Information Design for Prosocial Behavior

ALEXANDRE REIFFERS-MASSON, IMT Atlantique, France
RAJESH SUNDARESAN, Indian Institute of Science, India

We study the idea of information design for inducing prosocial behavior. As an example, we ground our study in the context of electricity consumption. Electricity utility providers would like to reduce the total power consumed by their residential consumer segment. Supply to this segment is often subsidized, and the saved power can be diverted to more profitable segments. Alternatively, the provider may be keen on earning carbon credits by inducing reduced consumption in this segment. The societal network in which the consumers reside may have a prevailing norm, for example, saving power is environment friendly and is considered good. Those that consume less are considered more prosocial and may derive a larger reputational benefit. How can the service provider, who is familiar with the prevailing norm and the consumption of all users, design suitable feedback signals that exploit reputation benefits to reduce net consumption? We call this a problem of information design and address this question in this paper. We consider a continuum of agents. Each agent has a different intrinsic motivation to reduce her power consumption. Each agent models the power consumption of the others via a distribution. Using this distribution, agents will anticipate their reputational benefit and choose a power consumption by trading off their own intrinsic motivation to do a prosocial action, the cost of this prosocial action and their reputation. We assume that the service provider can provide two types of quantized feedbacks of the power consumption. We study their advantages and disadvantages. For each feedback, we characterize the corresponding mean field equilibrium, using a fixed point equation. Besides computing the mean field equilibrium, we highlight that, in some situations, revealing less information can lead to more prosociality. We illustrate the result on London smart-meter data. We also introduce the notion of privacy and provide a new quantized feedback respecting agents' privacy concerns yet improving prosociality. The results of this study are not restricted to the framework of energy efficiency but are also applicable to congestion problems in road traffic and other resource sharing problems.

CCS Concepts: • **Mathematics of computing** → **Mathematical analysis**; • **Applied computing** → **Decision analysis**.

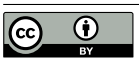
Additional Key Words and Phrases: information design, mean-field game, prosocial behavior

ACM Reference Format:
Alexandre Reiffers-Masson and Rajesh Sundaresan. 2026. Information Design for Prosocial Behavior. *Proc. ACM Meas. Anal. Comput. Syst.* 10, 2, Article 42 (June 2026), 26 pages. https://doi.org/10.1145/3805640

1 Introduction.

We ground our study of reputation-based information design for inducing prosocial behavior in the context of electricity consumption. Electricity utility providers are interested in the reduction of the power consumed by their residential consumer segment. Supply to this segment is often subsidized and if the electricity utility provider is able to save power, it will be possible to redirect it to more profitable segments. To do so, a classical approach in economics would be to implement taxes. However, in the residential consumer segment, these price-based policies can be difficult to

Authors' Contact Information: Alexandre Reiffers-Masson, alexandre.reiffers-masson@imt-atlantique.fr, IMT Atlantique, Plouzané, Finistère, France; Rajesh Sundaresan, rajeshs@iisc.ac.in, Indian Institute of Science, Bangalore, India.



ACM 2476-1249/2026/6-ART42
https://doi.org/10.1145/3805640

implement for political reasons (they are often unpopular) or for engineering reasons (for instance, lack of knowledge of demand elasticities for energy-efficient durable goods) .

Recently, considerable attention has been paid to non-price interventions and policies that *nudge* consumers to conserve energy. In general, these decentralized ways to change energy consumption behavior are inexpensive to implement and can be applied in any kind of society, whether developed or developing. One example in the context of demand response is the creation of a lottery with the distribution of energy coupons [25, 26, 35]. Another example is the use of social comparison and reputation to improve the prosocial behavior of an agent (which is, in this case, the reduction of power consumption) [3, 29, 30, 32]. Our work is motivated by such recent management programs. For instance, in [32], the authors conducted a randomized field experiment in a suburb of a large Indian city (Kerala), where residential consumers received personalized electricity reports based on smart-meter data. These reports provided structured feedback combining social comparison (by displaying the position of a household's consumption relative to the distribution of energy consumption of similar peers) and simplified recommendations to reduce electricity usage. The overall impact of the program was estimated to be around a 1% reduction in electricity consumption. Similar programs in the US [29, 30] have also been implemented, where a reduction of 2.0% in energy consumption has been observed. Other experiments have been performed and are summarized in [4]. The focus of our paper is to derive a game theoretical analysis of policies that use social comparison as a tool for improving prosocial behavior. This study is not the first of its kind and we highlight the difference with other related works in Section 7. Social comparison has been proved to be a powerful means to induce prosocial behavior. In the context of the reduction of the total power consumed by the residential consumer segment, we suggest using social comparison instead of monetary reward because reduced consumption can be portrayed as a prosocial action. There is then a reputation benefit that can be attached to the reduction effort. Contrary to what one might expect at first glance, monetary rewards may reduce the expected prosocial action in some settings, as demonstrated in [8].

Our game theoretical model and analysis can be described as follows. The intensity of a person's prosocial orientation will be determined by three factors: her intrinsic interest in doing a prosocial action, the cost associated with this action and finally the desire to be *perceived* as generous and altruistic. The last factor can be seen as the agent's desire to have a good reputation in the society. By society we mean a group of people upon whom the agents want to make a good impression. From this perspective, the society can be even the individual herself, when the individual is viewing and assessing her own action, in which case we talk about self-signaling and identity. As observed in [8], from a modeling perspective, there is no major difference between the society being a set of individuals or the agent herself.

The society assesses an individual's reputation based on its collective estimate of the agent's prosociality. This estimation is done using the knowledge of the actions of the different agents. This knowledge of actions can be total or partial. For example, in the context of household energy consumption, we can imagine two types of observation of the consumption of different households.

- In the first scenario, the electricity utility provider may publicize a full report to the society on the consumption of each agent[1]. The society is then able to estimate precisely how much more altruistic is one person's action than another person's action (thus ordering each person according to her respective altruism level).
- In the second scenario, perhaps to address privacy concerns, the report may only show that the consumption of an agent is larger or smaller than a threshold. Others in the society, then,

[1]While this might seem an overkill in the electricity consumption context, such things are possible in the context of donations, voluntarism, committee work that help run an institution, etc. Our study should be seen in this broader context.

are only able to say if a person belongs to the more generous group or to the less generous one.

These two examples have the merit of demonstrating that it is possible, by designing the information observed by members of the society, to mobilize societal opinion concerning an agent.

Our goal is to show the importance of information design to improve prosocial behavior. The initial model is inspired from [8] which we extend in this paper to more complex information designs. We compare the impact of the different parameters on efficiency and highlight the impact of the different information designs on prosocial behavior. From a mathematical point of view, extensions of the model [8] requires new proofs of existence, uniqueness under some conditions, and characterization of the mean field equilibria which we provide in this paper. Additionally, we discuss several possible extensions and future directions which we feel will be of use to researchers in the field.

1.1 Organization and Main Results.

The remainder of the paper is organized as follows. Section 2 introduces the game-theoretical model. We describe the main components of the societal network, which in our example is the complete (fully connected) network of all the agents, the agents' intrinsic motivations, their actions, the cost function associated with performing an action, and finally the reputational benefit associated with a given action. In this section, we also explain how the computation of reputational benefit can be understood as a signal extraction problem. We introduce the different feedback mechanisms (Type-A and Type-B) used to signal an agent's action. Section 3 is the main section of this paper. We prove the existence of, and study the properties of, mean field equilibria for the two feedback mechanisms. For each feedback, we characterize the mean field equilibrium using a fixed-point equation. For the second feedback (Type-B), sufficient conditions for uniqueness properties of the equilibrium follow naturally. One of the key insights of this section is that Type-B feedback induces discontinuities in agents' best responses, leading to a concentrated mass point at the threshold; some agents are better mobilized compared to the case of the first feedback. In Section 4, we show how important the choice of feedback and the choice of the partition (into less prosocial and more prosocial groups) can be in improving the expected prosocial action. More precisely, we prove that being "economical with the truth" (e.g., not revealing all information about agents' actions) can improve overall prosociality. We derive conditions under which Type-B feedback outperforms Type-A in aggregate welfare. We also compute explicitly the different mean field equilibria for each feedback considered, under the assumption that the intrinsic values follow a uniform distribution. Section 5 illustrates an instantiation of our framework using smart-meter electricity consumption data from London. We map household behavior into the model, infer intrinsic motivations, and empirically identify welfare-optimal feedback thresholds. The experiments reveal that even coarse, threshold-based feedback leads to significant behavioral shifts. Importantly, the effect of feedback depends on the strength β of reputational incentives. As β increases, both Type-A and Type-B produce substantial gains—up to 12% and 6.6% improvement in average prosociality, respectively. In Section 6, we formalize a notion of privacy, and then introduce a new feedback mechanism that improves privacy at the expense of the average level of prosociality. We provide sufficient conditions for the existence of an equilibrium under this new feedback, and demonstrate through an example that increasing the number of thresholds can lead to higher aggregate prosociality while preserving privacy. In Section 7, we discuss the related literature and highlight the major differences between our work and the unified framework suggested in [10, 11], as well as the norm-based approaches described in [2, 8, 9]. In Section 8, we discuss several possible extensions

of this work and future directions. Finally, Section 9 concludes the paper with a brief summary of our results.

2 Model.

Table 1. Main notations used throughout this paper

Symbol	*Meaning*
$\mathcal{I} \coloneqq [0,1]$	Agent set. $i \in \mathcal{I}$ is the index of a given agent.
$\mathcal{A} \coloneqq [0,+\infty)$	Action space. $a_i \in \mathcal{A}$ is the action of agent i.
$w_i \in \mathbb{R}_+$	Intrinsic motivation value representing propensity for prosocial actions. The empirical cdf of $(w_i)_{i\in I}$ is F.
$L(a_i)$	Quantized version of agent i's action provided by the service provider to the societal network.
$C(a_i)$	Cost function such that $C : [0,+\infty) \to \mathbb{R}_+$.

Table 2. Different L functions with the associated utilities

Feedback	$L(a_i)$	*Agents' objectives*
Type-A	a_i	$\max\limits_{a_i\in[0,+\infty)} \{a_i w_i - C(a_i) + \beta\mathbb{E}\left[w_i \mid a_i\right]\}$
Type-B	$1_{a_i\geq\theta}$	$\max\left\{\max\limits_{a_i\in[0,\theta]} a_i w_i - C(a_i) + \beta\mathbb{E}\left[w_i \mid a_i < \theta\right], \max\limits_{a_i\in[\theta,+\infty)} a_i w_i - C(a_i) + \beta\mathbb{E}\left[w_i \mid a_i \geq \theta\right]\right\}$

We assume that the societal network consists of a continuum number of agents, as we shall make precise soon. Each agent chooses an action from a set which is taken to be totally ordered in terms of prosocial behavior. An agent's choice of an action is based on three components, the *agent's intrinsic motivation* (for instance the prosocial orientation of the agent), a *cost* associated with her action and a *reputational benefit*. The reputational benefit captures the effect of judgments and reactions of other members of the society towards an agent. A social service provider (e.g. utility service provider or government) is interested in the maximization of the global level of prosocial actions. For each agent, the service provider can manipulate an agent's reputational benefit by designing the information other agents get about this agent's choice. We assume that the agent's intrinsic motivation is *private information* known only to herself but the distribution of agents' intrinsic motivation is *common knowledge*. Agents interact with each other through the information fed by the social service provider and the consequent reputational benefits they derive. Moreover, since the number of agents is infinite, the global level of prosocial actions in the societal network is the outcome of a *mean field equilibrium*. The service provider's problem is thus an *Information Design Problem*.

