Strategic Obfuscation:

Influence Through Information Costs

Luis Henrique Linhares Daniel Monte*

November 6, 2026

Click here for the latest version.

Abstract

We develop a model of strategic obfuscation in which an interested party (the Obfuscator) chooses the cost for a rationally inattentive decision-maker (the Learner) to process information. We characterize when the Obfuscator chooses full transparency, full obfuscation, or partial obfuscation. Endogenous information costs can lead to a counterintuitive implication: increasing the Learner's cost of mistakes can improve her resulting welfare as it leads the Obfuscator to favor more transparency. The framework links persuasion and rational inattention, shifting attention from commitment in disclosure to the design of information costs. The model helps explain fine print and product complexity and has implications for disclosure and transparency policy.

Keywords: strategic obfuscation, rational inattention, costly information

JEL Codes: D11, D82, D83

*Linhares: Boston University. E-mail: lhrlinha@bu.edu. Monte: São Paulo School of Economics - FGV. E-mail: daniel.monte@fgv.br. We are very thankful for conversations with Henrique de Oliveira, Lucas Maestri, Chiara Margaria, Laurent Mathevet, Jawwad Noor, Juan Ortner, Kareen Rozen, and Adrien Vigier.

1 Introduction

Obfuscation is ubiquitous. Lawyers required to disclose evidence may bury the relevant material amid a large volume of irrelevant documents. Terms and conditions are long and confusing, so reading them before accepting is the exception, not the rule. Direct product comparison is made harder when firms offer products that differ in size and flavor from those of their competitors (Richards et al., 2020). The reasoning is simple: a decision maker wants information to make a better choice, so they spend time and effort to acquire it. But if an interested party can affect how costly it is to obtain information, they can influence what is learned and, ultimately, the decision.

In this project we study strategic obfuscation. We seek to understand what drives an agent, the Obfuscator, to make information either easier or harder to process for another, whom we call the Learner.

Our model borrows from the Bayesian persuasion and rational inattention literatures (Kamenica and Gentzkow, 2011; Gentzkow and Kamenica, 2014; Matysková and Montes, 2023). In our setting, the Obfuscator can influence the Learner's choice solely by choosing how costly it is to obtain new, relevant information. The Learner is rationally inattentive (Sims, 2003): she chooses how much to learn by balancing the benefit of a better decision against the cost of learning. Our setting therefore bridges the persuasion, rational inattention (RI) and strategic obfuscation literatures in a novel way.

A key contribution of our paper is to focus on the design of which information can feasibly be learned. As such, we provide an approach to persuasion that does not require a strong commitment assumption.[1] We assume the cost of information is chosen before the Obfuscator observes the state of the world and thus, the chosen level of obfuscation is not a signal in itself.[2] We go over a comparison of Strategic Obfuscation with Bayesian Persuasion and

[1]Thus, our model contrasts with Bayesian Persuasion (see Kamenica and Gentzkow (2011) but also Kamenica (2019) and Bergemann and Morris (2019) for comprehensive surveys), in which a Sender commits to a disclosure policy. In our paper, the Obfuscator (essentially) chooses how the Learner experiments, within the RI framework. Indeed, Kamenica and Gentzkow (2011) use commitment in the information acquisition stage instead of in the information transmission stage as a possible motivation for the commitment assumption.

[2]We comment on this when presenting the model and describing practical examples. Martin (2017) and de Clippel and Rozen (2025) study strategic aspects of obfuscation. Where the Obfuscator cannot commit, we interpret our results as what he could achieve with commitment, providing an upper bound to his payoff.

Cheap Talk in Section 5.3.

The Obfuscator's efforts to raise costs can be interpreted in different ways: restricting access to information, decreasing the time available to make a decision, or simply dumping irrelevant data. Unlike other obfuscation papers, we do not focus on a single mechanism. Instead, we generalize how this behavior can appear when facing a rationally inattentive decision-maker. A field experiment on StubHub illustrates how making information harder to obtain can increase purchases. Blake et al. (2021) find that shoppers were 14.1% more likely to make a purchase when the platform delayed mandatory fees until checkout than when it showed the total price upfront. Shoppers could still see the full price before buying, but they had to go farther into the purchase process to find it, making tickets harder to compare.

Mathematically, the Obfuscator chooses a parameter $\lambda \geqslant 0$, at no direct cost to himself,[3] that scales the cost of information. We use an entropic cost, proportional to the expected reduction in Shannon entropy (Shannon, 1948).[4] After observing the cost, the Learner chooses how and what to learn. By setting $\lambda = 0$, the Obfuscator effectively lets the Learner correctly identify the state of the world. By setting a sufficiently large λ, the Obfuscator can make any information processing prohibitively costly, and the Learner must base her decisions on the common prior alone.

Our main results characterize the chosen level of obfuscation based only on how aligned interests are and on the previously available common information, i.e., the prior. Our characterization allows us to identify three obfuscation regimes and when they arise: (i) full transparency, when there is no cost of information; (ii) full obfuscation, when there is no learning; and (iii) partial obfuscation, when the Obfuscator confuses "just a little bit". In this last case, some information is acquired, but mistakes still happen with positive probability.

Partial obfuscation is particularly revealing about the mechanisms at the Obfuscator's disposal. We find they operate along two distinct margins. Along the *extensive margin*, increasing costs changes which options the Learner considers.[5] As information becomes more

[3]Hiding or providing information may be costly to the Obfuscator. We abstract from such costs. Appendix E briefly discusses an extended model in which the Obfuscator incurs a cost of obfuscation. Adding these costs, however, does not alter our main conclusions.

[4]Section 6 discusses how our results generalize when using the family of cost functions from Caplin et al. (2022).

[5]These are the same consideration sets as in Caplin et al. (2019).

expensive, she may drop options with a low probability of being beneficial. Removing some options may benefit the Obfuscator, but further obfuscation risks eliminating options he wants the Learner to investigate. The Obfuscator resolves this trade-off by selecting an intermediate cost that induces his optimal consideration set.

Along the *intensive margin*, obfuscation can distort how often the Learner chooses each action while keeping the considered options fixed. This happens when some states have costlier mistakes. The Learner then tolerates more frequent errors in states where mistakes are relatively harmless in order to save on attention. The Obfuscator can exploit this by choosing an intermediate cost that induces the Learner to make mistakes in his favor, yielding an expected payoff that strictly exceeds both full transparency and full obfuscation.

Specializing to two states and two actions lets us isolate this second mechanism. Whenever the Learner acquires information, both actions remain in consideration, so we effectively shut down the extensive margin effects. Partial obfuscation is then optimal exactly when the Learner is *skeptical*: she rejects the Obfuscator's preferred action under the prior, even though her threshold for taking it is below one-half.

To illustrate, consider a firm selling a product of uncertain value to a consumer. The product may be *Good* for her, in which case buying gives a benefit, or *Bad*, in which case she would regret the purchase. Both the firm and the consumer are initially unsure whether the product suits her and share a prior probability $\mu_0 \in (0,1)$ that it is Good. The firm always benefits from a sale. The payoff structure is

	not buy	buy
Good	$0, 0$	$1, B$
Bad	$0, 0$	$1, -C$.

The first entry in each cell represents the firm's payoff and the second the consumer's. Suppose $B > C > 0$, so missing a good product is costlier than buying a bad one. The consumer buys without further information when μ_0 exceeds the belief of indifference $\mu_{ind} = \frac{C}{B+C}$. In that case, the firm chooses a sufficiently high finite information cost and preserves the sale. This is the sales adage "don't talk yourself out of a sale".

If instead $\mu_0 < \mu_{ind}$, the consumer would not buy using the prior alone. Full transparency

lets her identify whether the product is Good or Bad and yields a sale only when it is Good. At a suitable intermediate cost, she saves on information by choosing a signal that recommends buying more often, accepting some bad purchases. This increases sales relative to full transparency. Raising costs further eventually makes learning unprofitable, and she returns to not buying. The firm therefore prefers partial obfuscation to either extreme.

Because transparency is often desirable from a social perspective, we show how this taxonomy can guide policy design. In particular, mandatory disclosure is often ineffective (think back to the long terms and conditions mentioned above). Thus, understanding when the Obfuscator chooses partial or total obfuscation reveals how alternative regulatory interventions can incentivize full transparency.

In the binary example, a policy that lowers the loss C from buying a Bad product helps the consumer if she makes a mistake. It can also make her more willing to buy without clear evidence, giving the firm more reason to obfuscate. Theorem 2 shows that, holding the payoffs from correct choices fixed, making a mistake sufficiently costly can induce full transparency and restore the consumer's first-best payoff. Proposition 8 identifies cases in which even a small increase in the cost of mistakes improves her equilibrium welfare. This is not a general prescription to make mistakes more costly. Rather, it illustrates why interventions must account for the Obfuscator's endogenous response.

Consider a policy that requires firms to give consumers more time to return a product for a refund. For a buyer who discovers that the product is a poor fit, the option to return it reduces the loss C. She may then be more willing to buy without checking the details. If the firm still profits from sales that closer inspection would have deterred, it may have more reason to make the product's drawbacks harder to discover before purchase.

The rest of the paper is organized as follows. Section 2 presents the model and both players' problems. Section 3 characterizes when partial obfuscation is optimal. Sections 4 and 5 specialize to two states and two actions and study welfare for the Learner. Section 5.3 compares Strategic Obfuscation with Bayesian Persuasion and Cheap Talk, and Section 6 examines other information cost functions. The Appendix contains the proofs and model extensions.

Related literature:

Information Design and the commitment assumption - In our framework, commitment concerns the cost of learning. In canonical Bayesian Persuasion models (Kamenica and Gentzkow (2011) and Bergemann and Morris (2019), for example), it concerns information transmission. Our paper is related to work by Lin and Liu (2024), Barros (2026), Lipnowski et al. (2022), Min (2021), Fréchette et al. (2022), Nguyen and Tan (2021), Perez-Richet and Skreta (2022) and Alonso and Camara (2024). Also related are Edmond and Lu (2021), who allow the Sender to confuse the receiver at a cost, and Bizzotto et al. (2020), who study persuasion with certification and costly information acquisition.

Rational Inattention - To solve the Learner's problem, we draw on the foundations for analyzing rationally inattentive agents (Matějka and McKay, 2015; De Oliveira et al., 2017; Caplin et al., 2019, 2022; Dean and Neligh, 2023). Ellis (2018) provides an axiomatic model of choice under limited attention and conditions under which behavior reflects an optimal choice of inattention. Following this literature, a growing body of work has applied rational inattention to industrial organization. Examples include Matějka and McKay (2012), Matějka (2016), and Boyacı and Akçay (2018). More recently, Janssen and Kasinger (2024) analyze a duopoly in which firms can obfuscate prices before competing for rationally inattentive consumers, while Cusumano et al. (2024) study competition between two firms selling to a rationally inattentive consumer. Asriyan et al. (2023) show how buyers' imperfect observation of product quality affects the complexity chosen by product designers.

Information Design and Rational Inattention - A related branch of the literature combines persuasion with Rational Inattention. Gentzkow and Kamenica (2014) provide a direct extension of the classic Bayesian Persuasion model in which the sender has to buy the signal at a cost given by Shannon Entropy. Matysková and Montes (2023) develop a model in which the Sender sends information at no cost. However, upon seeing a signal realization, the receiver may buy further information at a cost. Bloedel and Segal (2021) detail how to optimally persuade a rationally inattentive receiver, as the usual concavification results are not attainable. Other related papers are by Lipnowski et al. (2020), Le Treust and Tomala (2019), Fabbri (2024), and Wei (2021). Our contribution is to make the information cost endogenous and study its strategic choice.

Strategic Obfuscation - A growing literature on obfuscation, typically in IO contexts, motivates our economic problem. See Gabaix and Laibson (2006); Ellison (2005), Ellison and Wolitzky (2012), Kalayci and Potters (2011), Jin et al. (2022), Ben-Shahar and Schneider (2014), and Nocke and Rey (2024). More recently, de Clippel and Rozen (2025) develop and empirically test a model of communication in which obfuscation intensity is endogenous and may appear in settings with mandatory disclosure. The receiver draws inferences from the sender's obfuscation choice. In contrast, our results characterize outcomes when the cost of information is chosen before the Obfuscator learns the state, so that Learners make no inference from obfuscation per se.

2 The Model

There are two players, the Obfuscator (*he*) and the Learner (*she*). We denote by $\Omega = \{\omega_1, \ldots, \omega_N\}$ the finite set of payoff-relevant states of the world and $A = \{a_1, \ldots, a_N\}$ the set of actions. The state is drawn from a common, full-support prior $\mu_0 \in int(\triangle(\Omega))$, known by both players,[6] but neither player observes its realization before the information acquisition stage.

