi}$ has yielded g and $\tilde{\xi}$ has yielded s.

agent's belief following s can be weakly greater than μ^* is if s also arises when $\omega = G$ for a set of $x-$values that has Lebesgue measure greater than $\hat{p}$. But in this case, the agent believes that the principal is of type b, and so her payoff can be no greater than $V(\mu_b) < V(\mu_b)$.

The final part, that the pooling equilibrium gives us the unique Pareto-efficient point from the point of view of the different types of principal, is a consequence of the following proposition.

Proposition 12 *In any equilibrium of the informed principal game, the interim payoff of type g is no greater than $V(\mu_g)$. If $V(\mu_g) < 1$, then in any equilibrium, the payoff of type b is no greater than $V(\mu_b)$.*

Proof. If $\mu_g \geq \mu^*$, then $V(\mu_g) = 1$, the maximal feasible value. So assume $\mu_g < \mu^*$. In this case, the pooling equilibrium gives payoffs ($V(\mu_g)$ and $V(\mu_b)$ respectively. Each type can guarantee this payoff by choosing the transparent solution that is defined below, and this provides a lower bound on equilibrium payoffs to the two types. Consequently, if the payoff of any type s is strictly greater than $V(\mu_s)$, the ex ante value of the principal is strictly greater than $V(\pi)$, which is impossible given that $V(\pi)$ is generated by the ex ante optimal experiment.

■

3.2 The transparent solution

There always exists a separating equilibrium in the informed principal problem, which we call the transparent solution. We define the transparent solution as follows.

- The principal publicly conducts experiment $\hat{\xi}$.
- If signal s is realized, and induces belief $\mu(s)$, the principal continues with an optimal **?** experiment at this belief.
- The principal's interim value function is given by $V(\mu(s))$.
- The principal's ex ante value value function equals $\tilde{V}(\pi) = \sum_{\omega\in\Omega}\sum_{s\in S}\pi(\omega)\nu(s|\omega)V(\mu(s))$. In general, $\tilde{V}(\pi) \leq V(\pi)$, and the inequality can often be strict.

Observe that beliefs on path are moot in the transparent solution, since the principal fully discloses her private information. As before, the agent's beliefs after observing a deviation are that this comes from the principal of type b, unless the deviating experiment proves this to be impossible. The proof that any deviation is unprofitable mirrors that of the proof of proposition 11.

The following example provides an instance where $\tilde{V}(\pi) < V(\pi)$. Suppose that the experiment $\hat{\xi}$ induces beliefs $\hat{\mu}_H$ and $\hat{\mu}_L$, where $\hat{\mu}_H > \mu^* > \hat{\mu}_L$, with probabilities θ and $1-\theta$. Bayesian consistency requires that $\theta\hat{\mu}_H + (1-\theta\hat{\mu}_L = \pi$. Thus, $V(\hat{\mu}_H) = 1$ and $V(\hat{\mu}_L) = \frac{\hat{\mu}_L}{\mu^*}$, and $\tilde{V} = \theta + (1-\theta)\frac{\hat{\mu}_L}{\mu^*}$. On the other hand, $V(\pi) = \min\{\frac{\pi}{\mu^*}, 1\}$, which is strictly greater.

4 Robustness with an informed principal

We now considered the perturbed information design problem. That is, we augment the model of the informed principal in the previous section, Γ, by adding private payoff shocks (F_n) for the agent. The consequent game, Γ_n, is an information design problem has two sided private information – on the part of the principal, and on the part of the agent. Further, we maintain the genericity assumption that $\mu^* < 1$, that was essential for the symmetric information case. Our focus is on the limit outcomes when F_n converges to the Dirac distribution on 0. A robust equilibrium of the game Γ is the limit of a sequence of equilibria of the games Γ^n as $n \to \infty$.

Existence of a robust equilibrium is more complex, since the principals are no longer just solving an optimization problem. In particular, the agent's beliefs can change discontinuously with the principal's choice of experiment. Our proof of existence is constructive – essentially we will show that the transparent solution is robust.

We assume that the principal gets her private information via the experiment $\hat{\xi}$, which induces signals $\{\hat{s}_1, ..., \hat{s}_T\}$, and corresponding beliefs $\{\mu_1, \mu_2, ..., \mu_T\}$.

We assume that these beliefs are ordered according to first order stochastic dominance, i.e. $\mu_{i+1} \succ \mu_i$ for all $i < T$.