The three main components of the societal network are the following. i) Agents and actions: What is the set of agents, what are the agents' action spaces, and how do others in the society

interpret agents' actions? ii) Intrinsic value and cost function: What is the information known about the agents' intrinsic values and the cost functions? iii) Reputational benefit: How is reputational benefit quantified? In the rest of this section, we develop a model of the societal network and address each of these questions. The main symbols used in this paper are summarized in Table 1.

Agents and actions: The agent set is the continuum $\mathcal{I} \coloneqq [0,1]$. Let $i \in \mathcal{I}$ be a given agent. We suppose that each agent has a continuum of possible actions. We denote the action space as $\mathcal{A} \coloneqq [0,+\infty)$. The action of agent i is denoted by $a_i \in \mathcal{A}$. For all $(a_i, a_i') \in \mathcal{A}^2$, if $a_i > a_i'$ then agent i performs a greater prosocial action when she chooses a_i over a_i'. For instance if a_i captures the energy *savings* of agent i, then the savings of energy is greater under the greater prosocial action a_i than under the lesser prosocial action a_i'.

Intrinsic value and cost function: Each agent i is endowed with an intrinsic value w_i and gets a reward of value $a_i w_i$ for an action a_i. Larger the w_i, greater the propensity of the agent towards a more prosocial action. The cumulative distribution function (cdf) of w_i is F. For each agent i, performing the action a_i costs $C(a_i)$, where $C : [0,+\infty) \to \mathbb{R}_+$ is a convex and increasing function. F and C are common knowledge to all the agents and the service provider.

Reputational benefit: The reputational benefit is described by the following steps.

(1) For each $i \in \mathcal{I}$, the intrinsic motivation of agent i is given by her prosocial propensity w_i. When she chooses action a_i, she reveals some information about w_i to the service provider.
(2) Given agent i's action a_i, the service provider reveals $L(a_i)$, a quantized version of a_i, to all agents in the societal network. The function L is common knowledge. Our goal is to understand the consequence of various choices of L by the service provider. Since L controls the amount of information about agent i's action a_i, this is an information design problem. Examples are $L(a_i) = a_i$ or the privacy friendly feedback scheme $L(a_i) = 1_{a_i \geq \theta}$, with $\theta \in \mathbb{R}_+$.
(3) When an action a_i is taken and $L(a_i)$ is revealed to all, as discussed above, some information about the agent's private w_i is also revealed to the societal network yielding a reputational benefit $\beta \mathbb{E}\left[w_i \mid \{L(a_k)\}_{k \in \mathcal{I}}\right]$, with $\beta \in \mathbb{R}_+$. The notion that the reputation of an agent is based on the societal network's opinion about her intrinsic value w_i, arising from the information that her action a_i reveals, has already been used in [8, 9, 15].

The above is a model in the context of consumption of energy in a psychological experiment in [29], which we now describe. It encompasses the following steps.

Step 1: For a given day, the service provider measures the consumption of energy of each household in a neighborhood.
Step 2: If the consumption of a given household is below θ, the service provider puts a green flag in front of that house; otherwise, nothing is done.
Step 3: Each household observes the flag in front of every other house and estimates the intrinsic motivation of that other household.

For each $i \in \mathcal{I}$, we assume that the utility of agent i for action a_i is the sum of the rewards arising from her propensity for prosocial behavior, the cost function and the reputational benefit. Agent i is thus interested in maximizing:

$$\max_{a_i \in [0,+\infty)} U(a_i, w_i; a_{-i}, w_{-i}) \coloneqq a_i w_i - C(a_i) + \beta \mathbb{E}\left[w_i \mid \{L(a_k)\}_{k \in \mathcal{I}}\right], \tag{1}$$

with a_{-i} (resp. w_{-i}) being the actions (resp. intrinsic motivations) of all the agents except agent i.

Let G be the cdf of actions a_i, $i \in \mathcal{I}$. This results in a certain feedback profile $\{L(a_k)\}_{k \in \mathcal{I}}$. The best response of agent i to $\{L(a_k),\ k \in \mathcal{I}\}$, which is a function of G, is given by:

$$a_i^*(w_i; G) = \arg \max_{a_i \in [0,+\infty)} U(a_i, w_i; a_{-i}, w_{-i}). \tag{2}$$

We will assume that $a_i^*(w_i; G)$ is uniquely defined. See examples later. For a given G, let TG be the distribution of the induced best response actions. A mean field equilibrium is defined as follows.

Definition 1. *The distribution G^* is a mean field equilibrium if $TG^* = G^*$.*

We refer the technical reader to a mathematically rigorous reformulation in the Appendix C. The equilibrium G^* will naturally depend on the feedback signal L.

One objective of the service provider could be to optimize the aggregate prosocial action $\int_{\mathcal{A}} b\, dG^*(b)$. Let W_j be the expected prosocial action when type-j feedback is provided, with $j \in \{A, B\}$. Type-A feedback reveals the action of an agent to all agents in the society. This feedback does not preserve the privacy of an agent's action. Under type-B feedback, the society gets to know only whether an agent belongs to the more prosocial group or to the less prosocial one. Therefore this provides better privacy than type-A feedback. See Table 2 for a summary of the two feedback types and the utilities.

The key ideas are best conveyed in the simplest settings. To do this, we make the following assumptions about the cost function and the feedback functions.

Assumption A:

(1) The cost function $C(a) = \frac{1}{2}\alpha a^2$.
(2) $L(\cdot)$ is one of the functions defined in Table 2. In type-B, the service provider can additionally control one threshold parameter, denoted θ.

We discuss extensions to general convex costs in Section 8 and additional feedbacks in Section 6.

3 Equilibria for type-A and type-B feedback schemes.

In this section, we characterize and study the properties of mean field equilibria for reputational benefit feedback types A and B. For ease of notation, we write $\mathbb{E}\left[w_i \mid L_j(\cdot)\right]$ for $\mathbb{E}\left[w_i \mid \{L_j(a_k)\}_{k\in\mathcal{I}}\right]$, when the feedback is type-j, $j \in \{\text{A}, \text{B}\}$. We will also write $\mathbb{E}\left[w_i \mid L_j(\cdot), L_j(a_i)\right]$ to draw the reader's attention to the feedback $L(a_i)$.

3.1 Type-A equilibrium.

We begin with a characterization of type-A equilibrium which is also portrayed in Figure 1a.

Theorem 1. *Under assumption A, there is a mean field equilibrium for type-A feedback. The prosocial action of player i is the unique solution to the following equation:*

$$a_i = \frac{w_i}{\alpha} + \beta(1 - e^{-\frac{a_i}{\beta}}). \tag{3}$$

Moreover,

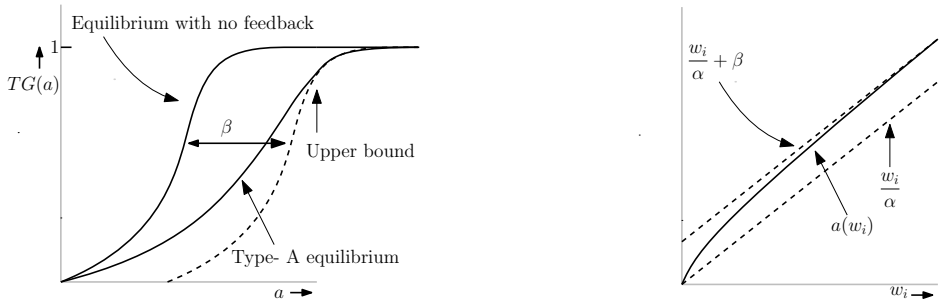
$$\mathbb{E}[w_i \mid L_A(\cdot)] = w_i \tag{4}$$

$$\frac{\partial \mathbb{E}[w_i \mid L_A(\cdot)]}{\partial a_i} = \frac{\beta}{\alpha}(1 - e^{-\frac{a_i}{\beta}}). \tag{5}$$

Proof. See Appendix B.1. □

Remark: Theorem 1 states that type-A feedback will result in an equilibrium that not only reveals a_i to the whole society but also reveals the intrinsic value w_i (since we will have $\mathbb{E}[w_i \mid L_A(\cdot)] = w_i$). In this case, the societal network learns everything about every agent's intrinsic motivation. Observe that feedback has pushed the agents towards more prosocial actions, and $w_i/\alpha \le a_i \le w_i/\alpha + \beta$. In the limiting cases, $\lim_{w_i\to+\infty} a_i = \frac{w_i}{\alpha} + \beta$ and $\lim_{w_i\to 0_+} a_i = \frac{w_i}{\alpha}$. See Figure 1b.

Remark: An examination of the proof below will indicate that it is nontrivial and nonstandard because it involves a functional fixed point equation. A similar exercise is carried out in Bénabou



(a) Distribution of $a^*(w_i; G)$ under Type-A feedback. (b) Equilibrium action $a^*(w_i; G)$ under Type-A feedback.

Fig. 1. Distribution and shape of $a^*(w_i; G)$ under Type-A feedback.

and Tirole for the two-dimensional Gaussian case [8]. In our case, we have extended their proofs for a general distribution and with a general cost function (see the next theorem). From the proof, we glean that uniqueness of the solution to equation (5) crucially hinges on the boundary condition $a_i = 0$ when $w_i = 0$.

In the next theorem, we will derive a Nash equilibrium, for a general convex cost function, under Type-A feedback. This result extends the previous theorem to a more general set-up.

Theorem 2. *Let us assume that $C'(0) = 0$, $C'(a_i) > 0$, $\overline{\alpha} > C''(a_i) > \underline{\alpha} > 0$, $\underline{\alpha} > \frac{\overline{\alpha}}{\overline{\beta}}$ and $C'''(a_i) < \frac{\underline{\alpha}}{\overline{\beta}}$). Then the prosocial action of player i is the unique solution of the following equation:*

$$a_i = (C')^{-1}(w_i + \beta(\Xi(a_i) - \Xi(0)e^{-\frac{a_i}{\beta}})), \tag{6}$$

where $\Xi(a_i) = e^{-\frac{a_i}{\beta}} \int_0^{a_i} C''(b)e^{\frac{b}{\beta}}db$.

Proof. See Appendix B.2. □

3.2 Type-B equilibrium.

As described in the previous section, agent i computes her prosocial action given her intrinsic motivation w_i and her cost function $C(\cdot)$. First we state a fixed point equation that characterizes the best response a_i to an action profile G for each w_i under the assumption that $L(\cdot)$ is of type-B. Following this, we demonstrate the existence of mean field equilibria (MFE).