The Learner wants to choose the action that matches the state. Her material payoffs are

$$u(a_i, \omega_j) = \alpha_j \mathbb{1}\{i = j\}, \tag{1}$$

where $\alpha_j > 0$ for every j. A correct choice gives a positive reward, while a mistake gives payoff zero. The rewards α_j may differ across states, so avoiding a mistake can be more important in some states than in others. The Obfuscator has a material utility function $v : A \to \mathbb{R}$ that depends only on the action chosen. His payoffs $v(a_1), \ldots, v(a_N)$ may all differ.[7]

We let $a^*(\mu) \coloneqq \arg\max_{a_i \in A} \mathbb{E}_\mu[u(a_i, \omega)] = \arg\max_{a_i \in A} \alpha_i \mu(\omega_i)$ denote the set of actions that are optimal for the Learner at belief μ. We break ties in favor of the Obfuscator, both

[6] We denote by $\triangle(\Omega)$ the space of all probability distributions on Ω, and by $int(X)$ the interior of set X, taken relative to the belief simplex when X is a set of beliefs.

[7] Appendix F presents the model in a version with arbitrary finite Learner payoffs, possibly different numbers of states and actions, and state-dependent Obfuscator payoffs.

when the Learner chooses an action and when she chooses how much to learn.[8] Our solution concept is the Obfuscator-preferred subgame perfect equilibrium, which we simply refer to as "equilibrium" throughout the paper.

The Learner chooses a Blackwell experiment π whose realizations are interpreted as recommended actions.[9] We will work directly with the induced distribution of posteriors $\tau \in \triangle(\triangle(\Omega))$ (Kamenica and Gentzkow, 2011).

The Learner is rationally inattentive. Upon choosing a distribution of posteriors τ, she incurs an information cost given by

$$c_\lambda(\tau) \coloneqq \lambda(H(\mu_0) - \mathbb{E}_\tau[H(\mu)]), \tag{2}$$

where $\lambda \geqslant 0$ is the marginal cost of information and $H : \triangle(\Omega) \to \mathbb{R}_+$ is the Shannon entropy (Shannon, 1948),

$$H(\mu) = -\sum_{m=1}^{N} \mu(\omega_m) \ln(\mu(\omega_m)).$$

Entropy measures uncertainty, so λ is the price of reducing it.[10]

There are two ways a signal can have zero cost: (i) the Obfuscator sets $\lambda = 0$ and learning is free, or (ii) the Learner chooses an uninformative signal, so the posterior equals the prior and there is no change in entropy.

Before either player knows the state, the Obfuscator chooses and commits to λ at no direct cost to himself. The Learner observes this cost and chooses τ. When recommendation a is realized, she updates to posterior μ^a by Bayes' rule.[11] The Learner then takes the recommended action, and payoffs are realized. Since the Obfuscator does not know the state when he chooses λ, the chosen cost does not itself reveal information. This setting is suited

[8]This is the same tie-breaking convention common in the Bayesian Persuasion literature (Kamenica and Gentzkow, 2011).

[9]We can take signal realizations to be action recommendations without loss of generality. Combining signals that induce the same action preserves the Learner's material payoff and weakly reduces information costs (Matějka and McKay, 2015).

[10]For beliefs on the boundary of $\triangle(\Omega)$, we let $0 \cdot \ln(0) \coloneqq \lim_{x \to 0^+} x \ln x = 0$, as usual.

[11]Let $\pi(a|\omega)$ be the probability of recommendation a in state ω. For a recommendation with positive probability, the posterior is

$$\mu^a(\omega) = \frac{\pi(a|\omega)\mu_0(\omega)}{\sum_{m=1}^{N} \pi(a|\omega_m)\mu_0(\omega_m)}.$$

to applications in which there is uncertainty on both sides: think of a firm interacting with a possible customer for the first time.[12]

2.1 The Learner's Rational Inattention Problem

Since the Obfuscator has to anticipate what the Learner will do given a value of λ, it is important to first understand how she will optimally choose to learn. Two opposing forces are in play: the Learner wants to learn the state of nature to choose the best possible action, but she also wants to avoid learning costs. In what follows, let $U(\mu) \coloneqq \max_{a_i \in A} \mathbb{E}_\mu[u(a_i, \omega)] = \max_i \alpha_i \mu(\omega_i)$ be the indirect utility for the Learner at a belief μ.

Given μ_0 and $\lambda \in \mathbb{R}_+$, the Learner solves

$$\max_{\tau \in \triangle(\triangle(\Omega))} \mathbb{E}_\tau \left[U(\mu)\right] - \lambda(H(\mu_0) - \mathbb{E}_\tau[H(\mu)]), \tag{3}$$
$$\text{s.t. } \mathbb{E}_\tau[\mu] = \mu_0.$$

The constraint imposes Bayes plausibility. We denote by $T_{\mu_0}(\lambda)$ the set of solutions to (3) and by τ_λ (with corresponding experiment π_λ) a solution selected under our tie-breaking rule. For its experiment π_λ, let $P_\lambda(a) \coloneqq \sum_{m=1}^N \mu_0(\omega_m)\pi_\lambda(a|\omega_m)$ be the unconditional probability that action a is recommended and chosen. In the direct representation, each posterior has one selected action. If several actions remain tied after applying both players' preferences, we select a fixed one. For $P_\lambda(a) > 0$, we have $P_\lambda(a) = \tau_\lambda(\mu^a)$.[13]

Following Caplin et al. (2019), we call $A(\tau_\lambda) = \{a \in A : P_\lambda(a) > 0\}$ the Learner's *consideration set*. In their analysis of rational inattention, limited consideration arises because an optimal experiment may recommend only some of the available actions. In our setting, the Obfuscator can influence which actions the Learner considers. The Obfuscator chooses λ to maximize his expected payoff from the induced action frequencies.

[12] In a similar fashion, Li and Shi (2017) provide a model in which a seller can allow the buyer to obtain more information about the valuation of a good. Neither the seller nor the buyer has perfect information about the valuation beforehand.

[13] The equivalent formulation in terms of conditional action probabilities is given in Corollary 1 of Matějka and McKay (2015).

2.2 The Obfuscator's Problem

The Obfuscator chooses the cost of learning to influence the Learner's eventual action. At a belief μ, his indirect utility is

$$V(\mu) \coloneqq \max_{a_i \in a^*(\mu)} v(a_i).$$

The maximization over $a^*(\mu)$ reflects the tie-breaking rule. If there are multiple optimal distributions of posteriors, the Learner also chooses the one favored by the Obfuscator. Thus, his expected utility from choosing λ is

$$\tilde{V}_{\mu_0}(\lambda) \coloneqq \max_{\tau \in T_{\mu_0}(\lambda)} \mathbb{E}_\tau[V(\mu)].$$

For a selected solution τ_λ, we can write this payoff as

$$\tilde{V}_{\mu_0}(\lambda) = \sum_{a_i \in A(\tau_\lambda)} P_\lambda(a_i) v(a_i). \tag{4}$$

Intuitively, the Obfuscator chooses the cost that induces the most favorable distribution over actions. We use the prior as an index to highlight that it affects the Obfuscator's choice.

The Obfuscator solves

$$\sup_{\lambda \in \mathbb{R}_+} \tilde{V}_{\mu_0}(\lambda). \tag{5}$$

We solve the game by backward induction. Subsection 2.3 characterizes the Learner's response to each information cost. Subsection 2.4 establishes when the Obfuscator is guaranteed an optimal choice, and Section 3 characterizes the resulting obfuscation regimes.

2.3 Solving the Learner's Problem

The Learner takes the marginal cost of information λ as given. She chooses which distinctions are worth paying for, weighing the benefit of a better decision against the cost of reducing uncertainty. We characterize her optimal experiment and the action probabilities it induces.

2.3.1 Optimal Information Acquisition

We use the posterior characterization in Caplin et al. (2019). Its conditions have three roles: (i) Bayes plausibility ensures that the posteriors average to the prior; (ii) for actions that are used, posterior likelihood ratios reflect the payoff advantages of the recommended actions; (iii) for actions left out of consideration, adding another recommendation cannot improve the Learner's net payoff. The following lemma states these conditions formally.[14]

Lemma 1 (Proposition 2 in Caplin et al. (2019))**.** *Fix $\lambda > 0$. A distribution τ_λ, with supported posteriors $\{\mu^a : a \in A(\tau_\lambda)\}$, solves the Learner's Rational Inattention Problem if and only if $\sum_{a \in A(\tau_\lambda)} \tau_\lambda(\mu^a)\mu^a(\omega) = \mu_0(\omega)$ for every $\omega \in \Omega$ and the following conditions hold:*

1. *For any $\omega \in \Omega$ and any $a_i, a_j \in A(\tau_\lambda)$,*

$$\frac{\mu^{a_i}(\omega)}{\exp\left(\frac{u(a_i,\omega)}{\lambda}\right)} = \frac{\mu^{a_j}(\omega)}{\exp\left(\frac{u(a_j,\omega)}{\lambda}\right)}. \tag{6}$$

2. *Given $a_i, a_j \in A$ such that $a_i \in A(\tau_\lambda)$, and $a_j \notin A(\tau_\lambda)$,*

$$\sum_{m=1}^{N} \mu^{a_i}(\omega_m) \exp\left(\frac{u(a_j,\omega_m) - u(a_i,\omega_m)}{\lambda}\right) \leqslant 1. \tag{7}$$

These conditions characterize the supported posteriors. We use them to identify the priors at which no learning is optimal and to derive the posterior thresholds in the binary case. The next lemma gives the action probabilities directly, which we use to characterize obfuscation in Section 3.

For a fixed λ, the Learner's solution can exhibit Locally Invariant Posteriors (Caplin et al., 2022): within a region of priors, the same posteriors remain optimal and changes in the prior are accommodated by changes in their probabilities through Bayes plausibility. This invariance is local; a change in the prior can also change the consideration set and posterior support. Moreover, in equilibrium, the prior affects the Obfuscator's choice of λ. Two priors that induce the same supported posteriors at a given cost may therefore lead to different information costs and different posterior supports once the Obfuscator chooses how much to obfuscate.

[14]The characterization comes from applying the KKT conditions to Equation (3). Matějka and McKay (2015) and Ravid (2020) also discuss how to solve this problem.

2.3.2 Optimal Action Probabilities

The Obfuscator's payoff depends on how often each action is chosen. Using the choice-probability approach of Matějka and McKay (2015), the matching payoffs give a convenient expression for these frequencies. For $\lambda > 0$, let

$$b_j(\lambda) \coloneqq \frac{1}{e^{\alpha_j/\lambda} - 1}. \tag{8}$$

The next lemma describes the optimal action probabilities without first solving for the posteriors.

Lemma 2 (Optimal Action Probabilities with Matching Payoffs)**.** *For every $\lambda > 0$, the Learner's optimal unconditional action probabilities are unique and satisfy*

$$P_\lambda(a_i) = \max\left\{\frac{\mu_0(\omega_i)}{\nu(\lambda)} - b_i(\lambda), 0\right\}, \tag{9}$$

where $\nu(\lambda) > 0$ is the unique value that makes $\sum_{i=1}^{N} P_\lambda(a_i) = 1$.

The proof is in Appendix A.1. An action receives positive weight when $\mu_0(\omega_i) > \nu(\lambda) b_i(\lambda)$; otherwise, it is excluded from consideration. Changing the information cost therefore affects both how often an action is chosen and whether it is chosen at all. Both effects matter for the Obfuscator's payoff in Equation (4).

2.4 Existence of a Solution to the Obfuscator's Problem

The Obfuscator can choose any nonnegative information cost, so the first question is whether his payoff attains its supremum at a finite cost. If no learning is uniquely optimal at some finite cost $\overline{\lambda}$, then at every higher cost the Learner still takes the same action based on her prior, and the Obfuscator receives the same payoff. Any such $\overline{\lambda}$ suffices to bound the search. In the binary case we use the first cost that induces no learning. We can therefore restrict his choice to the compact interval $[0, \overline{\lambda}]$, where upper semicontinuity ensures that a maximum is attained.

2.4.1 No Learning at a Given Cost

For a given $\lambda > 0$, let $I(\lambda; a_i)$ be the set of priors at which choosing a_i without learning is optimal. Formally, by Lemma 1, we have

$$I(\lambda; a_i) \coloneqq \left\{ \mu \in \triangle(\Omega) : \sum_{m=1}^{N} \mu(\omega_m) \exp\left(\frac{u(a_j, \omega_m) - u(a_i, \omega_m)}{\lambda} \right) \leqslant 1, \forall j \neq i \right\}. \quad (10)$$

These inequalities come from applying the excluded-action condition in Lemma 1 to an experiment that recommends only a_i.[15] Lemma 1 also implies that, for $\lambda > 0$, every posterior reached after a recommendation a_i lies in $I(\lambda; a_i)$. Thus,

$$\text{supp}(\tau_\lambda) \subseteq \bigcup_{a_i \in A} I(\lambda; a_i).$$

This gives a useful interpretation of optimal learning. The Learner may want to acquire information at the prior, but each induced posterior justifies its recommended action without further learning. At those beliefs, buying more information at the same marginal cost would not improve her net payoff. The no-learning regions therefore describe both when she acts on the prior alone and where an optimal experiment can lead her.