Consider the principal with signal $\hat{s}_T$.

We will show that robust equilibrium is essentially unique, and thus there is a unique robust value function, i.e. the ex ante value of the principal in the robust equilibrium, before she observes her private information. We denote this by $V^\dagger(\pi)$.

Our main result is the following proposition.

Proposition 13 *There is a (essentially) unique robust equilibrium of the informed principal game* Γ.

- $V^\dagger(\pi) = \tilde{V}(\pi)$, *i.e. the robust value function coincides with the transparent solution.*
- $V^\dagger(\pi) \leq V(\pi)$, *where the inequality can be strict, so that the payoff of the pooling solution is not robust.*

.

Main theorem.

Recall that a principal type is a belief, μ_k. Any equilibrium induces a distribution τ over the final beliefs, ν, of the agent, which must be Bayes-plausible. Let $\{\nu_1, \nu_2, ..., \nu_L\}$ denote the induced beliefs. The Bayes-plausibility condition is

$$\sum_{k=1}^{L} \tau(\nu_k)\nu_k = \pi.$$

Let τ_j denote the conditional distribution over the agent's final beliefs, given principal belief μ_j. Thus, $\tau_j(\nu_k)$ denotes the probability that the agent has belief ν_k, conditional on principal belief μ_j. Let $\mathbb{E}(\nu|\mu_j) = \sum_{k=1}^{L} \tau_j(\nu_k)\nu_k$ denote the expected belief of the agent given that the principal's belief type is μ_j.

Theorem 14 *Consider an informed principal, where assumption 10 is satisfied. In any robust Perfect Bayesian equilibrium of the informed principal game,* $\mathbb{E}(\nu|\mu_j) = \mu_j$ *for every* j.

Proof. Consider any perturbed game. We show that $\mathbb{E}(\nu|\mu_j) = \mu_j$ for every j. We show that Since the marginal distribution of the agent's beliefs in any PBE must be Bayes-plausible, if $\mathbb{E}(\nu|\mu_k) > \mu_k$ for some k, there must be some $j : \mathbb{E}(\nu|\mu_j) < \mu_j$. Suppose that principal type μ_j conducts the experiment ξ, thereby revealing her private information. She can follow this with an experiment $\hat{\tau}_j$, which first order stochastic dominates τ_j. This follows from the fact that $\sum_{k=1}^{L} \hat{\tau}_j\hat{(}\nu_k)\hat{\nu}_k = \mu_j$, thereby increasing her payoff. Since $\mathbb{E}(\nu|\mu_j) = \mu_j$ for every j in every perturbed game, the theorem follows. ■

First, we show that transparent solution is robust, i.e. the payoffs in the transparent solution for the two types of prinicipal, $V(\mu_g)$ and $V(\mu_b)$, can be approximated arbitrarily closely as F_n converges to the Dirac distribution. This follows from the proof of proposition 8. In the perturbed game, for each $s \in \{g, b\}$, the principal of type s can ensure herself a payoff that is within ϵ of $V(\mu_s)$ when n is sufficiently large, by choosing the transparent solution. Thus, in any robust equilibrium, the payoff of type s is no smaller than $V(\mu_s)$. Now, if $\mu_g \leq \mu^*$, then proposition 12 implies that this is the unique robust equilibrium payoff.

So let us turn to the case where $\mu_g > \mu^*$. For the principal of type g, we will show that the pooling solution is dominated by the transparent solution in any perturbed game. Consider an arbitrary perturbed game Γ^n. Suppose that the agent has belief μ

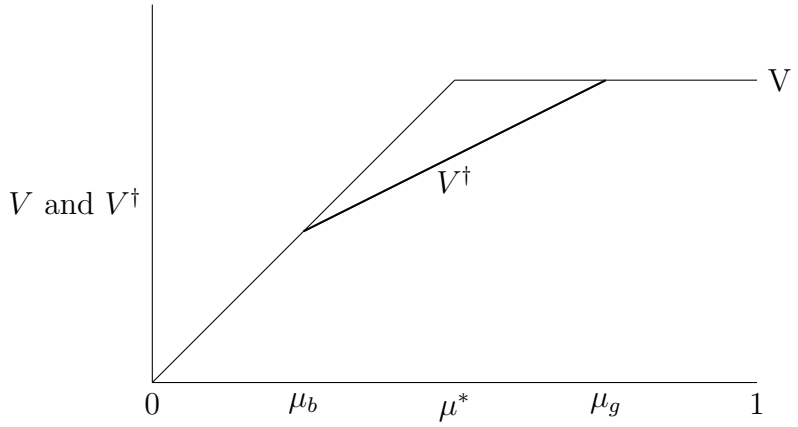
when he takes his action. Choosing Y is optimal if