For the type-B feedback function $L_B(\cdot)$, and with G being the cdf of actions, agent i's best response $a_i^*(w_i; G)$ is determined via:

$$\begin{aligned} &U_B(a_i^*(w_i; G), w_i) \coloneqq \\ &\max\left\{ \underbrace{a_{i1}^*(w_i; G)w_i - \alpha\frac{(a_{i1}^*(w_i; G))^2}{2} + \beta\mathbb{E}\left[w_i \mid L_B(\cdot), a_{i1}^* < \theta\right]}_{U_{1B}(a_{i1}^*, w_i)}, \right. \\ &\left. \underbrace{a_{i2}^*(w_i; G)w_i - \alpha\frac{(a_{i2}^*(w_i; G))^2}{2} + \beta\mathbb{E}\left[w_i \mid L_B(\cdot), a_{i2}^* \geq \theta\right]}_{U_{2B}(a_{i2}^*, w_i)} \right\}, \end{aligned} \tag{7}$$

where (if the maxima below exist):

$$a^*_{i1}(w_i;G) \coloneqq \arg\max_{a_i\in[0,\theta)} U_{1B}(a_i,w_i), \tag{8}$$

$$a^*_{i2}(w_i;G) \coloneqq \arg\max_{a_i\in[\theta,+\infty)} U_{2B}(a_i,w_i), \tag{9}$$

$$a^*_i(w_i;G) \coloneqq \begin{cases} a^*_{i1}(w_i;G) & \text{if } U_{1B}(a^*_{i1}(w_i;G),w_i) > U_{2B}(a^*_{i2}(w_i;G),w_i),\\ a^*_{i2}(w_i;G) & \text{otherwise.}\end{cases}$$

Here, under assumption of existence of the above maxima, the candidate action level $a^*_{i1}(w_i;G)$ corresponds to the optimal action of agent i if she were to perform an action below the threshold θ. On the contrary, if agent i were to perform an action θ or above, then the optimal choice would be $a^*_{i2}(w_i;G)$. The final choice $a^*_i(w_i;G)$ is the better of the two and thus the global optimum. In order to study the equilibria of this game, the first step is to derive an expression of $a^*_{i1}(w_i;G)$, $a^*_{i2}(w_i;G)$ and $a^*(w_i;G)$ as a function of w_i by assuming that the reputational benefits are given. Then the second step will be to derive a closed form expression of reputational benefits.

Let $c_1 \coloneqq \mathbb{E}[w_i \mid L_B(\cdot), a_i < \theta]$ and $c_2 \coloneqq \mathbb{E}[w_i \mid L_B(\cdot), a_i \geq \theta]$. Clearly c_1 is a function of $L_B(\cdot)$ of all agents and $1_{a_i<\theta}$ for the agent i, and hence is a constant for all $a_i < \theta$. Similarly c_2 is a constant for all $a_i \geq \theta$.

Proposition 1. *(Best response to G) Under assumption A, if $\Delta_B(G) \coloneqq c_2 - c_1$ is positive, then agent i's best response to G is*

$$a^*_i(w_i;G) \coloneqq \begin{cases} w_i/\alpha & \text{if } w_i \in [0,u) \cup [\alpha\theta,+\infty)\\ \theta & \text{if } u \geq w_i < \alpha\theta,\end{cases} \tag{10}$$

with $u \in [0,\alpha\theta]$ satisfying:

$$u = \left[\alpha\theta - \sqrt{2\alpha\beta\Delta_B(G)}\right]_+. \tag{11}$$

Proof. See Appendix B.3. □

The next corollary provides the distribution of the best response profile $a^*(w_i;G)$, and can be easily deduced from Proposition 1.

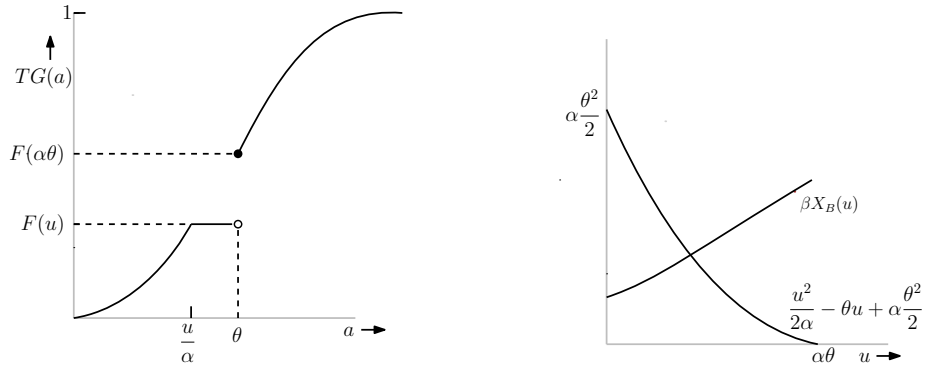
Corollary 1. *Let G be an action profile. Under assumptions A and $\Delta_B(G) \geq 0$, the best response profile TG is:*

$$TG(a) \coloneqq \begin{cases} F(\alpha a) & \text{if} \quad a \in [0,u/\alpha),\\ F(u) & \text{if} \quad a \in [u/\alpha,\theta),\\ F(\alpha a) & \text{if} \quad a \in [\theta,+\infty),\end{cases} \tag{12}$$

where u is given in (11) and depends on G through $\Delta_B(G)$.

The distribution of the best response of the agents is depicted in Figure 2a. The shape TG has the following properties. Firstly we note that if $w_i > \alpha\theta$, the prosocial action of agent i is equal to $\frac{w_i}{\alpha}$ which is the same as in the case when the reputation benefit is not part of her utility ($\beta = 0$). Also we can notice that by playing $\frac{w_i}{\alpha}$, she will reveal to the service provider her intrinsic motivation, since the service provider knows α. In the case of $w_i < \alpha\theta$, two situations can occur. If $w_i < u$, then again the best response of agent i will be to choose $\frac{w_i}{\alpha}$ and the same conclusion can be applied. On the other hand, if $w_i \in [u,\alpha\theta)$, then agent i chooses $a^*_i(w_i;G) = \theta$ which implies that now her reputation is equal to $\mathbb{E}[w_i \mid L_B(\cdot), a_i \geq \theta]$. From this observation, we can deduce that, first, the agent makes the bare minimum effort to go into the next higher reputation category and, second, the agent does not fully reveal w_i to the service provider [2]. This pattern of jump in

[2]The service provider knows the actions, and, from $a_i = \theta$, can infer that $w_i \in [\frac{u}{\alpha},\theta)$. However, the other agents only get to see $1_{a_i\geq\theta}$. The increased reputation benefit stems from this economical feedback of information to the other agents.



(a) Distribution of $a^*(w_i; G)$ under the type-B feedback. A segment of the population increase their prosocial action level to θ.

(b) Unique equilibrium for type-B feedback

Fig. 2. Illustration of type-B feedback equilibrium properties.

prosocial actions for a segment of the population has been also observed in donations [34], when there are categories of donations. Indeed, the amount of donations are not distributed according to the uniform distribution but rather according to a multimodal distribution where each mode corresponds to the minimum amount of donation needed to enter into a corresponding category. All others have either (1) sufficiently high intrinsic motivation ($w_i > \alpha\theta$) that playing lower without losing the reputational benefit only results in lower intrinsic benefit and so lower utility, or (2) sufficiently low intrinsic motivation ($w_i < u$) that playing higher will result in higher cost.

We assume that F has a density $f(\cdot)$ and a finite expectation $\mathbb{E}[w_i]$. Define Tc_1 and Tc_2 as the c_1 and c_2 associated with TG (see paragraph before Proposition 1). These are then:

$$Tc_1 \;\coloneqq\; \frac{\int_0^u w f(w)\,dw}{F(u)} = u - \frac{\int_0^u F(w)\,dw}{F(u)}. \tag{13}$$

$$Tc_2 \;\coloneqq\; \begin{cases} \dfrac{\int_u^{+\infty} w f(w)\,dw}{1-F(u)} = u + \dfrac{\int_u^{+\infty}[1-F(w)]\,dw}{1-F(u)}, & \text{if} \quad F(u) < 1, \\ u & \text{if} \quad F(u) = 1, \end{cases} \tag{14}$$

where (14) holds because the distribution F is assumed to have finite mean. Thus,

$$\Delta_B(TG) = Tc_2 - Tc_1 = \begin{cases} \dfrac{\int_u^{+\infty}[1-F(w)]\,dw}{1-F(u)} + \dfrac{\int_0^u F(w)\,dw}{F(u)}, & \text{if} \quad F(u) < 1 \\ u - \mathbb{E}[w_i], & \text{if} \quad F(u) = 1. \end{cases} \tag{15}$$

At equilibrium, $TG^* = G^*$ and hence $\Delta_B(TG^*) = \Delta_B(G^*)$. Since $\Delta_B(TG)$ depends only on F (known, fixed) and u (to be determined), let us write:

$$X_B(u) \coloneqq \begin{cases} \dfrac{\int_u^{+\infty}[1-F(w)]\,dw}{1-F(u)} + \dfrac{\int_0^u F(w)\,dw}{F(u)}, & \text{if } F(u) < 1, \\ u - \mathbb{E}[w], & \text{if } F(u) = 1. \end{cases} \tag{16}$$

Taking the derivative with respect to u and simplifying, we get

$$X_B'(u) = \begin{cases} f(u)\left[\dfrac{\int_u^{+\infty}[1-F(w)]\,dw}{(1-F(u))^2} - \dfrac{\int_0^u F(w)\,dw}{F(u)^2}\right] & \text{if} \quad F(u) < 1, \\ 1, & \text{if} \quad F(u) = 1. \end{cases} \tag{17}$$

From (16) and (17), we can conclude that $\lim_{u\downarrow 0} X_B(u) = \mathbb{E}[w]$, that $X_B(u) \geq u - \mathbb{E}[w] - \delta$ for a $\delta > 0$ and all u sufficiently large and hence $\lim_{u\uparrow+\infty} X_B(u) = +\infty$. Furthermore, we can also conclude that $X_B(u)$ is differentiable for all u with $F(u) < 1$, and further $X_B(u)$ is continuous for all $0 \leq u < +\infty$. From the relation in (11) and the condition $\Delta_B(TG^*) = \Delta_B(G^*)$, we see that the intersection (or lack thereof) of the curves $X_B(u)$ in (16) and $(u-\alpha\theta)^2/(2\alpha)$ will determine the equilibria.

Two cases can occur. The first case is when there is no intersection between $\beta X_B(u)$ and $(u-\alpha\theta)^2/(2\alpha)$. In this case, $\beta X_B(u)$ is always greater, and from an equilibrium perspective, $u = 0$ in (11), and all agents such that $w_i < \alpha\theta$ will play θ. The rest of the agents will play $\frac{w_i}{\alpha}$. The second case is when $\beta X_B(u)$ and $(u-\alpha\theta)^2/(2\alpha)$ intersect for some $u < \alpha\theta$, as in Figure 2b. Clearly, if F has a nontrivial density, then $X_B(\alpha\theta) > 0$. (If $F(\alpha\theta) = 1$, then $\mathbb{E}[w] < \alpha\theta$ since F has a nontrivial density and $X_B(\alpha\theta) > 0$ from the second case in (16). If $F(\alpha\theta) < 1$, then first formula in (16) applies and $X_B(\alpha\theta) \geq \left(\int_0^{\alpha\theta} F(w)\,dw\right)/F(\alpha\theta) > 0$.) A sufficient condition for an interior u^{MFE} is then $\theta > \sqrt{2\beta\mathbb{E}[w]/\alpha}$. We have thus established:

Theorem 3. *A mean field equilibrium always exists. Moreover, the following hold:*

(1) If $\beta X_B(u)$ and $(u-\alpha\theta)^2/(2\alpha)$ intersect at $u^{MFE} \in [0, \alpha\theta)$, then a mean field equilibrium exists with $a_i^*(w_i)$ as given in (10) and u^{MFE} solution of (11).

(2) If $\beta X_B(u)$ is greater than $(u-\alpha\theta)^2/(2\alpha)$ for all $u \in [0, \alpha\theta]$, then all agents such that $w_i < \alpha\theta$ will play θ and the rest of the agents will play $\frac{w_i}{\alpha}$.

(3) If $\theta > \sqrt{\frac{2\beta\mathbb{E}[w]}{\alpha}}$, there is at least one intersection for the curves $\beta X_B(u)$ and $\frac{(u-\alpha\theta)^2}{2\alpha}$.

Uniqueness of the intersection of the function $\beta X_B(u)$ and the function $\frac{u^2}{2\alpha} - \theta u + \alpha\frac{\theta^2}{2} = \frac{(u-\alpha\theta)^2)}{2\alpha}$ can be ensured if $\beta X_B(u)$ is increasing in u.