The no-learning regions expand as information becomes more costly: if $\lambda_1 < \lambda_2$, then $I(\lambda_1; a) \subseteq I(\lambda_2; a)$.[16] Thus, once acting on the prior is optimal, raising the information cost cannot make learning worthwhile again. Let $\delta_\mu \in \triangle(\triangle(\Omega))$ denote the distribution of posteriors that puts probability 1 on μ. Therefore, if the Learner chooses δ_μ as the optimal distribution of posteriors, it means that she chose to *learn nothing*, as the posterior equals the prior with probability 1. The next result distinguishes when this no-learning experiment is optimal from when it is the only optimum.

Proposition 1 (Uniqueness of No Learning)**.** *For $\lambda > 0$, we have that $\delta_{\mu_0} \in T_{\mu_0}(\lambda)$ if and only if $\mu_0 \in I(\lambda; a_i)$ for some $a_i \in A$. Moreover, if $\mu_0 \in int(I(\lambda; a_i))$, then $T_{\mu_0}(\lambda) = \{\delta_{\mu_0}\}$.*

The first part characterizes when no learning is optimal. The second ensures that, when the prior lies in the interior of $I(\lambda; a_i)$, every optimal experiment is uninformative and the

[15] No-learning regions at an exogenous information cost also appear in Caplin et al. (2019) and Matysková and Montes (2023). Here, we study them as λ varies because the cost is chosen by the Obfuscator.

[16] For a formal proof, see Proposition 3 in Matysková and Montes (2023). The zero-cost limit of these regions is described in Lemma 4 in the Appendix.

Learner chooses a_i with probability 1. The Obfuscator can then predict her response without having to select among several learning strategies. This uniqueness will be useful when we establish a finite bound on the information costs relevant to his problem.

At $\lambda = 0$, learning is free and the Learner can choose an optimal action in every state. At $\lambda > 0$, she acquires information whenever the prior lies outside all the no-learning regions. This leaves the question of whether any finite cost puts the prior in one of those regions.

2.4.2 The Limits of Full Obfuscation

No learning is optimal at cost λ when $\delta_{\mu_0} \in T_{\mu_0}(\lambda)$. If it is the unique optimum, that cost induces full obfuscation: the Learner takes her prior-optimal action without acquiring information. Proposition 1 gives the conditions for this at a fixed cost. We now ask which priors allow the Obfuscator to induce no learning by choosing a sufficiently high finite cost.

Proposition 2 (The Limits of Full Obfuscation)**.** *For any $\mu \in int(\triangle(\Omega))$, the following statements are equivalent:*

(i) $a^*(\mu)$ is a singleton;

(ii) there exists $\overline{\lambda} \in \mathbb{R}_{++}$ s.t. $\mu \in int(I(\overline{\lambda}; a_i))$ for some $a_i \in a^*(\mu)$.

Proposition 2 characterizes the prior beliefs for which the Learner always learns, no matter the cost. They are exactly the beliefs in which she has more than one optimal action available. We can see why: when the Learner is indifferent between actions, even very coarse information can help break the tie. Therefore, even if λ is arbitrarily large, the Learner can afford a cheap enough signal (by making it sufficiently uninformative) that is still useful. Consequently, inducing full obfuscation is impossible in these cases, and only in these cases.

Corollary 1. *Full obfuscation is possible, i.e., $\delta_{\mu_0} \in T_{\mu_0}(\lambda)$ for some $\lambda \in \mathbb{R}_+$, if and only if $a^*(\mu_0)$ is a singleton.*

2.4.3 Existence of a Solution for the Obfuscator's Problem

Suppose that $a^*(\mu_0) = \{a_k\}$ for some $k \in \{1, \ldots, N\}$. Proposition 2 yields a finite $\overline{\lambda} > 0$ such that $\mu_0 \in int(I(\overline{\lambda}; a_k))$. Since the no-learning regions expand with λ, Proposition 1

implies that every $\lambda \geqslant \overline{\lambda}$ induces no learning and action a_k with probability 1. Hence, $\tilde{V}_{\mu_0}(\lambda) = v(a_k) = V(\mu_0)$ throughout this range.[17]

No cost above $\overline{\lambda}$ can therefore improve on the payoff at $\overline{\lambda}$. We can restrict the supremum in Equation (5) to the compact interval $[0, \overline{\lambda}]$. The remaining step is attainment: our tie-breaking rule makes $\tilde{V}_{\mu_0}$ upper semicontinuous,[18] so it attains a maximum on this interval. This gives the following result.

Proposition 3 (Existence of a Solution)**.** *If $a^*(\mu_0)$ is a singleton, the Obfuscator's Problem has a solution $\lambda^* \in \mathbb{R}_+$.*

The formal proof is in the Appendix. It also applies directly to the broader payoff environment in Appendix F. The following example shows that the Obfuscator's problem may fail to have a solution when the Learner is indifferent between two or more actions at the prior.

Example 1 (Nonexistence at a Tied Prior). Consider two states and two actions with matching rewards $\alpha_1 = 2$ and $\alpha_2 = 1$, and let $\mu_0(\omega_1) = 1/3$. Both actions give the Learner an expected payoff of $2/3$ at the prior. Let $v(a_1) = 1$ and $v(a_2) = 0$. The Obfuscator would like to induce full obfuscation because the tie-breaking rule would then lead the Learner to choose a_1 with probability one. At every finite cost, however, she can benefit from buying a sufficiently coarse informative signal that helps her break the tie. For $\lambda > 0$, Equation (9) gives

$$\tilde{V}_{\mu_0}(\lambda) = P_\lambda(a_1) = \frac{1}{3}\left(1 + \frac{1}{1 + e^{1/\lambda}}\right).$$

Under full transparency, $P_0(a_1) = 1/3$. The Obfuscator's payoff increases toward $1/2$ as $\lambda \to \infty$. The supremum is approached only in the limit, so his problem has no solution. □

We do not view the restriction to a unique prior-optimal action as a limitation, since the Obfuscator's problem has a solution for almost every prior. For fixed matching rewards, any tie at the prior requires $\alpha_i \mu_0(\omega_i) = \alpha_j \mu_0(\omega_j)$ for some $i \neq j$. Each equality defines a lower-dimensional set in the belief simplex, and their finite union has measure zero.

[17] Any finite bound with this property suffices for the existence argument. In Section 4, we identify the smallest no-learning cost directly.

[18] See Appendix A.4 for the proof.

For the rest of the paper, we assume that $a^*(\mu_0)$ is a singleton unless stated otherwise. We can then write the Obfuscator's Problem as

$$\max_{\lambda \in [0,\overline{\lambda}]} \tilde{V}_{\mu_0}(\lambda). \tag{11}$$

We write $\mathcal{O} \coloneqq [0, \overline{\lambda}]$ for this compact search interval. We can now study the Obfuscator's incentives to obfuscate based on the model's primitives.

3 Transparency, Full Obfuscation, and Partial Obfuscation

The Obfuscator chooses the cost that gives him the most favorable distribution of actions. An intermediate cost can benefit him in two ways. It can remove some actions from consideration, or it can shift the probabilities of the actions that remain in consideration when the Learner's rewards differ. We isolate the second mechanism in the 2×2 model in Section 4.

We begin with the following definition.

Definition 1 (Obfuscation Regimes)**.** *We say that the Obfuscator's Problem has*

1. *a full transparency solution if $\lambda = 0$ is optimal;*
2. *a partial obfuscation solution if the maximum payoff strictly exceeds both $\tilde{V}_{\mu_0}(0)$ and the no-learning payoff $V(\mu_0)$;*
3. *a full obfuscation solution if the no-learning payoff $V(\mu_0)$ is maximal.*

We refer to full transparency and full obfuscation as *boundary obfuscation* solutions.

A useful benchmark is *Agreement at the Prior*: the Learner's prior-optimal action a_k also maximizes v over A. Full obfuscation then gives the Obfuscator $v(a_k) = \max_{a_i \in A} v(a_i)$, so no other information cost can improve his payoff. This holds with any number of matching states and actions and formalizes the sales intuition from the introduction.[19]

[19] Appendix F.3 discusses the aligned and zero-sum benchmarks in the broader payoff environment.

However, when the prior-optimal action does not maximize his payoff, the value of learning depends on which other actions it induces and how often they are chosen. We study this trade-off in the next subsection.

3.1 Equal Matching Rewards

Can partial obfuscation be optimal when mistakes have the same cost across states?

To answer this question, suppose that $N \geqslant 2$ and $\alpha_j = \Delta > 0$ for every j. We also assume that the prior probabilities are distinct. Relabel the states and their matching actions so that

$$\mu_0(\omega_1) > \cdots > \mu_0(\omega_N) > 0. \tag{12}$$

Thus, absent new information, the Learner chooses a_1. With equal rewards, the terms $b_j(\lambda)$ in Equation (9) coincide. The positive action probabilities therefore belong to the highest-prior options. As information becomes more expensive, options leave consideration in the order $a_N, a_{N-1}, \ldots, a_2$, until only a_1 remains.

This may give the Obfuscator a reason to stop at an intermediate cost. Removing some options from consideration may benefit him, while further obfuscation removes options he would like the Learner to keep investigating. He chooses the cost, and the Learner determines the resulting consideration set and choice frequencies.

Let $m_k \coloneqq \sum_{i=1}^{k} \mu_0(\omega_i)$ and use the convention $\mu_0(\omega_{N+1}) = 0$. Define

$$R_k \coloneqq \sum_{i=1}^{k} \frac{\mu_0(\omega_i) - \mu_0(\omega_{k+1})}{m_k - k\mu_0(\omega_{k+1})} v(a_i),\ k = 1, \ldots, N. \tag{13}$$

For $k < N$, R_k is the expected payoff at the cost where a_{k+1} has just left consideration and the Learner chooses among $a_1, \ldots, a_k$. At that cost the probability of each remaining action is proportional to how far its prior is above the prior of the action that just left. The endpoint candidates are $R_N = \sum_i \mu_0(\omega_i) v(a_i)$ under full transparency and $R_1 = v(a_1)$ under no learning. The proof is in Appendix A.5.

Proposition 4 (Equal Matching Rewards)**.** *Suppose that $\alpha_j = \Delta > 0$ for every j and that*

Equation (12) *holds. The Obfuscator's optimal payoff is*

$$\max_{\lambda \in \mathcal{O}} \tilde{V}_{\mu_0}(\lambda) = \max_{1 \leqslant k \leqslant N} R_k. \tag{14}$$

We have a partial obfuscation solution if and only if

$$R_k > \max\{R_1, R_N\} \text{ for some } k \in \{2, \ldots, N-1\}. \tag{15}$$

Full transparency is optimal if and only if $R_N = \max_k R_k$*, and full obfuscation is optimal if and only if* $R_1 = \max_k R_k$*.*

Note that R_k is defined only in terms of primitives. Between changes in consideration, the Obfuscator's payoff is monotone or constant. To find an optimal cost, we therefore need only compare the N candidates, from full transparency to the cost at which only a_1 remains.

The result also illustrates one of the mechanisms that induces partial obfuscation. With at least three options, an intermediate consideration set can outperform both learning the state perfectly and acting on the prior alone. What matters is which actions remain and how often they are chosen. With only two actions, however, there is no intermediate candidate between R_N and R_1. We look at the binary case in Section 4 to isolate the other mechanism that can induce partial obfuscation: payoff asymmetries.

One final note. Suppose a_k is the Obfuscator's favorite action, with $2 \leqslant k < N$. He may still prefer a cost at which the Learner considers more than the first k actions. For example, if $\mu_0(\omega_k)$ is close to $\mu_0(\omega_{k+1})$, the probability of choosing a_k may already be very small by the time a_{k+1} leaves consideration. The Obfuscator may then prefer a lower cost at which a_k is chosen more often.

The next example illustrates the candidate comparison in Proposition 4.

Example 2 (Choosing Insurance). A consumer buys one of three policies with the same premium through a comparison platform.[20] Exactly one policy covers her exposure. Neither the platform nor the consumer knows which policy provides coverage.[21] If the Learner buys the correct policy, she gets a reward of 1, and 0 otherwise. Call the policies High (H),

[20]Insurance comparison platforms such as Insureon and Simply Business let customers compare quotes from multiple insurers in one place.

[21]For instance, this may be the first time the consumer buys coverage through the platform.