$$\mu - (1 - mu)\ell + \eta \geq 0.$$

Thus the agent chooses Y with probability $F_n(\ell - \mu(1 + \ell)$. This probability is strictly increasing in μ. Consequently, the principal of type g is strictly worse off, in any perturbed game, by inducing a belief $\mu < \mu_g$.

■

The following figure illustrates our results. It plots the principal's ex ante value without private information, V, which would also be her value function as an informed principal in the pooling equilibrium. More importantly, the bold line plots the principal's *robust* value function, as a function of her prior, when she has private information, $V^\dagger$. This is drawn on the assumption that the principal's interim beliefs have a two point support, μ_g and μ_b, and consequently, the prior must lie in the interval spanned by these two beliefs. Thus, an increase in the prior corresponds to a greater probability of signal g.



There are several immediate implications of our result. First, the principal is always worse off, at least weakly, if the experiment $\hat{\xi}$ is more informative. Second, the extension to multiple signals is straightforward, and robustness implies the transparent solution.

Supermodular example

$$u(a, \omega) = a\omega - .5betaa^2$$

$$\omega = \beta a.$$

4.1 3 state example

Let us consider the following example. There are three equi-probable states of the world and three actions for the agent, whose payoffs are given in the table below. The principal has state independent preferences, with payoffs 0 from L, 3 from M and 4 from H. In the absence of any information, the agent would take action M.

	ω_1	ω_2	ω_3
L	3	2	0
M	2	3	2
H	0	2	3

Given the very low payoff to action L and the prior, the principal seeks to induce actions H and M. TLet λ_i denote the probability mass "transported" to signal s_H, the signal that recommends H. Clearly, each λ_i must be less than $\frac{1}{3}$. thus the principal seeks to maximize $\lambda_1+\lambda_2+\lambda_3$ subject to the two incentive compatibility conditions (that both H and M should be optimal when recommended). Optimality of action H at s_H requires:

$$\lambda_3 \geq \lambda_2 + 2\lambda_1. \tag{2}$$

Optimality of action M at s_M requires:

$$1 - \lambda_1 \leq (1 - \lambda_2) + 2(1 - \lambda_3). \tag{3}$$

Let us now analyze the ex ante optimal experiment. Let x_i denote the probability mass in state ω_i where action H is recommended. Similarly, let y_i (resp. z_i) denote the probability mass in state ω_i where action M (resp, L) is recommended. Thus $x_i+y_i+z_i = \frac{1}{3}, for$i=1,2,3.

The optimal ex ante experiment is the following: $\lambda_3 = 1, \lambda_2 = \lambda_1 = \frac{1}{3}$. Thus $1 - \lambda_2 = 1-\lambda_1 = \frac{2}{3}$. This experiment induces the belief $(\frac{3}{5}, \frac{1}{5}, \frac{1}{5})$, inducing H – this event occurs with probability $\frac{5}{9}$. With the remaining probability, it induces the belief $(0, \frac{1}{2}, \frac{1}{2})$, inducing M. Let us compare this with the optimal experiment that follows full revelation. With full revelation, when the state is revealed to be ω_2, there is no further information to be revealed, and the action M is taken. When the state is in $\{\omega_1, \omega_3\}$, then the optimal experiment induces beliefs $(\frac{2}{3}, 0, \frac{1}{3})$ and $(\frac{1}{3}, 0, \frac{2}{3})$, inducing actions H and M respectively, with equal probabilities.

Let us establish that both types of principal are strictly better off under the ex ante optimal experiment. This is clearly the case for type ω_2, since he now induces H with positive probability. For the type $\neg\omega_2$, the conditional probability that H is played is $2_{\overline{3}}$, which is greater than one-half.

Consequently, if the principals can commit to no further experiments after the first one is conducted, the ex ante optimal experiment is an equilibrium. However, when M is induced, the agent knows that type ω_1 is not possible. Then type ω_2 can benefit in the perturbed game by revealing his type.