4 Shaping of feedback.

We now consider two examples to highlight the need for a systematic study of information design.

4.1 Being economical with the truth can improve the net prosocial action.

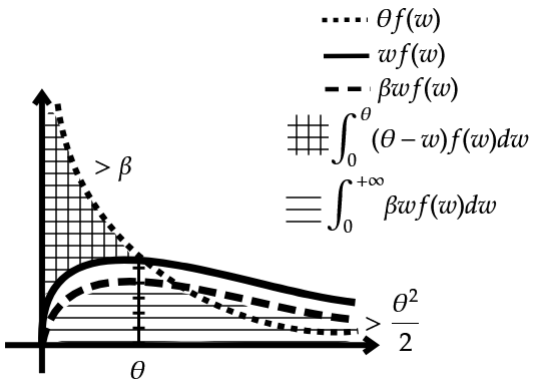
In type-A feedback the service provider revealed all the actions. In type-B feedback, the service provider revealed only the categories (more prosocial or less prosocial) to which an individual belonged. Let W_A and W_B denote the net prosocial actions of the population under type-A feedback and type-B feedback respectively. In the next proposition, we will give conditions such that W_B is greater than W_A, a surprising result at first glance.

Proposition 2. *If $f(w)$ is decreasing in $w \in [0, +\infty)$ and if there exist θ, α and β such that the following conditions are satisfied:*

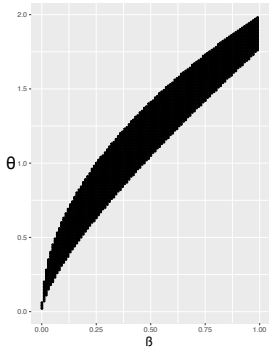
$$\int_0^{\alpha\theta} \left(\theta - \frac{w}{\alpha}\right) f(w)\,dw \quad > \quad \beta, \tag{18}$$

$$\beta\mathbb{E}[w] > \frac{\alpha\theta^2}{2}, \tag{19}$$

then $W_B > W_A$.



(a) Graphical illustration of (18) and (19) when $\alpha = 1$.



(b) Parameter pairs (β, θ) satisfying Proposition 2 (Weibull).

Fig. 3. Situations when $W_B > W_A$.

Proof. See Appendix B.4. □

We now provide an example where the conditions of Proposition 2 are satisfied. In Figure 3a, we first understand why these two conditions are in opposition. Indeed, if we want (18) to be satisfied we will need a small β so that the area of the crossed region is at least β. But if β is too small we will reduce the area of the square-dotted region which will affect (19). Similar conclusions can be drawn for θ. By increasing θ, it will be easier to satisfy (18), but more difficult to satisfy (19).

Let us consider that the density $f(w) = \frac{k}{\lambda}(\frac{w}{\lambda})^{k-1} e^{-(\frac{w}{\lambda})^k}$ (Weibull distribution) with a scale parameter (λ) equal to 1 and shape parameter (k) equal to 0.5. When the shape parameter is lower than 1, $f(w)$ is decreasing in w. Take $\alpha = 1$. The results of a numerical simulation in Figure 3b, indicate that the pairs (β, θ) in the darkened area satisfy the conditions of Proposition 2. For these parameters, being "economical with the feedback", by sharing only the quantized feedback, improves the net prosocial action.

4.2 The uniform distribution case and an explicit optimization.

In this subsection, we study the example when F has the uniform distribution over $[0, 1]$. Under this assumption, an agent's intrinsic motivation is her rank in the society, i.e., $F(w_i) = w_i = i$. The utility of agent i can be rewritten as follows:

$$U(a_i, i, a_{-i}) = a_i i - C(a_i) + \beta \mathbb{E}[i \mid \{L(a_k)\}_{k \in \mathcal{I}}]. \tag{20}$$

For simplicity, assume $\alpha\beta < 1$. Since we will be interested in making individuals jump to larger prosocial actions, from Figure 2a, we have that $\alpha\theta \leq 1$.

It can be easily checked that $X_B(u) \equiv 0.5$. The mean field is determined by the intersection between a constant function $\beta X_B(u) \equiv \beta 0.5$ and the decreasing polynomial function $(u - \alpha\theta)^2/(2\alpha)$ in the interval $u \in [0, \alpha\theta]$. If there is no intersection then $u^{MFE} = 0$. From this, we deduce that $u^{MFE} = [\theta\alpha - \sqrt{\alpha\beta}]_+$. Furthermore, by Theorem 3:

$$a^*(i) \coloneqq \begin{cases} i/\alpha & \text{if } i \in [0, [\alpha\theta - \sqrt{\alpha\beta}]_+) \cup [\alpha\theta, 1] \\ \theta & \text{otherwise.} \end{cases} \tag{21}$$

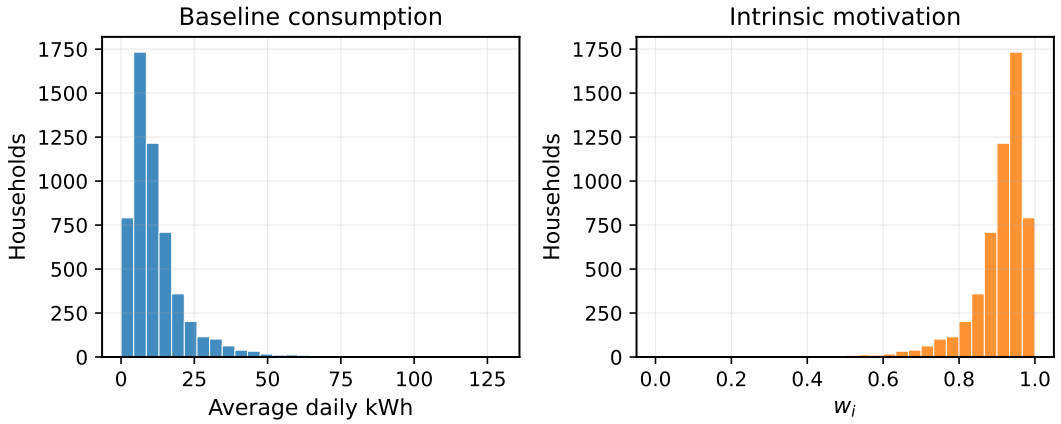


Fig. 4. Baseline consumption (left) and intrinsic motivation w_i (right) for March 2013.

The net prosocial action under type-B feedback is given by:

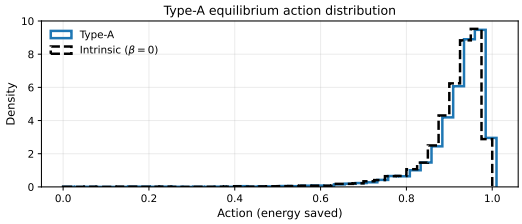
$$\begin{aligned} W_B &= \frac{1}{2\alpha}[w^2]_0^{[\alpha\theta-\sqrt{\alpha\beta}]_+} + \theta(\alpha\theta - [\alpha\theta - \sqrt{\alpha\beta}]_+) + \frac{1}{2\alpha}[w^2]_{\alpha\theta}^1 \\ &= \frac{1}{2\alpha} + \frac{1}{2}\min\left\{\alpha\theta^2, \beta\right\}. \end{aligned} \tag{22}$$

Note that if $\theta \leq \sqrt{\beta/\alpha}$, then W_B is increasing in θ else W_B is independent of θ. Therefore the optimal θ to use will be any θ that satisfies $\frac{1}{\alpha} \geq \theta \geq \sqrt{\frac{\beta}{\alpha}}$.

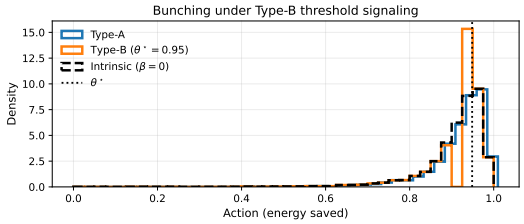
5 Numerical Experiments on smart meter data

In this section, we evaluate the different feedback mechanisms on real smart-meter data in order to illustrate the performance of our approach and how it can be instantiated in a real-world scenario. We use the publicly available *London Smart Meter* dataset [7]. We use an already sanitized version of the dataset made available on Kaggle [22]. This dataset contains half-hourly electricity consumption records for 5,567 London households over the period of November 2011 to February 2014. In this paper, we focus on March 2013, even if our conclusions are the same for the other months. To compute the intrinsic motivation of each household, we first aggregate consumption at the daily level and restrict attention to households with complete records over the considered period. For each month (in our case March 2013) and each household, we define the baseline consumption level as the average daily energy usage over that month. Intrinsic motivation parameters are then constructed by reversing and normalizing baseline consumption, namely $w_i = \frac{\max_j b_j - b_i}{\max_j b_j - \min_j b_j}$, where b_i denotes the baseline consumption of household i. Thus, households with lower baseline consumption are assigned higher intrinsic motivation. This normalization ensures that $w_i \in [0, 1]$ while preserving relative heterogeneity. Figure 4 reports the empirical distribution of baseline consumption and the induced distribution of intrinsic motivations.

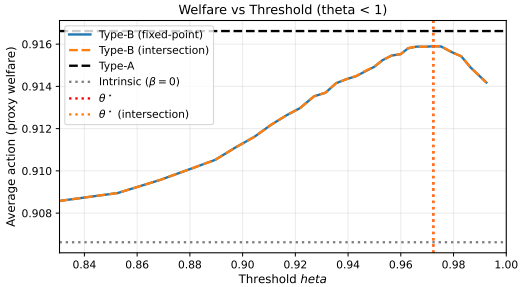
Unless stated otherwise, all experiments use $\alpha = 1$ and $\beta = 0.01$. Note that the average action is approximately 0.907 in the absence of feedback ($\beta = 0$). To compute the equilibria associated with the different feedback mechanisms, we iterate the corresponding fixed-point maps until convergence (see Appendix D, Algorithms 1).



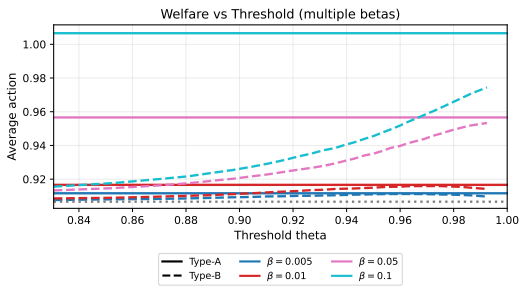
(a) Type-A (full revelation): equilibrium action density.



(b) Type-A vs Type-B comparison (distributional shift under thresholding).



(c) Welfare (average action) vs. threshold θ.



(d) Welfare (average action) vs. threshold θ for different β.

Fig. 5. Type-B threshold signaling on the London smart-meter data.

We first consider the Type-A mechanism, corresponding to full information revelation. Figure 5a reports the resulting action distribution. Note that, the distribution is smooth and exhibits no mass points. The average action under Type-A feedback is approximately 0.9165 in this case. We now focus on Type-B feedback and its comparison with Type-A feedback. We study how welfare varies with the threshold level. Figure 5c plots the average equilibrium action as a function of the threshold θ. Welfare is non-monotonic in θ and attains a maximum at an interior value $\theta^\star = 0.966$ in that case, it is corresponding to try to make everyone jump. The mean action obtained under Type-B feedback at $\theta^\star$ is about 0.916, which is lower than the value 0.9165 achieved under Type-A feedback.