Middle (M), and Low (L) in order of their prior probabilities of providing coverage, given by $(0.5, 0.4, 0.1)$.

Each policy is offered by a different insurance company. Because of different contractual arrangements with the different insurers, the platform earns commissions of \$10, \$30, and \$5 from selling H, M, and L, respectively.

The platform chooses how much comparison assistance it provides in its report. For instance, standardized definitions and side-by-side comparisons make investigation easier.

Let $r = e^{1/\lambda}$ for $\lambda > 0$. When all three policies are in the consideration set, the conditions of Caplin et al. (2019) in Lemma 1 give the following posteriors, with entries ordered High, Middle, and Low:

$$\left\{ \mu^H = \frac{(r,1,1)}{r+2},\ \mu^M = \frac{(1,r,1)}{r+2},\ \mu^L = \frac{(1,1,r)}{r+2} \right\}.$$

Bayes plausibility then gives the purchase probabilities:

$$P_\lambda(H) = \frac{r}{2(r-1)},\ P_\lambda(M) = \frac{2r-1}{5(r-1)},\ P_\lambda(L) = \frac{r-8}{10(r-1)}.$$

As λ rises, r falls. At $r = 8$, Low's purchase probability reaches zero and it leaves the consideration set. The consumer then buys High with probability $4/7$ and Middle with probability $3/7$. The platform earns $R_2 = 130/7 \approx 18.57$, compared with $R_3 = 17.50$ under full transparency and $R_1 = 10$ under full obfuscation. By Proposition 4, the platform chooses $\lambda^* = 1/\ln 8$. Figure 1 shows how further obfuscation shifts purchases from Middle to High and lowers the expected commission until the consumer buys High without learning.

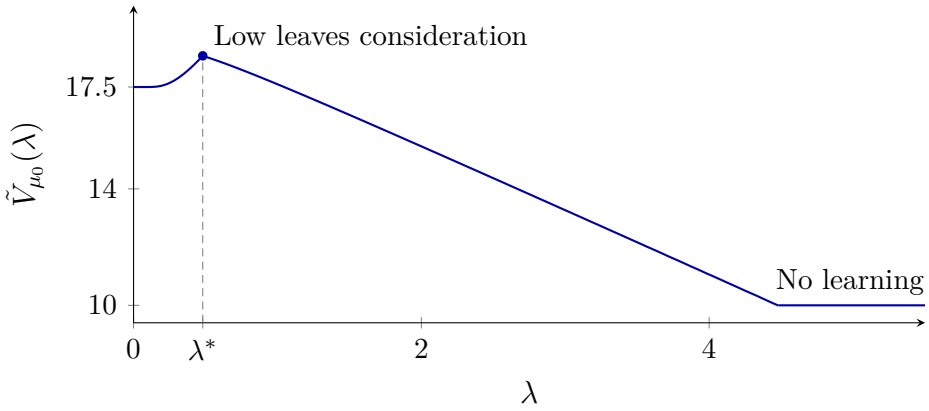


Figure 1: The platform's expected commission in dollars as a function of the information cost.

□

3.2 Characterizing Partial Obfuscation with Unequal Rewards

We now provide the more general characterization of partial obfuscation by allowing the rewards α_j to differ. A rare state can remain worth investigating for the Learner when mistakes in that state are sufficiently costly. Moreover, the different learning margins can respond differently to the information cost. The Obfuscator's payoff can then peak while the consideration set stays unchanged, so the boundary comparison in Proposition 4 is no longer sufficient.

To characterize partial obfuscation in this setting, we combine two requirements. The induced action probabilities must be optimal for the Learner, and their payoff must exceed both endpoint payoffs for the Obfuscator. Write $a^*(\mu_0) = \{a_{i_0}\}$ and let

$$\overline{V} \coloneqq \max\left\{\sum_{j=1}^{N} \mu_0(\omega_j)v(a_j),\ v(a_{i_0})\right\}. \tag{16}$$

This is the better payoff from full transparency and full obfuscation.

For $\lambda > 0$ and a nonempty candidate set $K \subseteq A$, let $m_K \coloneqq \sum_{a_j \in K} \mu_0(\omega_j)$. Using $b_j(\lambda)$ from Equation (8), define for every $a_i \in A$

$$E_{Ki}(\lambda) \coloneqq \mu_0(\omega_i)\left(1 + \sum_{a_j \in K} b_j(\lambda)\right) - m_K b_i(\lambda). \tag{17}$$

Think of K as a guess about which actions the Learner considers. Requiring the candidate weights in K to sum to one gives $\nu_K(\lambda) = m_K/(1 + \sum_{a_j \in K} b_j(\lambda))$. The term $E_{Ki}(\lambda)$ is proportional to the signed gap between the prior on ω_i and the cutoff $\nu_K(\lambda)b_i(\lambda)$ implied by Equation (9). Thus, if an excluded action's prior is above its cutoff, the Learner would gain by giving it positive weight. Because $\sum_{a_i \in K} E_{Ki}(\lambda) = m_K$, dividing by m_K gives candidate weights that sum to one:

$$P_\lambda^K(a_i) = \begin{cases} E_{Ki}(\lambda)/m_K, & a_i \in K, \\ 0, & a_i \notin K. \end{cases} \tag{18}$$

Nonnegative weights for included actions and the excluded-action inequalities below are exactly the conditions for these probabilities to solve the Learner's problem.[22]

The Obfuscator's gain relative to the better endpoint is

$$G_K(\lambda) \coloneqq \frac{1}{m_K} \sum_{a_i \in K} E_{Ki}(\lambda) v(a_i) - \overline{V}. \tag{19}$$

For a candidate satisfying the Learner's optimality conditions, the first term is the Obfuscator's expected payoff $\tilde{V}_{\mu_0}(\lambda)$. A positive $G_K(\lambda)$ means that the resulting choice frequencies give him a higher payoff than either full transparency or full obfuscation. We can therefore state the following characterization.

Proposition 5 (Characterizing Partial Obfuscation)**.** *For $N \geqslant 2$, the Obfuscator's Problem has a partial obfuscation solution if and only if there are $\lambda \in (0, \overline{\lambda})$ and a nonempty candidate set $K \subseteq A$ such that*

$$\begin{aligned} E_{Ki}(\lambda) &\geqslant 0 \quad \textit{for every } a_i \in K, \\ E_{Kj}(\lambda) &\leqslant 0 \quad \textit{for every } a_j \notin K, \\ G_K(\lambda) &> 0. \end{aligned} \tag{20}$$

The proof is in Appendix A.6.

We can test for partial obfuscation by checking these inequalities over the finitely many candidate sets K and the continuous range of costs λ. This removes the need to solve for the

[22] A candidate set may include an action with zero weight. The actual consideration set consists of its actions with $E_{Ki}(\lambda) > 0$. Allowing zero weights keeps the characterization valid at consideration-set boundaries.

Learner's experiment at each cost.

Unlike the equal-reward case, the inequalities do not restrict the optimum to a change in the consideration set. With unequal rewards, changing the cost can shift choice frequencies in the Obfuscator's favor or against him before any action leaves. Thus, changes in consideration and asymmetries in the stakes of mistakes can both contribute to partial obfuscation. The binary case lets us isolate the latter force and express the characterization directly in terms of payoffs and the prior.

4 The 2×2 Model

In this section, we specialize the model from Section 2 to two states and two actions. We maintain state-independent Obfuscator preferences and the same information acquisition technology.

This specialization allows us to (i) provide a simple geometric interpretation of the Obfuscator's payoff, (ii) express the conditions for partial obfuscation coming from payoff asymmetries, and (iii) perform comparative statics and study welfare. The first two points are useful for understanding the forces that lead to partial obfuscation, while the last point is important for policy implications.

Let $A = \{a_H, a_L\}$ and $\Omega = \{\omega_H, \omega_L\}$. With two states, we identify a belief $\mu \in \triangle(\Omega)$ with the probability that the state is ω_H, i.e., with the number $\mu = \mu(\omega_H) \in [0, 1]$. We use the same convention for the common prior $\mu_0 \in (0, 1)$. We label the Obfuscator's strictly preferred action a_H. Without loss, we normalize $v(a_H) = 1$ and $v(a_L) = 0$ throughout.

Let $B = \alpha_H > 0$ and $C = \alpha_L > 0$ be the Learner's rewards for choosing the matching action in states ω_H and ω_L, respectively. To simplify the exposition, we subtract C from both action payoffs in state ω_L. By Bayes plausibility, every feasible experiment's expected payoff falls by the same amount, $(1 - \mu_0)C$, so the Learner's problem and the Obfuscator's problem are unchanged. Lemma 9 in Appendix B.1 gives the details.

The following table gives the Learner's normalized payoffs.

	a_L	a_H
ω_H	0	B
ω_L	0	$-C$

Think back to the product example in the introduction. Buying a bad product is an ex post mistake that costs the Learner C, while buying a good product yields B. At belief μ, the expected payoff from buying is $\mu B - (1-\mu)C$. The Learner is therefore indifferent between the two actions at

$$\mu_{ind} = \frac{C}{B+C}. \tag{21}$$

The assumption that $a^*(\mu_0)$ is a singleton now becomes $\mu_0 \neq \mu_{ind}$. The Learner's Problem and the Obfuscator's Problem are as before.

4.1 Solving the Learner's Problem

In Subsection 2.3, we characterized the Learner's optimal posteriors using Lemma 1. We now apply those conditions to the two-state, two-action setting.

When the Learner acquires information, the two supported posteriors are pinned down by the cost of information and her payoffs. Once we have these posteriors, we find their probabilities through Bayes plausibility. To write the solution, define the following auxiliary functions for $\lambda > 0$:[23]

$$\underline{\mu}(\lambda) \coloneqq \frac{1 - e^{-C/\lambda}}{e^{B/\lambda} - e^{-C/\lambda}}, \tag{22}$$

$$\overline{\mu}(\lambda) \coloneqq \frac{e^{C/\lambda} - 1}{e^{C/\lambda} - e^{-B/\lambda}}. \tag{23}$$

As illustrated in Figure 2, $\underline{\mu}(\lambda)$ and $\overline{\mu}(\lambda)$ are the binary counterparts of the no-learning regions defined in Section 2.4.1: $I(\lambda; a_L) = [0, \underline{\mu}(\lambda)]$ and $I(\lambda; a_H) = [\overline{\mu}(\lambda), 1]$.[24] To understand how these thresholds characterize the Learner's solution for a given $\lambda > 0$, we can go case by case.

[23]The exponential function arises from the use of Entropy as the cost function.

[24]See also Caplin et al. (2019); Matysková and Montes (2023).

If (i) $\mu_0 \in [0, \underline{\mu}(\lambda)]$, the cost of further information outweighs its benefits. The Learner acquires an uninformative signal, keeps the prior as her posterior, and chooses a_L with probability 1. Analogously, if (ii) $\mu_0 \in [\overline{\mu}(\lambda), 1]$, she acquires no information and chooses a_H with probability 1. If, however, (iii) $\mu_0 \in (\underline{\mu}(\lambda), \overline{\mu}(\lambda))$, learning is worthwhile. The optimal experiment induces two posteriors: $\mu^{a_L} = \underline{\mu}(\lambda)$ after recommendation a_L, and $\mu^{a_H} = \overline{\mu}(\lambda)$ after recommendation a_H.

Bayes plausibility determines how often each recommendation occurs:

$$P_\lambda(a_H) = \frac{\mu_0 - \underline{\mu}(\lambda)}{\overline{\mu}(\lambda) - \underline{\mu}(\lambda)}, \; P_\lambda(a_L) = 1 - P_\lambda(a_H). \tag{24}$$

Indeed, $\mu_0 = P_\lambda(a_H)\overline{\mu}(\lambda) + P_\lambda(a_L)\underline{\mu}(\lambda)$. The thresholds describe what the Learner believes after learning, while these probabilities tell us how often she takes each action.

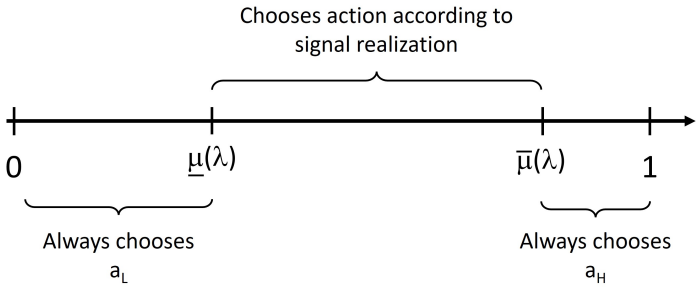


Figure 2: Given λ, the functions $\underline{\mu}$ and $\overline{\mu}$ define the choice by the Learner depending on where the prior lies in the $[0, 1]$ interval.

To see how the Obfuscator will choose λ, it is then sufficient to understand how the thresholds $\underline{\mu}(\lambda)$ and $\overline{\mu}(\lambda)$ change with λ. To ensure continuity, we define

$$\underline{\mu}(0) \coloneqq \lim_{\lambda \to 0^+} \underline{\mu}(\lambda) = 0, \text{ and } \overline{\mu}(0) \coloneqq \lim_{\lambda \to 0^+} \overline{\mu}(\lambda) = 1.$$

These limits follow from $B, C > 0$. At the other extreme, as $\lambda \to \infty$, we get

$$\lim_{\lambda \to \infty} \underline{\mu}(\lambda) = \lim_{\lambda \to \infty} \overline{\mu}(\lambda) = \frac{C}{B + C} = \mu_{ind}. \tag{25}$$

At $\lambda = 0$, the thresholds are 0 and 1. As information becomes more expensive, $\underline{\mu}(\lambda)$ increases and $\overline{\mu}(\lambda)$ decreases toward μ_{ind}. Both functions are continuous and strictly monotone.[25] Thus, $\mu_{ind} \in [\underline{\mu}(\lambda), \overline{\mu}(\lambda)]$ for every finite $\lambda \geqslant 0$.

[25] See Matysková and Montes (2023). Continuity also follows directly from Equations (22) and (23).

The interval of priors at which learning is worthwhile therefore shrinks as λ increases. At high costs, only a Learner sufficiently close to indifference still acquires information. This is the binary version of the no-learning result in Section 2.4.2. Figure 3 illustrates the movement of the thresholds when $\mu_{ind} = 0.5$.

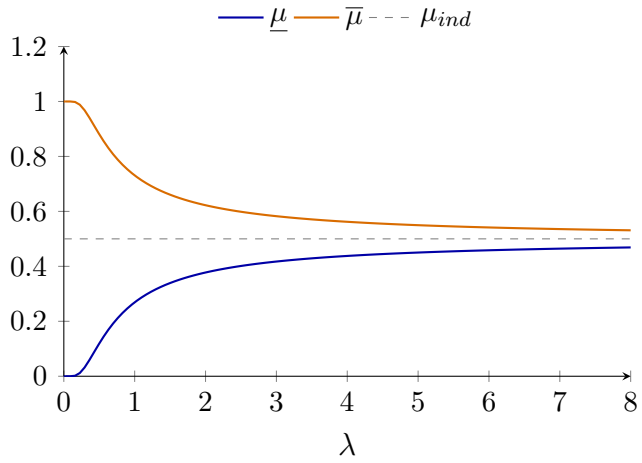


Figure 3: The thresholds $\overline{\mu}$ and $\underline{\mu}$ as functions of λ when $B = C = 1$, so $\mu_{ind} = 0.5$.

4.2 The Obfuscator's Problem

Now that we know how the Learner responds to information costs in the binary case, we can write the Obfuscator's payoff more explicitly. His indirect utility $V(\mu)$ is a step function. It equals zero for $\mu < \mu_{ind}$ and one for $\mu \geqslant \mu_{ind}$, with the equality case determined by our tie-breaking rule. Under the normalization that $v(a_H) = 1$ and $v(a_L) = 0$, his expected payoff is simply the probability that the Learner chooses a_H:

$$\tilde{V}_{\mu_0}(\lambda) = P_{\tau_\lambda}(a_H) = P_\lambda(a_H).$$

Proposition 2 already ensures that a finite cost can induce no learning. In the binary case, we take $\overline{\lambda}$ to be the first cost at which a posterior threshold reaches the prior. Since the thresholds are continuous and strictly monotone, this cutoff is the unique cost satisfying

$$\begin{cases} \underline{\mu}(\overline{\lambda}) = \mu_0, & \text{if } \mu_0 < \mu_{ind}, \\ \overline{\mu}(\overline{\lambda}) = \mu_0, & \text{if } \mu_0 > \mu_{ind}. \end{cases} \tag{26}$$

At $\lambda = 0$, the Learner acquires a fully informative signal. For $\lambda \in (0, \overline{\lambda})$, the prior lies strictly between the two thresholds and she acquires some information. For $\lambda \geqslant \overline{\lambda}$, she acts on the prior, so $\tilde{V}_{\mu_0}(\lambda) = V(\mu_0)$. We can therefore use the cutoff in Equation (26) as the upper bound in the compact problem in Equation (11).

4.3 A Geometric Interpretation for the Obfuscator's Problem

Bayes plausibility also gives a geometric interpretation of the Obfuscator's payoff. Let $m(\lambda)$ be the slope of the chord joining $(\underline{\mu}(\lambda), V(\underline{\mu}(\lambda)))$ and $(\overline{\mu}(\lambda), V(\overline{\mu}(\lambda)))$. For $\lambda \in [0, \overline{\lambda}]$, the height of the chord at μ_0 gives $\tilde{V}_{\mu_0}(\lambda)$ because the weights are exactly the probabilities in Equation (24), with the endpoint cases given by continuity. At full transparency the posteriors are 0 and 1. At full obfuscation the posterior equals the prior, so the payoff is $V(\mu_0)$. Unlike Bayesian Persuasion, the Obfuscator cannot choose the two endpoints independently.

As λ rises, the two thresholds move closer to μ_{ind} and the height of the chord at μ_0 changes. Figure 4 shows a weakly increasing payoff (top row), a weakly decreasing payoff (middle row), and a payoff that first rises and then falls (bottom row).

Using $\mathcal{O} = [0, \overline{\lambda}]$ with the binary cutoff above, we can write the payoff as

$$\tilde{V}_{\mu_0}(\lambda) = \begin{cases} m(\lambda)(\mu_0 - \underline{\mu}(\lambda)) + V(\underline{\mu}(\lambda)) \text{ if } \lambda \in \mathcal{O}, \\ V(\mu_0), \text{ otherwise.} \end{cases} \tag{27}$$

We can also find $\overline{\lambda}$ as the unique cost λ that solves

$$m(\lambda)\Big(\mu_0 - \underline{\mu}(\lambda)\Big) + V(\underline{\mu}(\lambda)) = V(\mu_0).$$

For the chord to have height $V(\mu_0) \in \{0, 1\}$ at the prior, one of its endpoints must be at μ_0. This is exactly the threshold condition in Equation (26).

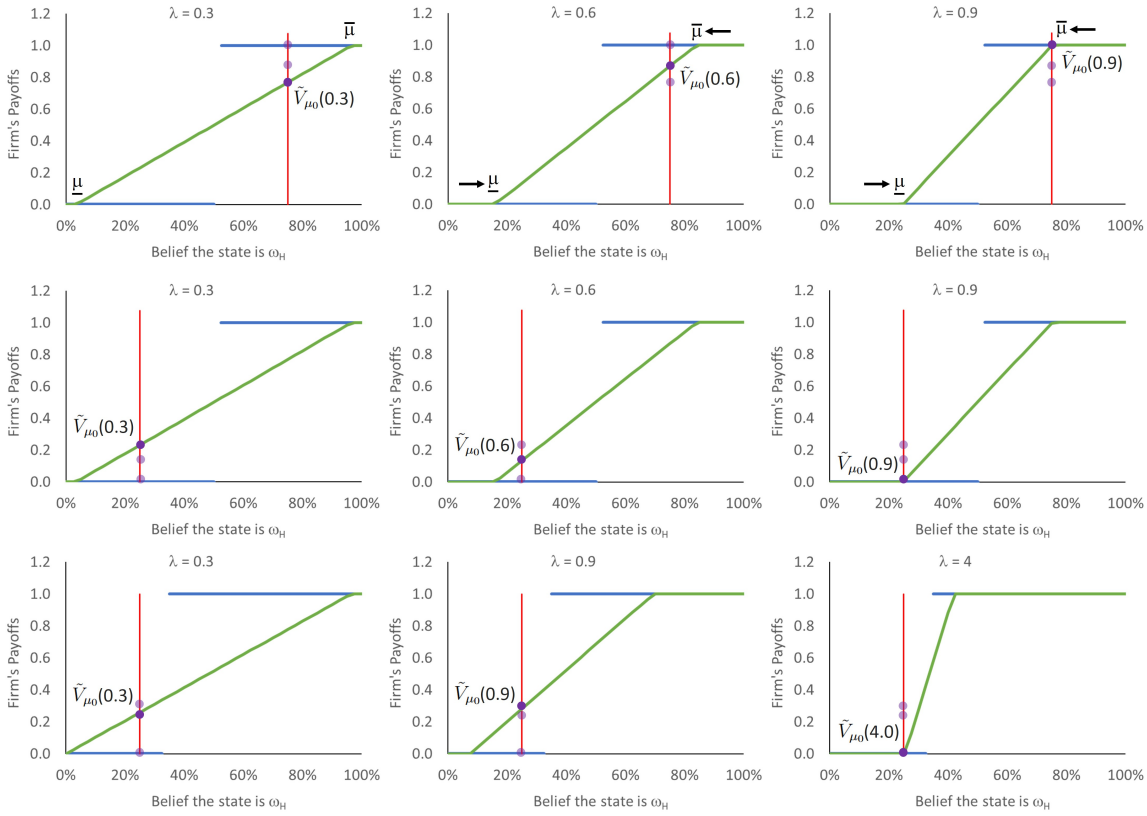


Figure 4: The blue step is V, the green line gives the expected payoff at each prior for the given λ, and the red line marks μ_0. The dark point gives $\tilde{V}_{\mu_0}(\lambda)$. We set $v(a_H) = 1$ and $v(a_L) = 0$ throughout. The top row uses $(\mu_0, B, C) = (0.75, 1, 1)$, the middle row $(0.25, 1, 1)$, and the bottom row $(0.25, 2, 1)$.

5 Obfuscation and Welfare

5.1 Characterization

We now use the binary geometry to express the obfuscation cases from Definition 1 directly in terms of the Learner's payoffs and prior. The general analysis already shows how consideration-set changes can induce partial obfuscation with equal rewards. With two actions, Proposition 4 rules out that case, so we focus on the additional role of asymmetric mistake costs.

In the binary case, the prior and the belief of indifference are enough to identify the obfuscation regime.

Definition 2 (Skeptical Learner)**.** *We say that we have a* Skeptical Learner *if the indifference belief is lower than* $\frac{1}{2}$ *and the initial prior is even lower, i.e.,* $\mu_0 < \mu_{ind} < 1/2$.

A *Skeptical Learner* would not choose the Obfuscator's preferred action a_H absent new information, even though her threshold for choosing it is below 1/2. There is a wide range of beliefs for which the two would agree, but her prior falls below that already-low threshold. The distinction matters: a low prior alone does not imply partial obfuscation. The location of the indifference belief also matters.

Theorem 1 (Obfuscation Characterization)**.** *In the* 2×2 *case,*

(i) we have a Partial Obfuscation solution if and only if we have a Skeptical Learner;

(ii) we have a Full Obfuscation solution if and only if we have Agreement at the Prior, i.e. $a^*(\mu_0) = \arg\max_{a_i \in A} v(a_i)$*;*

(iii) otherwise, we have a Full Transparency solution.

The three cases have a simple interpretation. If $\mu_0 > \mu_{ind}$, the Learner chooses a_H without new information, and full obfuscation preserves the Obfuscator's favorite action. If $\mu_0 < \mu_{ind}$, the Obfuscator needs some learning to induce a_H. For a Skeptical Learner, $B > C$, and the upper threshold initially moves inward faster than the lower one, in the sense that $\underline{\mu}(\lambda)/(1 - \overline{\mu}(\lambda)) \to 0$ as $\lambda \downarrow 0$. Bayes plausibility then puts more weight on the recommendation to buy than under full transparency. At sufficiently high costs, the Learner stops buying altogether. With an unfavorable prior and $\mu_{ind} \geqslant 1/2$, full transparency is optimal. Appendix B.2 establishes these comparisons.

The characterization identifies the regime without requiring us to solve for the optimal cost. To compute that cost under partial obfuscation, we can use a first-order condition.

Proposition 6 (At Most One Interior Stationary Point)**.** *In the* 2×2 *case, there exists at most one* $\lambda^* \in (0, \overline{\lambda})$ *such that* $\frac{d\tilde{V}_{\mu_0}(\lambda^*)}{d\lambda} = 0$.

Under partial obfuscation, an optimum lies in $(0, \overline{\lambda})$ and must satisfy $d\tilde{V}_{\mu_0}/d\lambda = 0$. Proposition 6 allows at most one such point, so the optimal cost is unique. Outside this case, a stationary point must be compared with the boundary payoffs before treating it as an optimum. The Skeptical Learner condition therefore provides a simple way to identify partial obfuscation without differentiating the more complicated payoff function.[26]

[26] Higher-order derivatives can also help distinguish maxima from minima, but are more computationally demanding.

Two benchmarks help put the characterization in context. The first is the binary implication of Proposition 4.