Figure 5d compares the equilibrium action distributions under Type-A and Type-B feedback, using the welfare-optimal threshold $\theta^\star$ for different threshold θ and different $\beta \in \{0.005, 0.01, 0.05, 0.01, 0.1\}$. The first observation is that for all the β and for all θ, on this dataset, Type-A always leads to a higher welfare than Type-B. Moreover, we can observe that the greater β is, the greater is the optimal threshold, going up to the point where $\theta \approx 1$; in this extreme case, any agent that it is not acting like the most prosocial individual will be in the less prosocial segment of the society.

Next, observe that, when $\beta = 0.1$, Type-A (resp. Type-B with optimal θ) can lead to a relative increase of about 12% (6.6%) in average prosociality compared to the case with no incentives, a significant impact. When the reputational weight is small (e.g., $\beta = 0.01$ on this dataset), the improvement of the average prosocial action over the no-feedback baseline is present though small.

The experiments on this real dataset clearly demonstrate the existence of a non-trivial welfare-optimal threshold, highlighting the role of information design even in weak-incentive regimes. For larger reputational incentives (here we are speaking of $\beta = 0.1$ or higher), both Type-A and Type-B feedback mechanisms lead to a significant improvement of the average social behavior compared to the baseline, confirming the effectiveness of reputation-based nudges.

6 A new feedback for trading off privacy for efficiency.

Our initial intuition was that type-A feedback will maximize the level of aggregate prosocial action. Surprisingly, as highlighted in Figure 3b for the example Weibull distribution, this is not always the case. Additionally, type-A feedback may not satisfy the privacy concerns of agents. In this section, we will first define a privacy measure associated with a feedback. We then extend type-B feedback by allowing multiple thresholds and study how such new feedback can lead to a higher average prosocial behavior while preserving privacy.

6.1 A privacy measure.

Our measure of privacy is the extent to which the society is uncertain about an agent's intrinsic motivation, averaged across the population. Recall that the society observes $L(a_i^*(w_i))$, therefore $\mathbb{E}[w_i \mid L(\cdot)]$ is the minimum mean squared error estimate of w_i. We define the privacy measure as the mean square error over the population:

$$V(L) = \int_0^{+\infty} (w_i - \mathbb{E}[w_i \mid L(\cdot)])^2 f(w_i)\, dw_i. \tag{23}$$

When $V(L) = 0$, the society is able to infer precisely the true value of w_i for each agent i. The higher $V(L)$ the greater is our measure of privacy.

6.2 Type-Bm feedback (multi-threshold type-B feedback).

We explore the following feedback:

$$L_{Bm}(x) = n - 1 \text{ if } x \in [\theta_{n-1}, \theta_n), \tag{24}$$

with $0 = \theta_0 < \theta_1 < \cdots < \theta_N = +\infty$. Let us consider the following candidate equilibrium, where the strategy of agent i, if $w_i \in [v_{n-1}, v_n)$, with $v_n \in [\theta_{n-1}, \theta_n)$ for all $n \in \{1, \ldots, N\}$, is given by:

$$\arg \max_{a_i \in [\theta_{n-1}, \theta_n)} \{a_i w_i - C(a_i) + \beta \mathbb{E}[w_i \mid L_{Bm}(\cdot), a_i \in [\theta_{n-1}, \theta_n)]\}, \tag{25}$$

with $0 = v_0 < v_1 < \cdots < v_{\max}$, where $v_{\max}$ is the smallest v with $F(v) = 1$. Only a quantized signal of agent i's action is revealed, whose $w_i \in [v_{n-1}, v_n)$, is revealed, and therefore her reputational benefit comes from

$$\begin{aligned} Y(v_n, v_{n-1}) &\coloneqq \mathbb{E}[w_i \mid L_{Bm}(\cdot), a_i \in [\theta_{n-1}, \theta_n)] \\ &= \mathbb{E}[w_i \mid w_i \in [v_{n-1}, v_n)] = \frac{\int_{v_{n-1}}^{v_n} w f(w)\, dw}{F(v_n) - F(v_{n-1})}. \end{aligned}$$

Note that $Y(v_{n+1}, v_n) \geq Y(v_n, v_{n-1})$ for all $n \in \{1, \ldots, N-1\}$.

An agent i, with $w_i \in [v_{n-1}, v_n)$, who is evaluating a deviation from the candidate equilibrium, will face the following optimization problem:

$$\max_{n \in \{1, \ldots, N\}} \left\{ \left[\frac{w_i}{\alpha} \right]_{\theta_{n-1}}^{\theta_n} w_i - \frac{\alpha}{2} \left(\left[\frac{w_i}{2\alpha} \right]_{\theta_{n-1}}^{\theta_n} \right)^2 + \beta Y(v_n, v_{n-1}) \right\}, \tag{26}$$

with $[x]_a^b = \min\{\max\{x, a\}, b\}$. This is obtained by first solving the optimization problem (25) within the interval $a_i \in [\theta_{n-1}, \theta_n)$, the solution to which is $\left[\frac{w_i}{\alpha} \right]_{\theta_{n-1}}^{\theta_n}$, followed by an optimization over $n \in \{1, 2, \ldots, N\}$.

If we assume that for all $n' \geq n^*$ which may depend on i,

$$\theta_{n'} \left(w_i - \frac{\alpha \theta_{n'}}{2} \right) + \beta Y(v_{n'+1}, v_{n'}) > \theta_{n'+1} \left(w_i - \frac{\alpha \theta_{n'+1}}{2} \right) + \beta Y(v_{n'+2}, v_{n'+1}), \tag{27}$$

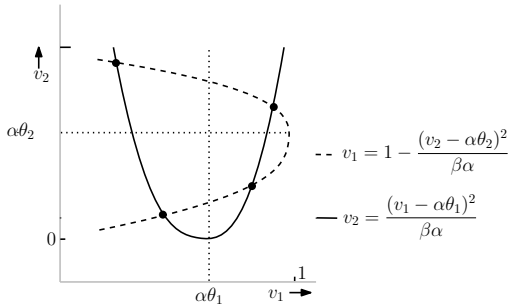


Fig. 6. Fixed point equations for Type-Bm feedback, under the condition that $v_1 \in (0, \alpha\theta_1)$ and $v_2 \in (\alpha\theta_1, \alpha\theta_2)$.

with n^* such that $\frac{w_i}{\alpha} \in [\theta_{n^*-1}, \theta_{n^*})$, then this optimization problem can be rewritten as:

$$\max \left\{ \frac{w_i^2}{2\alpha} + \beta Y(v_{n^*}, v_{n^*-1}), \quad \theta_{n^*}(w_i - \frac{\alpha\theta_{n^*}}{2}) + \beta Y(v_{n^*+1}, v_{n^*}) \right\}. \tag{28}$$

Now the deviation of player i can be restricted to a choice in $[\theta_{n^*-1}, \theta_{n^*})$ (in fact $\frac{w_i}{\alpha}$) or a choice in $[\theta_{n^*}, \theta_{n^*+1})$ (in fact θ_{n^*}). The equilibrium condition is therefore equivalent to the case where player i with $w_i = v_n$ is indifferent to choosing $\frac{v_n}{\alpha}$ or θ_n. Therefore, we need that, for all $n \in \{1, \ldots, N-1\}$, v_n satisfies the following vector fixed point equation:

$$\frac{v_n^2}{2\alpha} + \theta_n \left(\frac{\alpha\theta_n}{2} - v_n \right) = Y(v_{n+1}, v_n) - Y(v_n, v_{n-1}), \tag{29}$$

with $v_n \in [\alpha\theta_{n-1}, \alpha\theta_n)$ for all $n \in \{1, \ldots, N-1\}$.

Consider now an example where F is a uniform distribution over the interval $[0, 1]$. The previous fixed-point equation can be rewritten as a system of equations:

$$\frac{(v_n - \alpha\theta_n)^2}{2\alpha} = \beta \frac{v_{n+1} - v_{n-1}}{2}, \tag{30}$$

i.e. $\alpha\theta_n - v_n = \sqrt{\alpha\beta(v_{n+1} - v_{n-1})}$ (because $v_n < \alpha\theta_n$), with $v_0 = 0$ and $v_N = 1$. We can observe that a sufficient condition for the existence of a vector $(v_1, \ldots, v_{N-1})$ that satisfies the previous fixed-point equation is given by:

$$\frac{\alpha}{2}(\theta_{n-1} - \theta_n)^2 > Y(\alpha\theta_{n+1}, \alpha\theta_n) - Y(\alpha\theta_{n-1}, \alpha\theta_{n-2}). \tag{31}$$

This is coming from the fact that $Y(\alpha\theta_n, \alpha\theta_{n-1}) < Y(v_{n+1}, v_n) < Y(\alpha\theta_{n+1}, \alpha\theta_n)$ and that $\frac{v_n^2}{2\alpha} + \theta_n(\frac{\alpha\theta_n}{2} - v_n)$ is a decreasing function in $v_n \in [\alpha\theta_{n-1}, \alpha\theta_n)$, where the maximum value is equal to $\frac{\alpha}{2}(\theta_{n-1} - \theta_n)^2$ and the infimum value is 0. In the case of the uniform distribution and $\alpha = 1$, this condition can simply be satisfied when $\alpha(\theta_{n-1} - \theta_n)^2 > \beta(\theta_{n+1} - \theta_{n-1})$, for all n. Therefore, it always exists β^* such that for $\beta \leq \beta^*$ a equilibrium characterized by (30) exists. For two thresholds, the implicit plots associated to the fixed point systems (30) are depicted in Figure 6. We fix $\theta_1 = \frac{1}{3}$, $\theta_2 = \frac{2}{3}$, $\alpha = 1$ and $\beta = 0.1$. We obtain $v_1 = 0.14$ and $v_2 = 0.37$. With these parameters $W_{Bm} = 0.562$. As a comparison, type-B feedback with $\theta = 1/3$ yields a lower $W_B = 0.55$. Moreover, to illustrate the trade-off between the aggregate prosocial action and the privacy measure, we propose the following numerical experiment. We will restrict ourself to N thresholds, with $N \in \{3, 4, 5, 6\}$. Thresholds are of the form $\theta_n = \frac{n}{N+1}$. Moreover we assume that F follows a uniform distribution having a support in $[0, 1]$, $\alpha = 1$ and $\beta = 0.01$. The results are tabulated in Table 3. We observe that the higher is N (more information is revealed), the higher is the net prosocial action. This

	$N=3$	$N=4$	$N=5$	$N=6$
Aggregate prosocial action (in e^{-1})	5.07	5.08	5.08	5.09
Privacy measure (in e^{-3})	5.8	3.7	2.6	1.9

Table 3. Value of the aggregate prosocial action and the privacy measure when for a given number of thresholds $N \in \{3,4,5,6\}$, $\theta_n = \frac{n}{N+1}$, $\alpha = 1$, $\beta = 0.01$.

observation is in line with the observation made previously for the uniform distribution while comparing type-A feedback and type-B feedback . Moreover, as expected, we observe that the privacy measure decreases with N. This last observation is justifying the need to design carefully type-Bm feedback, to ensure that the net-prosocial action is as high as possible while respecting a certain level of privacy for the agents.

7 Related works.

This section is dedicated to the related works and especially the main difference between our work with (1) the information design paradigm framed in [11] and (2) the model suggested in [8].