Lemma 3. *If the Learner has Symmetric Preferences (i.e., $\mu_{ind} = \frac{1}{2}$), then we have a boundary obfuscation solution.*

When $B = C$, the two thresholds move symmetrically as λ increases, so the resulting payoff is monotone. The belief of indifference reflects the relative stakes in the two states. Holding C fixed, a higher B lowers μ_{ind}, so the Learner needs less evidence that the state is ω_H before choosing a_H.[27]

The second benchmark uses the geometry in Section 4.3. For this result alone, we allow the Obfuscator to have state-dependent preferences, as in Appendix F.

Proposition 7 (Concavity and Obfuscation)**.** *In the 2×2 case, if V is concave (convex), then $\tilde{V}_{\mu_0}$ is weakly increasing (weakly decreasing) in λ.*

As λ increases, the posteriors move inward while Bayes plausibility keeps their average fixed. Reducing this spread raises the Obfuscator's expected payoff if V is concave and lowers it if V is convex. In the baseline binary model, V is a nonconstant step function, so neither curvature condition applies.

Partial obfuscation requires an interior information cost to outperform both boundary choices, so $\tilde{V}_{\mu_0}$ must be non-monotonic. The key force is that the Obfuscator need not always prefer less learning: full transparency may reveal too much, but full obfuscation may leave the Learner acting on an unfavorable prior. In such cases, an intermediate information cost allows the Learner to acquire some information while still inducing mistakes in the direction the Obfuscator prefers. This mechanism helps interpret practices such as add-on fees that are formally disclosed but placed in fine print, or product comparisons in which relevant attributes are available but surrounded by many additional features. These examples are best understood as ex-ante choices of complexity, format, or ease of comparison. Although partial obfuscation could also result from full obfuscation being prohibitively costly, our baseline

[27] For related evidence on information acquisition, Brown and Jeon (2024) document that when faced with choosing among different types of health insurance, patients with a higher likelihood of needing insurance tend to gather more information about the plans before making a decision.

model shows that such costs are not necessary: partial obfuscation can be optimal even when the Obfuscator can choose any λ at no direct cost.

5.2 Welfare and Mistake Costs

The binary characterization also lets us study how the cost of mistakes affects the Obfuscator's incentive to provide transparency. Using the payoffs from Section 4, let $u_0^* = \mu_0 B$ denote the Learner's ex-ante payoff under full transparency. It is the highest payoff she can achieve.

Consider a policy that makes buying a bad product more costly by lowering $u(a_H, \omega_L)$ from $-C$ to $-C - \varepsilon$ and holding the other three payoffs fixed. In the original matching payoffs, the same policy lowers the payoff from buying a bad product from 0 to $-\varepsilon$ and leaves the payoff from declining it fixed at its original level. Thus, the adjustment used to normalize the payoffs is the same before and after the policy. We evaluate welfare using the normalized material payoffs, net of information costs. The next theorem shows how making a mistake more costly can induce the Obfuscator to provide full transparency.

Theorem 2 (Policy Design for Transparency)**.** *Assume we do not have Full Transparency. Then, there exists $\underline{\varepsilon} > 0$ such that reducing $u(a_H, \omega_L)$ by $\varepsilon > \underline{\varepsilon}$ induces a full information equilibrium (the Obfuscator sets $\lambda^* = 0$) that achieves u_0^*. Moreover,*

(i) under Partial Obfuscation, $\underline{\varepsilon} = B - C$, and the result holds with equality, i.e., for all $\varepsilon \geqslant \underline{\varepsilon}$;

(ii) under Full Obfuscation, $\underline{\varepsilon} = \max\{\frac{\mu_0}{1-\mu_0} B - C, B - C\}$.

It may seem strange that strictly reducing the Learner's payoff from an action in a state can improve the Learner's equilibrium welfare. We are effectively increasing the costs of mistakes. However, the intuition is straightforward: increasing the costs of mistakes ensures that the Learner requires a higher posterior probability that the product is good before she buys, leaving the Obfuscator no choice but to reduce λ. Moreover, the high cost of a mistake has no direct impact on welfare because under the induced full information, the Learner never makes a mistake. Applying Theorem 2 only requires knowing things that are common

knowledge: payoffs and the prior. It does not require experimentation from the policy-maker and still leaves learning in the hands of the Learner.

The return-window example in the introduction illustrates the reverse incentive. Reducing the loss from a bad purchase can give the firm more reason to obfuscate. In evaluating such a policy, we need to account for how the Obfuscator responds as well as how much the policy protects the Learner.

Can a small increase in the cost of a mistake also improve welfare? The next result answers this question for some Skeptical Learners. Recall that a Skeptical Learner is characterized by the condition

$$\mu_0 < \mu_{ind} < \frac{1}{2} \iff \frac{\mu_0}{1-\mu_0} B < C < B. \tag{28}$$

For such a Skeptical Learner, let

$$\begin{aligned} W(\lambda, C) \coloneqq & \max_{\tau} \mathbb{E}_{\mu\sim\tau}[\max\{0, (B+C)\mu - C\}] - \lambda(H(\mu_0) - \mathbb{E}_{\mu\sim\tau}[H(\mu)]), \\ & \text{s.t. } \mathbb{E}_{\mu\sim\tau}[\mu] = \mu_0. \end{aligned}$$

denote the Learner's welfare, net of information costs, when parameters are (λ, C). Correspondingly, for a given equilibrium level of obfuscation $\lambda^*(C)$,[28] we can further define $W^*(C) \coloneqq W(\lambda^*(C), C)$.

A higher C has two effects. Holding the information cost fixed, buying in the bad state becomes more costly. But the Obfuscator responds by lowering $\lambda^*(C)$, which makes learning cheaper. Equation (B.4) in Appendix B.6 separates these two effects.

Proposition 8 (Welfare and Mistake Costs)**.** *Assume we have a Skeptical Learner, so* $C \in (\frac{\mu_0}{1-\mu_0} B, B)$*. Then, there exists* $\eta > 0$ *such that if* $C \in \left(\frac{\mu_0}{1-\mu_0} B,\ \frac{\mu_0}{1-\mu_0} B + \eta\right) \cup (B - \eta, B)$*, then* $\frac{dW^*(C)}{dC} > 0$*.*

Therefore, if the conditions of Proposition 8 hold, then there exists $\delta > 0$ such that the Learner's welfare is increasing in ε (the increase in the cost of mistakes) for all $\varepsilon \in (0, \delta)$. We do not have an analytical proof that applies to all values of C in the Skeptical interval. The result near the two boundaries suggests that the property may hold more generally, but that

[28] Proposition 6 gives uniqueness under partial obfuscation.

remains a conjecture.

Proposition 8 gives additional economic meaning to the definition of a Skeptical Learner and, therefore, to Theorem 1. Under full obfuscation, a sufficiently small increase in the cost of mistakes leaves the prior in the agreement-at-the-prior region, so the Obfuscator continues to fully obfuscate. The Learner then faces a worse payoff when she makes a mistake in the bad state. Under partial obfuscation, when C is sufficiently close to either boundary of the Skeptical interval, even a small increase in mistake costs improves the Learner's equilibrium welfare. The result highlights a strategic channel through which making mistakes more costly induces greater transparency and can leave the Learner better off.

5.3 Comparison to Communication Games

We now compare our framework with communication games, holding payoffs and the common prior fixed. What changes is who controls information. In Bayesian Persuasion (BP) (Kamenica and Gentzkow, 2011), the sender commits to an experiment that determines what the receiver learns about the state. In Cheap Talk (CT) (Crawford and Sobel, 1982), an informed sender chooses a message without commitment. In Strategic Obfuscation (SO), the uninformed Obfuscator commits to an information cost, and the Learner then chooses her own experiment. We use sender-favoring tie-breaking in BP and SO. In CT, the receiver may mix among best responses after a message.

We first consider the binary environment from Section 4, with a unique optimal action at the prior. We allow the Obfuscator's utility $v : A \times \Omega \to \mathbb{R}$ to depend on the state unless stated otherwise. The Obfuscator corresponds to the sender and the Learner to the receiver.[29]

Definition 3 (Sender Payoffs)**.** *Let s^P be the optimal payoff for the Sender under Bayesian Persuasion; let s^C be the Sender's best equilibrium payoff under Cheap Talk; and let s^O be the optimum payoff for the Obfuscator under Strategic Obfuscation.*

With state-independent preferences,[30] the sender's or Obfuscator's payoff depends only on the receiver's or Learner's action. This gives the following comparison.

[29]We assume that the set of messages for Cheap Talk is rich enough to include at least $|A|$ messages.
[30]Also called transparent motives, as in Lipnowski and Ravid (2020).

Proposition 9 (Payoff Ranking)**.** *The following holds:*

(i) $s^O \leqslant s^P$*. Under State-Independent Preferences for the sender/Obfuscator,* $s^O = s^P$ *if, and only if, we have Agreement at the Prior, i.e.,* $a^*(\mu_0) \subseteq \arg\max_{a_i \in A} v(a_i)$*.*

(ii) Under State-Independent Preferences for the sender/Obfuscator, $s^C \leqslant s^O$*.*

Every posterior distribution induced by SO is Bayes-plausible, so BP gives an upper bound on the Obfuscator's payoff. Under agreement at the prior, an uninformative BP signal is optimal, as is any signal that keeps every posterior in the region where the receiver chooses the sender's preferred action. Appendix C proves the payoff equality characterization.

With two states and two actions and state-independent sender preferences, the best CT payoff is the babbling payoff. If the sender strictly prefers one action, every type must face the same probability of that action after each on-path message. An interior probability would require indifference at every such posterior, which would imply indifference at the prior. The receiver therefore takes her prior-optimal action, which SO can also induce by setting a sufficiently high cost.

With more than two actions the comparison with CT can go either way. Some posterior distributions can be supported in CT but cannot arise from optimal learning at a common cost λ. Conversely, the Obfuscator can make information free even when an informed sender would prefer to conceal the state. The next two examples illustrate these possibilities in the broader environment of Appendix F.

Example 3 (Cheap Talk outperforming Strategic Obfuscation). There are two players, player 1 (sender or Obfuscator), and player 2 (receiver or Learner). Player 2 chooses an action from the set $A = \{a, b, c, d\}$, which determines both players' payoffs. The state $\omega \in \{High, Low\}$ is drawn with prior $\mu_0 = \Pr(\omega = High)$. Table 1 shows payoffs in the order (sender, receiver), with states in the rows. We use the same convention in the next example.

Table 1: Payoffs for the players.

		Receiver action			
		a	b	c	d
State	High	1, -2	0, -1	1, 0	0, 1
	Low	1, 1	0, 0.9	1, 0	0, -1

Given the optimal action for Player 2 at each belief, $a^*(\mu) = \arg\max_{x \in A} \mathbb{E}_\mu[u(x, \omega)]$, Figure 5 plots the indirect utility $V(\mu)$ for Player 1, using the sender-favored best response at ties. We fix a prior of 0.4. Under Cheap Talk, we can use the characterization in Lipnowski and Ravid (2020), and there exists an equilibrium in which Player 1 induces the distribution of posteriors τ given by $\tau(0.05) = \frac{2}{9}$ and $\tau(0.5) = \frac{7}{9}$ that yields a payoff of 1. The goal is to induce posteriors that fall in the small areas where the payoff for Player 1 is 1. Under Strategic Obfuscation, however, a payoff of 1 for Player 1 is not possible. For any $\lambda \in \mathbb{R}_+$, the induced distribution of posteriors τ_λ must assign positive probability to recommending action b or d.

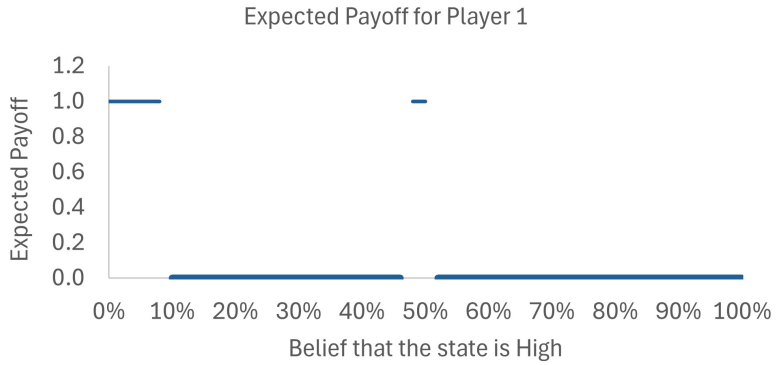


Figure 5: The sender's payoff at each posterior, using the sender-favored best response at ties.