Information on information design: This work lies within the framework of information design. In [11], the authors suggested a unified framework for the information design problem. They consider a finite set of agents with a finite set of actions. The payoff of each agent depends on its own type and a state of the nature unknown to her. The designer knows the state and also the type associated with each agent. He will decide an information structure to satisfy his own objective. To provide a characterization of the set of equilibria achieved by controlling the information, the authors define the concept of obedience. Obedience is given by the set of "recommendations" provided by the designer that agents will follow (in a Bayesian setting). This specific set of strategies can be proved to be equal to the set of Bayesian correlated equilibria, which is characterized by linear algebraic constraints. Once this set is known, a Bayesian correlated equilibrium is chosen to maximize the objective of the designer and an extended information structure is suggested by the designer such that the desired Bayesian Nash equilibrium is equal to a Bayesian correlated equilibrium. This can be proven to be equivalent to a linear optimization problem. For a general overview of information design in the previously described set up, from the application point of view or from the algorithmic point of view, see the survey papers [11, 19, 20, 33]. This body of works has significant differences with our approach. In [11], the authors assume that the designer knows the state of nature and will design a probabilistic signal mechanism such that induces agents to behave according to the desire of the designer. A typical example is how to create an incentive for companies to invest in a risky investment by hiding the risky nature of the investment using information design.

We will now describe recent works in information design which can be related to our work, from a conceptual point of view, but are not similar in the resolution and the mathematical proofs. Before starting to describe the different works, we want to point out to the readers the following: One of the main characteristics of these lines of works is the fact that it is possible to reframe the information design problem as a convex optimisation problem, or at least to show that the optimal information design lies on the boundary of a convex set. It is not possible, in our case, even in the simple case of a quadratic cost function, to have a similar result. In [27], the authors study an information design problem where the service provider will decide or not to reveal the state of the queue to Bayesian customers. Based on this observation, the customers will decide to join or not the queue. In this paper, the authors prove that the optimisation problem facing by the service provider can be rewritten as an infinite linear program in the queue's steady-state distribution. They also

prove that it is better for the service provider to conceal the information about the state of the queue. A recent line of work in information design applied to game in networks has been suggested in [13, 14]. They introduce, in a network game with strategic complement, an information designer which can decide to reveal the state of the "world" and therefore, enforce a particular behavior among the agents. The concept of Bayesian Persuasion [6] has also been extended to the case of risk-conscious agents. In our work, we design the information that is supplied to the society. And the designer is indeed committed to a given feedback. However our model is not captured within the framework of information design and the references mentioned above. The reason is the following – our game can be seen as a three players game as mentioned in the appendix C, the designer, a sender (the agent), and a receiver (the society). The designer has to find the optimal design, under the constraint that the sender and the receiver are at the Nash equilibrium (see [16] for a similar set of Nash equilibria). This problem cannot be reformulated as a convex optimisation problem. Indeed, (29) suggests that we cannot characterize the set of Nash equilibria as a convex set. We also use a deterministic signal mechanism instead of a probabilistic mechanism which does not allows us to convexify the set of feedbacks. Beyond static models of information design, recent work studies how optimal signals can be learned over time under model uncertainty. For instance, [12] consider a sequential persuasion setting in which a sender learns to influence a receiver using no-regret type algorithms.

Social comparison and norm based approach: Our model builds on [8], where the authors introduce a game-theoretical framework in which agents choose the intensity of their prosocial action based on intrinsic motivation, monetary rewards, costs, and reputational concerns. In this setting, agents have private information about their intrinsic motivations (prosocial and monetary) and infer the motivations of others through observed actions, leading to a signal extraction problem for reputation. In contrast to [8], where the designer acts *through monetary incentives*, our work focuses on information design. We consider a setting in which the designer observes all actions and controls the information revealed to the population to influence reputational incentives and aggregate prosocial behavior. Extensions of this framework have been studied in [2, 9]. In [8], the authors highlight crowding-out effects, where increasing monetary rewards may decrease prosocial behavior. In [9], the model is extended to allow for imperfect knowledge of the distribution of intrinsic motivations, and the designer can influence behavior by revealing or withholding information about this distribution. In [2], the focus is on the visibility of actions: making actions observable to a larger fraction of the population increases reputational incentives but raises privacy concerns.

Our work is complementary to these approaches. A key common feature is that reputation is modeled as a signal extraction problem. However, the control available to the designer is fundamentally different. In [2, 8, 9], the information structure is either fixed or controlled through the extent of visibility (e.g., the fraction of the population observing an action) or through agents' prior information. In contrast, we directly design the form of the feedback signal by revealing a quantized version of the action. This difference has several implications. First, while existing works assume that actions are either observed or not (and fully revealed when observed), in our model, the feedback is coarse but global, leading to a different type of inference problem. Second, this modification induces qualitatively different equilibrium behavior: in particular, we show that quantized feedback generates discontinuities in best responses and mass points in the equilibrium distribution. Third, it requires new mathematical arguments to establish the existence and characterization of mean field equilibria under such feedback. More generally, most of the existing literature assumes that the information structure is given and studies how incentives or equilibrium interactions shape behavior. In contrast, our approach consists of designing the information itself. By doing so, we

show that modifying the granularity of information can have an impact on equilibrium outcomes. In particular, we prove that revealing less information can increase the aggregate level of prosocial behavior, a result that cannot be captured by models where the information structure is fixed.

Other related works: There is a wide range of applications of information design such as routing game [1, 17], queueing game [5, 31], economic applications concerning persuasion [10, 23], deception in UAVs swarms [18], cyber deception [24] and matching markets [28]. These works are not using the framework of information design defined in [11] or the one defined in our paper. But still, information (size of the queue, roads available) is hidden to improve the efficiency of a given system (delay in a queue, congestion in a city), and in that respect is related to our work.

8 Discussion and extensions.

In this section, we will discuss the different extensions of our work.

Link with field experiments: The field experiment in Kerala [32] provides an illustration of a feedback similar to Type-A – households are given perfect information of their own relative position in the population. Empirically, this intervention led to an average reduction of approximately 1% in electricity consumption. This study shows that this reduction arises through a combination of mechanisms: social comparison and simple feedbacks (and adapted solutions) make energy-saving actions more interesting and easier to implement. Our model provides a framework to better understand the different information designs that can be adopted, and how they could be improved to further reduce electricity consumption.

Equilibrium selection and learning algorithm. An interesting question would be to understand if there is a natural decentralized learning algorithm that converges to the equilibrium induced by the different feedback. For instance, we can imagine a mechanism that mimics the behavior of an agent in the society. She will learn her reputation (or rank) and will adapt her prosocial action over time depending on her current reputation level.

Non-linear reputational benefit. In this paper we study linear reputation benefit in the sense that $\mathbb{E}[w_i \mid L(\cdot)]$ is the expectation of a linear function of w_i. Therefore, we did not observe effects when an agent who is already perceived as prosocial relaxes on the intensity of her prosocial action because it is costly to maintain a high reputation in the society. It would be interesting to extend this work to the case where the reputational benefit is equal to $\mathbb{E}[s(w_i) \mid L(\cdot)]$, where $s(\cdot)$ is a concave function of w_i. Type-B feedback will have similar results, but for type-A, a careful study of the ordinary differential equation that appears in the proof of Theorem 1 will be needed.

Common resource sharing. In the current model, the interaction between the agents is only captured through the reputational benefit. However, it could be that the lesser the level of prosocial actions of agents in the society, the higher of the cost of an agent's action. Indeed, consider a routing game on a network with three roads and with one of them being really cheap (Braess's Paradox). The norm maybe that taking the costly road is a prosocial action. The cost of an agent's action also depends on the fraction of the population that take the same road. In this new framework, we need to extend the classical framework of routing game by adding the reputational benefit to the cost function. There will be a need to adapt the classical results of routing games, and the ones of this paper, to prove the existence of an equilibrium.

Creation of a collective identity. Our focus in this paper has been on how to use social comparison and reputational benefit for inducing prosocial behavior at the individual level. From a larger perspective, however, it would be interesting to explore ways to create and maintain a collective identity or a collective awareness. For example, a group of people in a neighborhood could be induced to form a team and meet team goals of reduced consumption via social comparison with other similar groups. Another example could be to provide the collective with a reduced consumption goal. Then one could indicate how much an individual's prosocial action has contributed towards the collective goal, such as how many kg of CO_2 emission has been saved. The challenge in such collectives is to ensure that interest is sustained over a sufficiently long duration. Methods that can help sustain such interest could be a topic of future research.

Social network for immediate impact. Education can help improve the "starting point", for example, shift the distribution F of the propensity for prosocial action. Policies to improve this starting point may however not be effective in the short term. In contrast, our approach has been to make use of the social network to obtain a more immediate impact. Networks are profoundly changing the way people aggregate preferences, and our approach to make use of social comparison to induce prosocial behavior is in line with this. It may be interesting to see how to combine the long-term policy-based approaches with the social comparison approaches.

9 Conclusions

In this paper, our goal has been to show the importance of information design to improve prosocial behavior. We considered two types of feedback, one without privacy that revealed all actions to all agents in the network, and another that provided only quantized information about agents' actions. We extended the initial model of [8] to more complex information designs which required us to derive new proofs of existence and characterization of the mean field equilibria. When the intrinsic motivation is drawn from the uniform distribution we obtained an explicit expression for the mean field equilibria. Moreover, we identified a setting where it was beneficial to be economical with the feedback information to improve the expected prosocial action. We also demonstrate on a real data how reputational incentive (and its design) can be beneficial to increase prosocial behavior in a society. To refine our design, we formalize a notion of privacy, and then introduce a new feedback mechanism that improves privacy. Sufficient conditions for the existence of an equilibrium are provided and we demonstrate through an example what could be the advantage of this new feedback. Finally, we suggest several possible extensions and future directions which we feel will be of use to researchers in the field.

Acknowledgement: Rajesh Sundaresan acknowledges support from the SERB through grant no. CRG/2019/002975, and from the Centre for Networked Intelligence (a CISCO CSR initiative) of the Indian Institute of Science. From Alexandre Reiffers-Masson side, this work was supported by the C.V. Raman Charpak Fellowship.

References

[1] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar. Informational braess' paradox: The effect of information on traffic congestion. *arXiv preprint arXiv:1601.02039*, 2016.

[2] S. N. Ali and R. Bénabou. Image versus information: Changing societal norms and optimal privacy. Technical report, NBER, 2016.

[3] H. Allcott. Social norms and energy conservation. *J. Public Econ.*, 95(9-10):1082–1095, 2011.

[4] H. Allcott and S. Mullainathan. Behavior and energy policy. *Science*, 327(5970):1204–1205, 2010.

[5] E. Altman and T. Jimenez. Admission control to an m/m/1 queue with partial information. In *Int'l Conf. Anal. Stoch. Model. Tech. Appl.*, pages 12–21, 2013.

[6] J. Anunrojwong, K. Iyer, and D. Lingenbrink. Persuading risk-conscious agents: A geometric approach. *Available at SSRN 3386273*, 2020.
[7] G. L. Authority. Smartmeter energy consumption data in london households. https://data.london.gov.uk/dataset/smartmeter-energy-consumption-data-in-london-households-vqm0d/. Accessed: 2025-12-22.
[8] R. Bénabou and J. Tirole. Incentives and prosocial behavior. *Am. Econ. Rev.*, 96(5):1652–1678, 2006.
[9] R. Bénabou and J. Tirole. Laws and norms. Technical report, NBER, 2011.
[10] D. Bergemann and S. Morris. Information design, bayesian persuasion, and bayes correlated equilibrium. *Am. Econ. Rev.*, 106(5):586–91, 2016.
[11] D. Bergemann and S. Morris. Information design: A unified perspective. *J. Econ. Lit.*, 57(1):44–95, 2019.
[12] M. Bernasconi, M. Castiglioni, A. Marchesi, N. Gatti, and F. Trovò. Sequential information design: Learning to persuade in the dark. *Adv. Neural Inf. Process. Syst.*, 35:15917–15928, 2022.
[13] O. Candogan. Optimality of double intervals in persuasion: A convex programming framework. *Available at SSRN 3452145*, 2019.
[14] O. Candogan and K. Drakopoulos. Optimal signaling of content accuracy: Engagement vs. misinformation. *Oper. Res.*, 68(2):497–515, 2020.
[15] G. G. Corneo. The theory of the open shop trade union reconsidered. *Labour Econ.*, 4(1):71–84, 1997.
[16] V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.
[17] S. Das, E. Kamenica, and R. Mirka. Reducing congestion through information design. In *Proc. 55th Allerton Conf. Commun., Control, Comput.*, pages 1279–1284, 2017.
[18] S. de Charentenay, A. Reiffers-Masson, G. Coppin, C. Lesueur, and J. Petit-Frère. Cooperative deception in swarms against a smart observer. In *Int'l Conf. Game Theory AI Secur.*, pages 215–234, 2025.
[19] S. Dughmi. Algorithmic information structure design: A survey. *ACM SIGecom Exch.*, 15(2):2–24, 2017.
[20] J. Horner and A. Skrzypacz. *Learning, Experimentation, and Information Design*, volume 1 of *Econometric Society Monographs*, pages 63–98. Cambridge University Press, 2017.
[21] I. Jewitt. Notes on the shape of distributions. Technical report, Mimeo, Oxford Univ., 2004.
[22] Kaggle. Smart meters in london (sanitized dataset). https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london. Kaggle dataset. Accessed: 2025-12-22.
[23] E. Kamenica and M. Gentzkow. Bayesian persuasion. *Am. Econ. Rev.*, 101(6):2590–2615, 2011.
[24] A. Kumar, S. Brahma, and C. A. Kamhoua. Towards a nudge-theoretic cyber deception framework. In *Proc. 61st Allerton Conf. Commun., Control, Comput.*, 2025.
[25] J. Li, B. Xia, X. Geng, H. Ming, S. Shakkottai, V. Subramanian, and L. Xie. Energy coupon: A mean field game perspective on demand response in smart grids. *ACM SIGMETRICS Perform. Eval. Rev.*, 43(1):455–456, 2015.
[26] J. Li, B. Xia, X. Geng, H. Ming, S. Shakkottai, V. Subramanian, and L. Xie. Mean field games in nudge systems for societal networks. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 3(4):15, 2018.
[27] D. Lingenbrink and K. Iyer. Optimal signaling mechanisms in unobservable queues. *Oper. Res.*, 67(5):1397–1416, 2019.
[28] M. Ostrovsky and M. Schwarz. Information disclosure and unraveling in matching markets. *Am. Econ. J.: Microecon.*, 2(2):34–63, 2010.
[29] P. W. Schultz, J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius. The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.*, 18(5):429–434, 2007.
[30] P. W. Schultz, J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius. The constructive, destructive, and reconstructive power of social norms: Reprise. *Perspect. Psychol. Sci.*, 13(2):249–254, 2018.
[31] A. Sharma, K. Jagannathan, and L. R. Varshney. Queuing approaches to principal-agent communication under information overload. *IEEE Trans. Inf. Theory*, 63(9):6041–6058, 2017.
[32] R. Sundaresan et al. Developing a framework for using electricity consumption data to drive energy efficiency in the residential sector. Technical report, Indian Inst. Sci., Bangalore, 2017.
[33] I. Taneva. Information design. *Am. Econ. J.: Microecon.*, 11(4):151–85, 2019.
[34] J. Tirole. *Economics for the common good.* 2017.
[35] B. Xia, H. Ming, K.-Y. Lee, Y. Li, Y. Zhou, S. Bansal, S. Shakkottai, and L. Xie. Energycoupon: A case study on incentive-based demand response in smart grid. In *Proc. 8th Int'l Conf. Future Energy Syst.*, pages 80–90, 2017.

A Appendix Overview

This appendix collects deferred proofs, a game-theoretic reformulation, and algorithmic details referenced in the main text.

B Deferred Proofs

B.1 Type-A equilibrium (proof of Theorem 1)

Proof. We now provide the proof of Theorem 1. Since we will consider unilateral deviations, let us view $\mathbb{E}[w_i \mid L_A(\cdot)]$ as a function of a_i, while keeping all other actions fixed. Consider an agent with a specific w_i. For her not to deviate, we must ensure the first order optimality condition which is

$$w_i = \alpha a_i - \beta \frac{\partial \mathbb{E}[w_i \mid L_A(\cdot)]}{\partial a_i}. \tag{32}$$

Since G^* is an equilibrium action profile, (32) must hold for all $i \in \mathcal{I}$. Let us search for those strategies that make the right-hand side of (32) monotone increasing in a_i. By this assumption, whose validity we shall later check for our final solution, revealing a_i is as good as revealing w_i to all agents. This implies that at equilibrium $\mathbb{E}[w_i \mid L_A(\cdot)] = w_i$ for every player $i \in \mathcal{I}$ (i.e., every w_i). Plugging this into (32), we have the following:

$$\mathbb{E}[w_i \mid L_A(\cdot)] = \alpha a_i - \beta \frac{\partial \mathbb{E}[w_i \mid L_A(\cdot)]}{\partial a_i}. \tag{33}$$

Let us define $x(a_i) = \mathbb{E}[w_i \mid L_A(\cdot)]$. We then obtain the differential equation:

$$\dot{x}(a_i) + x(a_i)/\beta = \alpha a_i/\beta. \tag{34}$$

The solution to this linear differential equation is given by:

$$x(a_i) = \alpha(a_i - \beta) + \zeta e^{-\frac{a_i}{\beta}}.$$

For agent i with $w_i = 0$, we must have $a_i = 0$. Indeed as stated previously the reputation of this agent is equal to 0. So any nonzero action only adds to cost, and therefore 0 is the best response. This boundary condition, using (32) and (35), yields $\zeta = -\alpha\beta$, and so

$$x(a_i) = \alpha(a_i - \beta(1 - e^{-\frac{a_i}{\beta}})). \tag{35}$$

By rearranging (35), we have:

$$a_i = \frac{w_i}{\alpha} + \beta(1 - e^{-\frac{a_i}{\beta}}). \tag{36}$$

It is easy to see there is a unique solution a_i to (36) (intersection of a linear function and a concave increasing function that starts at a strictly positive value w_i/α but saturates at $w_i/\alpha + \beta$, see Figure 7). Finally, we check that the monotonicity assumption holds for the obtained $x(a_i)$. From (35), we have $\dot{x}(a_i) = \alpha(1 - e^{-\frac{a_i}{\beta}}) \geq 0$. This concludes the proof.

Moreover we can observe that because $\frac{w_i}{\alpha} - a_i + \beta(1 - e^{-\frac{a_i}{\beta}})$ is decreasing in a_i which satisfied the initial concavity assumption stated at the beginning of the proof. □

B.2 Type-A with general cost (proof of theorem 2)

Proof. The proof follows the same lines as the one of Theorem 1. Let us view $\mathbb{E}[w_i \mid L_A(\cdot)]$ as a function of a_i. Let us consider an agent with a specific w_i. For her not to deviate, the first order optimality condition should be satisfied:

$$w_i = C'(a_i) - \beta \frac{\partial \mathbb{E}[w_i \mid L_A(\cdot)]}{\partial a_i}. \tag{37}$$

Since G^* is an equilibrium action profile, (37) must hold for all $i \in \mathcal{I}$. Now we will focus on strategies that make the right-hand side of (37) monotone increasing in a_i. This implies that at equilibrium

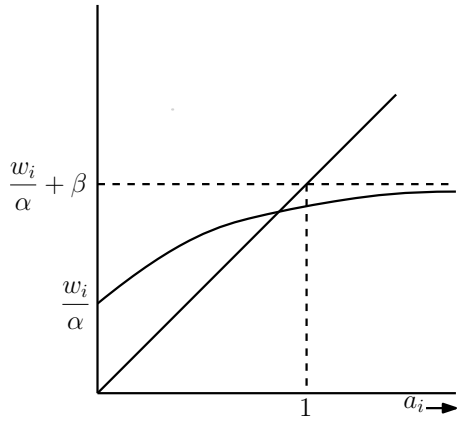


Fig. 7. Intersection between a_i and $\frac{w_i}{\alpha}+\beta(1-e^{-\frac{a_i}{\beta}})$.

$\mathbb{E}[w_i \mid L_A(\cdot)] = w_i$ for every player $i \in \mathcal{I}$ (i.e., every w_i). Plugging $\mathbb{E}[w_i \mid L_A(\cdot)] = w_i$ into (37), we have the following differential equation (in a_i):

$$\mathbb{E}[w_i \mid L_A(\cdot)] = C'(a_i) - \beta \frac{\partial \mathbb{E}[w_i \mid L_A(\cdot)]}{\partial a_i}. \tag{38}$$

Let us define $x(a_i) = \mathbb{E}[w_i \mid L_A(\cdot)]$. By taking the derivative in a_i, in the previous equation, we obtain:

$$\ddot{x}(a_i) = \frac{1}{\beta}(C''(a_i) - \dot{x}(a_i)). \tag{39}$$

The solution to this linear differential equation is given by:

$$\dot{x}(a_i) = \zeta e^{-\frac{a_i}{\beta}} + e^{-\frac{a_i}{\beta}} \int_0^{a_i} C''(b) e^{\frac{b}{\beta}} db. \tag{40}$$

For agent i with $w_i = 0$, we must have $a_i = 0$. Indeed as stated previously the reputation of this agent is equal to 0. So any nonzero action only adds to cost, and therefore 0 is the best response. This boundary condition, using (37) and (40), yields $\zeta = -\Xi(0)$, where $\Xi(a_i) = e^{-\frac{a_i}{\beta}} \int_0^{a_i} C''(b) e^{\frac{b}{\beta}} db$. By rearranging (40), we have:

$$a_i = (C')^{-1}(w_i + \beta(\Xi(a_i) - \Xi(0)e^{-\frac{a_i}{\beta}})). \tag{41}$$

Note that if $\underline{\alpha} \le C''(a_i) \le \overline{\alpha}$, then $\underline{\alpha} \ge \Xi(a_i) \ge \overline{\alpha}$. And therefore, if $\underline{\alpha} > \frac{\overline{\alpha}}{\beta}$, then $\Xi(a_i)$ is strictly increasing in $a_i > 0$. Moreover if $C'''(a_i) < \frac{\underline{\alpha}}{\beta}$ then $\Xi(a_i)$ is concave in a_i. It is easy to see there is a unique solution a_i to (41) (intersection of a linear function and a concave increasing function that starts at a strictly positive value $(C')^{-1}(w_i)$. Finally, we check that the monotonicity assumption holds for the obtained a_i. From (40), we have $C'(a_i) - \beta\dot{x}(a_i) = C'(a_i) + \beta(\Xi(a_i) - \Xi(0)e^{-\frac{a_i}{\beta}}) \ge 0$. Its derivative $C''(a_i) + \beta(\Xi'(a_i) - \Xi(0)e^{-\frac{a_i}{\beta}})$ is positive. This concludes the proof. □

B.3 Type-B best response (proof of propositon 1)

Proof. From measurability considerations, we have

$$\begin{aligned} c_1 &:= \mathbb{E}[w_i \mid L_B(\cdot), a_i < \theta] \text{ is independent of } a_i, \\ c_2 &:= \mathbb{E}[w_i \mid L_B(\cdot), a_i \ge \theta] \text{ is independent of } a_i. \end{aligned}$$