□

Example 4 (Cheap Talk underperforming Strategic Obfuscation). Consider now that Player 2 chooses an action from the set $A = \{a, b, c\}$, and that the payoffs are given by Table 2. The corresponding expected utility for Player 1 is plotted in Figure 6. For a prior of 40%, a payoff of 0.5 is the maximum equilibrium payoff that a sender can get under Cheap Talk (Lipnowski and Ravid, 2020). However, under SO, the Obfuscator can choose $\lambda = 0$, which leads to full information and an expected payoff of 0.7.

At posterior 0.75, the receiver is indifferent between b and c. Mixing equally gives the sender 0.5, as does choosing a at posterior 0.25. Placing weight 0.3 on 0.75 recovers the prior 0.4. A common sender payoff above 0.5 would require every on-path posterior to be at least 0.75, which Bayes plausibility rules out.

Table 2: Payoffs for the players.

		Receiver action		
		a	b	c
State	High	0.5, -1	0, 0.5	1, 1
	Low	0.5, 1	0, 0.5	1, -1

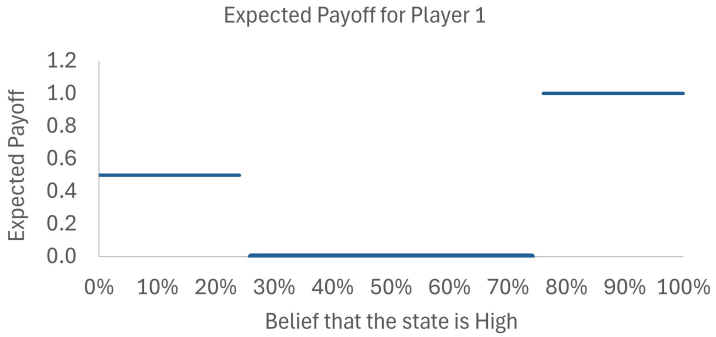


Figure 6: The sender's payoff at each posterior, using the sender-favored best response at ties. Cheap Talk also allows mixtures at the ties.

□

6 Generalizing the Information Cost Function

So far, we have used Shannon entropy throughout the general model and its binary specialization. We now keep the 2×2 setting and ask which results survive when we change the information cost function. We use the class of posterior-separable costs introduced in Caplin et al. (2022), maintain state-independent preferences for the Obfuscator with $v(a_H) > v(a_L)$, and retain the normalization $u(a_H, \omega_H) = B,\ u(a_L, \omega_H) = 0,\ u(a_H, \omega_L) = -C,\ u(a_L, \omega_L) = 0$.

Definition 4 (Posterior-Separable Costs, Caplin et al. (2022))**.** *A posterior-separable cost function $c_\lambda : \triangle(\triangle(\Omega)) \to \mathbb{R}$ with parameter $\lambda \in \mathbb{R}_+$ is weakly uniformly posterior-separable if, for each subset $\overline{\Omega} \subset \Omega$, there exists a strictly convex function $S_{\overline{\Omega}} : \triangle(\overline{\Omega}) \to \overline{\mathbb{R}}$ such that, for all $\mu_0 \in \triangle(\Omega)$ and Bayes-plausible $\tau \in \triangle(\triangle(\Omega))$,*

$$c_\lambda(\mu_0, \tau) = \lambda \left(\sum_{\mu \in \operatorname{supp} \tau} \tau(\mu) S_{\operatorname{supp} \mu_0}(\mu) - S_{\operatorname{supp} \mu_0}(\mu_0) \right). \tag{29}$$

The Shannon benchmark corresponds to the convex generator $S_E(\mu) = -H(\mu)$. Equation (29) then reproduces the entropy-reduction cost in Equation (2). We retain the parameter λ to make the Obfuscator's choice explicit; changing S changes the relative cost of different experiments.[31] The next proposition, for which we omit the formal proof, shows that the link between agreement at the prior and optimal full obfuscation survives this change.

Proposition 10 (Full Obfuscation, Generalized)**.** *In the 2x2 case with a cost function as in Equation* (29)*, we have full obfuscation if, and only if, we have agreement at the prior.*

The intuition is the same as before. If the Learner's prior-optimal action is a_H, full obfuscation gives the Obfuscator his maximal payoff. If instead it is a_L, full obfuscation gives him his lowest payoff, and full transparency is strictly better because the prior is interior. Under agreement at the prior, full obfuscation also attains the Bayesian Persuasion payoff. Thus, the comparison in part (i) of Proposition 9 continues to hold.

We now allow the Learner to be indifferent between the two actions at the prior. In the Shannon model, Proposition 2 shows that she then acquires information at every finite cost. The next proposition extends the binary result to differentiable posterior-separable costs. The proofs of the following two propositions are in the Appendix.

Proposition 11 (Indifference Blocks Obfuscation)**.** *Assume that $c_\lambda(\mu_0, \tau)$ is given as in Equation* (29) *and that $S_{\operatorname{supp}\mu_0}$ is differentiable at μ_{ind}. If $\mu_0 = \mu_{ind}$, then $\delta_{\mu_0} \notin T_{\mu_0}(\lambda)$ for any $\lambda \in \mathbb{R}_+$.*

One may ask how special partial obfuscation is within this broader family of information costs. Are there cost functions other than Entropy that allow for it? The next result gives a positive answer.

Proposition 12 (Robustness to Cost Functions)**.** *Assume that we have a Skeptical Learner, i.e., $\mu_0 < C/(B+C) < 1/2$, and let $J \subset (0,1)$ be a compact interval containing, in its interior, the two posteriors induced by optimal partial obfuscation under Shannon costs. Then there exists a neighborhood $\mathcal{N}$ of the Shannon generator S_E in the $C^2(J)$ topology such that,*

[31] For a fixed $\lambda > 0$, the scale can be absorbed into the generator by replacing S with λS.

for every strictly convex generator $S \in \mathcal{N}$ that is continuous on $[0,1]$, the corresponding cost

$$c_\lambda^S(\mu_0, \tau) = \lambda\Big(\mathbb{E}_\tau[S_{\operatorname{supp}\mu_0}(\mu)] - S_{\operatorname{supp}\mu_0}(\mu_0)\Big),$$

also admits a partial obfuscation solution.

Partial obfuscation therefore does not depend on choosing exactly the Shannon cost function. Small changes in the generator preserve the Obfuscator's strict gain over both full transparency and full obfuscation.

References

Alonso, R. and O. Camara (2024). Organizing data analytics. *Management Science 70*(5), 3123–3143.

Asriyan, V., D. Foarta, and V. Vanasco (2023). The good, the bad, and the complex: Product design with imperfect information. *American Economic Journal: Microeconomics 15*(2), 187–226.

Barros, L. (2026). The limits of information acquisition in cheap talk. Working paper.

Ben-Shahar, O. and C. E. Schneider (2014). *More Than You Wanted to Know*. Princeton University Press.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Bizzotto, J., J. Rüdiger, and A. Vigier (2020). Testing, disclosure and approval. *Journal of Economic Theory 187*, 105002.

Blake, T., S. Moshary, K. Sweeney, and S. Tadelis (2021). Price salience and product choice. *Marketing Science 40*(4), 619–636.

Bloedel, A. W. and I. Segal (2021). Persuading a rationally inattentive agent. Working paper.

Boyacı, T. and Y. Akçay (2018). Pricing when customers have limited attention. *Management Science 64*(7), 2995–3014.

Brown, Z. Y. and J. Jeon (2024). Endogenous information and simplifying insurance choice. *Econometrica 92*(3), 881–911.

Caplin, A., M. Dean, and J. Leahy (2019). Rational inattention, optimal consideration sets, and stochastic choice. *Review of Economic Studies 86*(3), 1061–1094.

Caplin, A., M. Dean, and J. Leahy (2022). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. *Journal of Political Economy 130*(6), 1676–1715.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica 50*(6), 1431–1451.

Cusumano, C. M., F. Fabbri, and F. Pieroth (2024). Competing to commit: Markets with rational inattention. *American Economic Review 114*(1), 285–306.

de Clippel, G. and K. Rozen (2025). Communication, Perception and Strategic Obfuscation. working paper.

De Oliveira, H., T. Denti, M. Mihm, and K. Ozbek (2017). Rationally inattentive preferences and hidden information costs. *Theoretical Economics 12*(2), 621–654.

Dean, M. and N. Neligh (2023). Experimental tests of rational inattention. *Journal of Political Economy 131*(12), 3415–3461.

Edmond, C. and Y. K. Lu (2021). Creating confusion. *Journal of Economic Theory 191*, 105145.

Ellis, A. (2018). Foundations for optimal inattention. *Journal of Economic Theory 173*, 56–94.

Ellison, G. (2005). A Model of Add-On Pricing. *The Quarterly Journal of Economics 120*(2), 585–637.

Ellison, G. and A. Wolitzky (2012). A search cost model of obfuscation. *RAND Journal of Economics 43*(3), 417–441.

Fabbri, F. (2024). Attention holdup. Working paper.

Fréchette, G. R., A. Lizzeri, and J. Perego (2022). Rules and commitment in communication: An experimental analysis. *Econometrica 90*(5), 2283–2318.

Gabaix, X. and D. Laibson (2006). Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets. *The Quarterly Journal of Economics 121*(2), 505–540.

Gentzkow, M. and E. Kamenica (2014). Costly Persuasion. *American Economic Review 104*(5), 457–462.

Janssen, A. and J. Kasinger (2024). Obfuscation and rational inattention. *The Journal of*

Industrial Economics 72(1), 390–428.

Jin, G. Z., M. Luca, and D. Martin (2022). Complex disclosure. *Management Science 68*(5), 3236–3261.

Kalayci, K. and J. Potters (2011). Buyer confusion and market prices. *International Journal of Industrial Organization 29*(1), 14–22.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics 11*, 249–272.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Le Treust, M. and T. Tomala (2019). Persuasion with limited communication capacity. *Journal of Economic Theory 184*, 104940.

Li, H. and X. Shi (2017). Discriminatory information disclosure. *American Economic Review 107*(11), 3363–85.

Lin, X. and C. Liu (2024). Credible persuasion. *Journal of Political Economy 132*(7), 2228–2273.

Lipnowski, E., L. Mathevet, and D. Wei (2020). Attention management. *American Economic Review: Insights 2*(1), 17–32.

Lipnowski, E. and D. Ravid (2020). Cheap talk with transparent motives. *Econometrica 88*(4), 1631–1660.

Lipnowski, E., D. Ravid, and D. Shishkin (2022). Persuasion via weak institutions. *Journal of Political Economy 130*(10), 2705–2730.

Martin, D. (2017). Strategic pricing with rational inattention to quality. *Games and Economic Behavior 104*, 131–145.

Matějka, F. (2016). Rationally inattentive seller: Sales and discrete pricing. *Review of Economic Studies 83*(3), 1125–1155.

Matějka, F. and A. McKay (2012). Simple market equilibria with rationally inattentive consumers. *American Economic Review 102*(1), 24–29.

Matějka, F. and A. McKay (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review 105*(1), 272–298.

Matysková, L. and A. Montes (2023). Bayesian persuasion with costly information acquisition.

Journal of Economic Theory 211, 105678.

Min, D. (2021). Bayesian persuasion under partial commitment. *Economic Theory 72*, 743–764.

Nguyen, A. and T. Y. Tan (2021). Bayesian persuasion with costly messages. *Journal of Economic Theory 193*, 105212.

Nocke, V. and P. Rey (2024). Consumer search, steering, and choice overload. *Journal of Political Economy 132*(5), 1684–1739.

Perez-Richet, E. and V. Skreta (2022). Test design under falsification. *Econometrica 90*(3), 1109–1142.

Ravid, D. (2020). Ultimatum bargaining with rational inattention. *American Economic Review 110*(9), 2948–63.

Richards, T. J., G. J. Klein, C. Bonnet, and Z. Bouamra-Mechemache (2020). Strategic obfuscation and retail pricing. *Review of Industrial Organization 57*, 859–889.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27*(3), 379–423.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics 50*(3), 665–690.

Wei, D. (2021). Persuasion under costly learning. *Journal of Mathematical Economics 94*, 102451.

Appendix

We collect the proofs in the order of the main text, placing supporting lemmas before the arguments that use them. The model extensions follow the proofs. When a proof applies to the broader payoff environment in Appendix F, we write $N = |A|$ and $M = |\Omega|$. The matching model is the case $M = N$. In those proofs, $v(a, \omega)$ also allows state-dependent Obfuscator payoffs, with $v(a, \omega) = v(a)$ in the main model.

A Proofs for the General Matching Model

The proofs in this section cover the Learner's solution and equilibrium existence in Subsections 2.3 and 2.4, as well as the obfuscation characterization in Section 3. Throughout this section, write $p_j = \mu_0(\omega_j)$ and $v_i = v(a_i)$. We identify a candidate action set with its matching index set when writing $i \in K$. We also abbreviate $P_i(\lambda) = P_\lambda(a_i)$ and $P_i^K(\lambda) = P_\lambda^K(a_i)$. Vectors $P(\lambda)$ and $P^K(\lambda)$ collect these components.