Case 1: Consider an individual with $w_i < \alpha\theta$. Her nonreputational utility function is depicted in Figure 8. Since c_2 is a constant independent of a_i when $a_i \geq \theta$, it is optimal for the agent not to

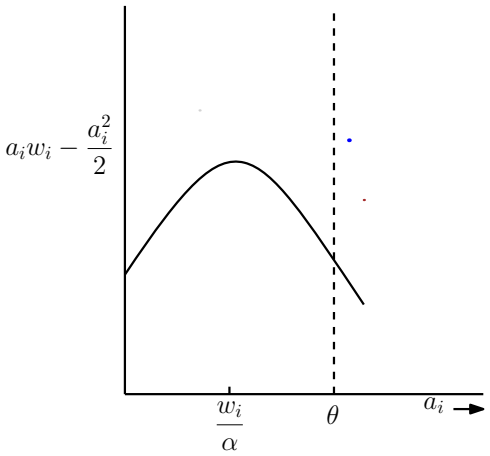


Fig. 8. Case 1: $\theta > \frac{w_i}{\alpha}$

play any a_i other than θ in the set $[\theta, +\infty)$. Now let us consider $a_i < \theta$ versus $a_i = \theta$. When $a_i < \theta$, it is optimal for her to play $a_i = w_i/\alpha$, and she derives the utility:

$$U_B(a_{i1}^*(w_i; G), w_i) = \frac{w_i^2}{2\alpha} + \beta c_1. \tag{42}$$

When $a_i = \theta$, she derives the utility:

$$U_B(\theta, w_i) = \theta w_i - \frac{\alpha\theta^2}{2} + \beta c_2. \tag{43}$$

Clearly then, those individuals with $w_i \in [0, \alpha\theta]$ such that

$$\begin{aligned} \theta w_i - \frac{\alpha\theta^2}{2} + \beta c_2 \geq \frac{w_i^2}{2\alpha} + \beta c_1 \quad &\Leftrightarrow \quad \frac{w_i^2}{2\alpha} - \theta w_i + \frac{\alpha\theta^2}{2} \leq \beta(c_2 - c_1) \\ &\Leftrightarrow \quad \frac{(w_i - \alpha\theta)^2}{2\alpha} \leq \beta(c_2 - c_1) \\ &\Leftrightarrow \quad \mid w_i - \alpha\theta \mid \leq \sqrt{2\alpha\beta(c_2 - c_1)} \\ &\Leftrightarrow \quad -\sqrt{2\alpha\beta(c_2 - c_1)} \leq w_i - \alpha\theta \\ &\qquad \leq \sqrt{2\alpha\beta(c_2 - c_1)}, \end{aligned}$$

will play θ. Since we are considering $w_i < \alpha\theta$, agents with $w_i \in [u, \alpha\theta]$ where:

$$u = \left[\theta\alpha - \sqrt{2\alpha\beta c_2 - c_1}\right]_+ \tag{44}$$

will play θ. Others with $w_i < u$ will play $a_i = w_i/\alpha$. Hence,

$$a^*(w_i; G) \coloneqq \begin{cases} \frac{w_i}{\alpha}, & \text{if } w_i \in [0, u), \\ \theta, & \text{if } w_i \in [u, \alpha\theta). \end{cases} \tag{45}$$

Case 2: Consider now an individual with $w_i \geq \alpha\theta$. Her nonreputational utility function is depicted in Figure 9. Since c_1 is a constant independent of a_i for $a_i < \theta$, and since $c_2 > c_1$, it is optimal for

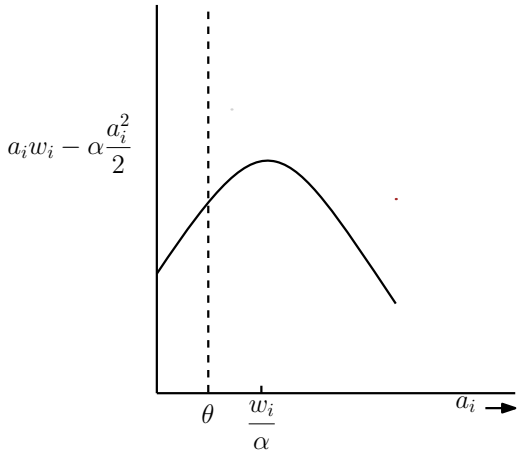


Fig. 9. Case 2: $\theta \leq \frac{w_i}{\alpha}$

her not to play any a_i smaller than θ. It is also clear that if $a_i \geq \theta$, she should pick $a_i = \frac{w_i}{\alpha}$. Hence $a^*(w_i; G) = w_i/\alpha$ for all i with $w_i \geq \alpha\theta$.

Finally, summarizing both cases, we get:

$$a^*(w_i; G) \coloneqq \begin{cases} \theta, & \text{if } w_i \in [u, \alpha\theta), \\ \frac{w_i}{\alpha}, & \text{otherwise,} \end{cases} \tag{46}$$

with u as in (44). This proves Proposition 1.

□

B.4 When Type-B outperforms Type-A (proof of proposition 2)

Proof. If $f(w)$ is decreasing then $X_B(u)$ is increasing for $u \in (0, \alpha\theta)$ (see [21]). If $\beta\mathbb{E}[w] = \lim_{u\to 0} \beta X_B(u) > \frac{\alpha\theta^2}{2}$, then u is always equal to 0 according to Proposition 1. Observe from Theorem 3 that $W_A \leq \frac{\mathbb{E}[w]}{\alpha} + \beta$. Hence,

$$\begin{aligned} W_B - W_A &\geq W_B - \frac{\mathbb{E}[w]}{\alpha} - \beta \\ &\geq \int_0^u \frac{w}{\alpha} f(w)\, dw + \theta \int_u^{\alpha\theta} f(w)\, dw + \int_{\alpha\theta}^{+\infty} \frac{w}{\alpha} f(w)\, dw - \frac{\mathbb{E}[w]}{\alpha} - \beta \\ &= \theta \int_u^{\alpha\theta} f(w)\, dw - \int_u^{\alpha\theta} \frac{w}{\alpha} f(w)\, dw - \beta > 0. \end{aligned}$$

where the last equality follows from (18). This concludes the proof. □

C Game reformulation *à la* Crawford and Sobel

This section follows the strategic information transmission framework of Crawford and Sobel [16].

The socio-economic setting behind the computation of the reputational benefit is the following. (1) The society observes for each agent $i \in [0, 1]$ a quantized feedback $L(a_i)$ of her action. (2) Using this collection of observations, the society estimates, for each agent i, the intrinsic motivation $E[w_i \mid \{L(\cdot)\}]$. (3) Agent i cares about her reputation and derives a reputational benefit $E[w_i \mid \{L(\cdot)\}]$.

From a strict mathematical perspective, we have been ambiguous in not specifying the measure space over which the continuum of random variables $\{a_i\}_{i\in I}$ are measurable. The σ-algebras over which the conditional measures of w_i given $\{L(a_i)\}_i$ are defined has been also left unclear. This is the intuitive language adopted in some literature for ease of exposition without obscure mathematical formalism.

We now provide the necessary mathematical formalism. The game is composed of only two players, the agent (leader/sender) and the society (follower/receiver). Let w be the type of the agent distributed according to $F(\cdot)$. The realization w is known only to the agent, but $F(\cdot)$ is common knowledge. Let $a \in [0, +\infty)$ be the action of the agent. Let $L(a)$ be revealed to the society. Let $y \in [0, +\infty)$ be the action of the society. The players' respective utilities are given by the following.

(1) The agent's utility:

$$U(a, w, y) = aw - C(a) + \beta y, \tag{47}$$

(2) The society's utility:

$$V(w, y) = -(w - y)^2. \tag{48}$$

Ideally, the society wants to choose y equal to w. However, the society does not know w and only observes $L(a)$.

The strategy of the agent, denoted $a(w)$, can be relaxed to be a statistical experiment $\sigma_a : w \to [0, +\infty)$. We now suggest a framework, based on Crawford & Sobel [16], in which the results of this paper can be interpreted in a mathematically rigorous way.

The statistical experiment σ_a yields on an the associated statistical experiment $\sigma_L : w \to [0, +\infty))$. After observing $L(a)$, the society will update the prior belief F to the a posterior belief, whose density is given by

$$\mathbb{P}(w \mid L(a)) = \frac{\sigma_L(L(a) \mid w) f(w)}{\int_0^{+\infty} \sigma_L(L(a) \mid v) f(v)\, dv}. \tag{49}$$

From (49), we can deduce that the society's best reply to σ_a is given by

$$y^*(\sigma_a) = \arg \max_{y\in[0,+\infty)} - \int_0^{+\infty} \mathbb{P}(w \mid L(a))(w - y)^2\, dw = \mathbb{E}[w \mid L(a)]$$

, the minimum mean squared error estimate under the joint law coming from F, σ_a for (X, A) and the associated F, σ_L for (X, L). The best reply to $y(\sigma_a)$ for the agent is given by σ_a^* which is any measure supported on the set $\arg\max_{a\in[0,+\infty)} aw - C(a) + \beta y^*(\sigma_a)$. The equilibrium we study in this paper is $(\sigma_a^*, y^*(\sigma_a^*))$. In the setting of Crawford and Sobel, the L was under the control of the leading player. Our setting differs in that L is under the control of the information designer.

D Algorithms

Algorithm 1 Fixed-point computation of the Type-B equilibrium

Require: Intrinsic types $(w_i)_{i=1}^N$, threshold θ, parameters α, β
Ensure: Equilibrium actions $(a_i)_{i=1}^N$ and reputation level $\bar{w}$

Compute baseline best responses:

$$a_i^{\mathrm{br}} \leftarrow \frac{w_i}{\alpha}, \quad i = 1, \dots, N$$

Initialize the reputation level:

$$\overline{w}^{(0)} \leftarrow \mathbb{E}\big[w_i \mid a_i^{\mathrm{br}} > \theta\big],\ \underline{w}^{(0)} \leftarrow \mathbb{E}\big[w_i \mid a_i^{\mathrm{br}} \le \theta\big]$$

for $k = 0, 1, 2, \dots$ until convergence **do**
  for $i = 1, \dots, N$ **do**
    if $\frac{w_i}{\alpha} \le \theta$ **then**
      Compute utilities given $\bar{w}^{(k)}$:

$$U_i^{\mathrm{jump}} \leftarrow w_i\theta - \frac{\alpha}{2}\theta^2 + \beta\bar{w}^{(k)},\ U_i^{\mathrm{br}} \leftarrow w_i\big(\frac{w_i}{\alpha}\big) - \frac{\alpha}{2}\big(\frac{w_i}{\alpha}\big)^2$$

      Best-response actions:

$$a_i^{(k)} \leftarrow \begin{cases} \theta & \text{if } U_i^{\mathrm{jump}} \ge U_i^{\mathrm{br}}, \\ \frac{w_i}{\alpha} & \text{otherwise} \end{cases}$$

    else

$$a_i^{(k)} \leftarrow \frac{w_i}{\alpha}$$

    end if
  end for
  Update reputation level:

$$\overline{w}^{(k+1)} \leftarrow \mathbb{E}\big[w_i \mid a_i^{(k)} > \theta\big],\ \underline{w}^{(k+1)} \leftarrow \mathbb{E}\big[w_i \mid a_i^{(k)} \le \theta\big]$$

  if $|\overline{w}^{(k+1)} - \overline{w}^{(k)}| < \varepsilon$ **then**
    break
  end if
end for
return $(a_i^{(k)}, \overline{w}^{(k)})$

Received January 2026; revised March 2026; accepted April 2026