A.1 Optimal Action Probabilities

Proof of Lemma 2. Fix $\lambda > 0$. Let $\Delta_N = \{x \in \mathbb{R}_+^N : \sum_i x_i = 1\}$. We first show that the Learner's optimal net payoff is

$$\lambda \max_{x \in \Delta_N} F_\lambda(x),\ F_\lambda(x) = \sum_j p_j \ln\left(1 + \frac{x_j}{b_j(\lambda)}\right). \tag{A.1}$$

Let Q_{ij} be the probability of recommending and taking action a_i in state ω_j. For each j, $\sum_i Q_{ij} = 1$. Its unconditional action probabilities are $q_i = \sum_j p_j Q_{ij}$. The mutual information is

$$I(Q) = \sum_j p_j \sum_i Q_{ij} \ln \frac{Q_{ij}}{q_i},$$

with the usual zero-probability conventions. This equals $H(\mu_0) - \mathbb{E}_\tau H(\mu)$. The Learner's net payoff is

$$J_\lambda(Q) = \sum_{i,j} p_j Q_{ij} u(a_i, \omega_j) - \lambda I(Q). \tag{A.2}$$

Working with such direct recommendations is without loss. Choose an optimal action after each original signal, then merge signals that induce the same action. Material payoffs are preserved and information costs weakly fall by concavity of entropy. Conversely, an experiment generated by a direct channel is available in the original problem. A maximizing channel must recommend optimal actions at its induced posteriors; otherwise, changing the action after that recommendation would improve the payoff.

For $x \in \Delta_N$, set

$$Z_j(x) = \sum_i x_i e^{u(a_i,\omega_j)/\lambda} = 1 + \frac{x_j}{b_j(\lambda)}.$$

Write $D_{\mathrm{KL}}(z\|x) = \sum_i z_i \ln(z_i/x_i)$ for relative entropy, with the usual zero-probability conventions. For any probability vector z supported on the support of x, the Gibbs inequality gives

$$\sum_i z_i u(a_i,\omega_j) - \lambda D_{\mathrm{KL}}(z\|x) \leqslant \lambda \ln Z_j(x), \tag{A.3}$$

with equality exactly when $z_i = g_i = x_i e^{u(a_i,\omega_j)/\lambda}/Z_j(x)$. The difference between the right and left sides is exactly $\lambda D_{\mathrm{KL}}(z\|g) \geqslant 0$.

Given any channel Q with marginal q, apply (A.3) with $x = q$ in every state and average using p_j. It follows that

$$J_\lambda(Q) \leqslant \lambda F_\lambda(q) \leqslant \lambda \max_{x\in\Delta_N} F_\lambda(x).$$

The function F_λ is continuous on the compact simplex, so choose a maximizer x^*. Define a channel by

$$Q^*_{ij} = \frac{x^*_i e^{u(a_i,\omega_j)/\lambda}}{Z_j(x^*)}, \tag{A.4}$$

and let q^* be its actual marginal. The entropy decomposition

$$\sum_j p_j D_{\mathrm{KL}}(Q^*_{\cdot j}\|x^*) = I(Q^*) + D_{\mathrm{KL}}(q^*\|x^*)$$

and equality in (A.3) imply

$$J_\lambda(Q^*) = \lambda F_\lambda(x^*) + \lambda D_{\mathrm{KL}}(q^*\|x^*).$$

The upper bound already established forces equality and $q^* = x^*$. Thus the upper bound is

attainable, and the optimizing weights really are unconditional action probabilities.

Strict concavity follows from

$$\frac{\partial F_\lambda}{\partial x_i} = \frac{p_i}{b_i + x_i}, \ \frac{\partial^2 F_\lambda}{\partial x_i^2} = -\frac{p_i}{(b_i + x_i)^2} < 0,$$

with all cross partials zero. Hence x^* is unique. The same upper-bound argument shows that the marginal of any optimal channel must maximize F_λ, so all optimal experiments induce the same action probabilities.

The simplex contains strictly positive feasible vectors. The KKT conditions are therefore necessary, and concavity makes them sufficient. With ν the multiplier on $\sum_i x_i = 1$, they are

$$\frac{p_i}{b_i + x_i} \begin{cases} = \nu, & x_i > 0, \\ \leqslant \nu, & x_i = 0. \end{cases} \tag{A.5}$$

At least one action is used, so $\nu > 0$. Rearranging yields (9). The sum of its right-hand sides is continuous in $\nu > 0$, tends to infinity as $\nu \downarrow 0$, and equals zero for sufficiently large ν. It is strictly decreasing wherever it is positive. Hence there is exactly one multiplier at which the sum equals one. □

Posterior consistency. For each used action, the channel in (A.4) yields

$$\mu^{a_i}(\omega_j) = \frac{p_j e^{\alpha_j \mathbb{1}\{i=j\}/\lambda}}{1 + P_j(\lambda)/b_j(\lambda)}. \tag{A.6}$$

These probabilities sum to one because the channel has marginal $P_i(\lambda)$. Bayes plausibility follows from the channel construction. Thus the optimal weights correspond to feasible, Bayes-plausible experiments.

A.2 No Learning and Its Uniqueness

The following lemma describes the zero-cost limit of the no-learning regions. Write $I(0; a_i) \coloneqq \lim_{\lambda \to 0^+} I(\lambda; a_i)$, and let $\Omega(a_i) = \{\omega \in \Omega : u(a_i, \omega) \geqslant u(a_j, \omega) \text{ for every } j \neq i\}$ be the set of states in which a_i is optimal. Under matching payoffs, $\Omega(a_i) = \{\omega_i\}$, so the limiting region consists of the belief assigning probability 1 to that state.

Lemma 4. *We have that* $\lim_{\lambda\to 0^+} I(\lambda; a_i) = \{\mu \in \triangle(\Omega) : \text{supp}(\mu) \subseteq \Omega(a_i)\}$.

Proof of Lemma 4. The sets $I(\lambda; a_i)$ are closed and shrink as $\lambda \downarrow 0$, so their limit is their intersection over $\lambda > 0$. If $\text{supp}(\mu) \subseteq \Omega(a_i)$, then $u(a_j, \omega) - u(a_i, \omega) \leqslant 0$ at every state with positive probability. Thus, for every a_j and $\lambda > 0$,

$$\sum_{\omega\in\Omega} \mu(\omega) \exp\left(\frac{u(a_j, \omega) - u(a_i, \omega)}{\lambda}\right) \leqslant 1.$$

Hence $\mu \in I(\lambda; a_i)$ for every positive cost.

Conversely, if $\text{supp}(\mu) \nsubseteq \Omega(a_i)$, there are ω_m and a_j such that $\mu(\omega_m) > 0$ and $u(a_j, \omega_m) > u(a_i, \omega_m)$. Then

$$\sum_{\omega\in\Omega} \mu(\omega) \exp\left(\frac{u(a_j, \omega) - u(a_i, \omega)}{\lambda}\right) \geqslant \mu(\omega_m) \exp\left(\frac{u(a_j, \omega_m) - u(a_i, \omega_m)}{\lambda}\right) \longrightarrow \infty$$

as $\lambda \downarrow 0$. The defining inequality for $I(\lambda; a_i)$ therefore fails at sufficiently small costs, which proves the converse. □

The next results apply to the finite-state environment, including the broader payoff specification in Appendix F. For $\lambda = 0$, we use $I(0; a_i) = \lim_{\lambda\to 0^+} I(\lambda; a_i)$, characterized in Lemma 4.

Lemma 5. *Fix* $a_i \in A$ *and* $\lambda \geqslant 0$. *Then,* $\mu \in I(\lambda; a_i) \implies a_i \in a^*(\mu)$.

Proof of Lemma 5. For $\lambda = 0$, the result follows from the characterization of $I(0; a_i)$ in Lemma 4. Let $\lambda > 0$ and suppose $a_i \notin a^*(\mu)$. There is an action a_j such that $\mathbb{E}_\mu[u(a_j, \omega) - u(a_i, \omega)] > 0$. By Jensen's inequality,

$$\mathbb{E}_\mu\left[\exp\left(\frac{u(a_j, \omega) - u(a_i, \omega)}{\lambda}\right)\right] \geqslant \exp\left(\frac{\mathbb{E}_\mu[u(a_j, \omega) - u(a_i, \omega)]}{\lambda}\right) > 1.$$

Thus $\mu \notin I(\lambda; a_i)$, proving the contrapositive. □

Lemma 6. *For any prior* μ_0 *and* $\lambda \geqslant 0$, *the set of solutions for the Learner's problem is convex.*

Proof of Lemma 6. Fix μ_0 and $\lambda \geqslant 0$. Let $\tau^1, \tau^2 \in T_{\mu_0}(\lambda)$ and $t \in [0, 1]$. The mixture

$\tau^t = t\tau^1 + (1-t)\tau^2$ satisfies Bayes plausibility because

$$\mathbb{E}_{\tau^t}[\mu] = t\mathbb{E}_{\tau^1}[\mu] + (1-t)\mathbb{E}_{\tau^2}[\mu] = \mu_0.$$

The Learner's objective,

$$L_\lambda(\tau) = \mathbb{E}_\tau[U(\mu)] + \lambda\mathbb{E}_\tau[H(\mu)] - \lambda H(\mu_0),$$

is affine in τ. Therefore $L_\lambda(\tau^t) = tL_\lambda(\tau^1) + (1-t)L_\lambda(\tau^2)$ equals the optimal value, so τ^t is also a solution. For $t \in (0,1)$ its support is $\text{supp}(\tau^1) \cup \text{supp}(\tau^2)$. □

Corollary 2. *Fix $\lambda > 0$ and let $\tau^1, \tau^2 \in T_{\mu_0}(\lambda)$. Then Equation (6) holds for all $a_i, a_j \in A(\tau^1) \cup A(\tau^2)$.*

Proof of Corollary 2. Fix $\alpha \in (0,1)$. By Lemma 6, τ^3 defined by $\tau^3(\mu) = \alpha\tau^1(\mu) + (1-\alpha)\tau^2(\mu)$ for all μ is also a solution. Necessity implies that Equation (6) holds for any $a_i, a_j \in A(\tau^3) = A(\tau^1) \cup A(\tau^2)$. □

Proof of Proposition 1. Fix $\lambda > 0$. Applying Lemma 1 to an experiment that recommends only a_i gives exactly the inequalities defining $I(\lambda; a_i)$. Bayes plausibility holds because the only posterior is μ_0, and Equation (6) is vacuous. This proves the first statement.

Now suppose $\mu_0 \in int(I(\lambda; a_i))$. For every $j \neq i$, the corresponding inequality is strict:

$$\sum_{\omega \in \Omega} \mu_0(\omega) \exp\left(\frac{u(a_j, \omega) - u(a_i, \omega)}{\lambda}\right) < 1.$$

Indeed, each constraint is linear in beliefs and cannot hold with equality on an interior neighborhood unless the two actions have identical payoff vectors, which our assumptions rule out.

Suppose there were another optimal distribution $\tau \neq \delta_{\mu_0}$. It must recommend an action $a_j \neq a_i$ with positive probability: an informative experiment followed by the same action everywhere would incur a positive information cost with no material benefit. By Corollary 2,

$$\mu^{a_j}(\omega) = \mu_0(\omega) \exp\left(\frac{u(a_j, \omega) - u(a_i, \omega)}{\lambda}\right) \text{ for every } \omega \in \Omega. \tag{A.7}$$

Summing over states and using the strict inequality above, the putative posterior μ^{a_j} would sum to less than one. Hence no other optimal distribution exists. □

A.3 The Limits of Full Obfuscation

Proof of Proposition 2. Fix $\mu \in int(\triangle(\Omega))$. First suppose $a^*(\mu) = \{a_i\}$. For each $j \neq i$, write $d_j(\omega) = u(a_j, \omega) - u(a_i, \omega)$, so $\mathbb{E}_\mu[d_j(\omega)] < 0$. Since the state space is finite, a Taylor expansion at $t = 0$ gives

$$\mathbb{E}_\mu[e^{td_j(\omega)}] = 1 + t\mathbb{E}_\mu[d_j(\omega)] + O(t^2) < 1$$

for all sufficiently small $t > 0$. There are finitely many actions, so choose a common t for which all these inequalities are strict and let $\overline{\lambda} = 1/t$. By continuity in beliefs, the inequalities remain satisfied on a neighborhood of μ. Thus $\mu \in int(I(\overline{\lambda}; a_i))$, proving (i) implies (ii).

For the converse, suppose that $a^*(\mu)$ contains more than one action. For any $a_i \in a^*(\mu)$, choose another $a_j \in a^*(\mu)$. Then $\mathbb{E}_\mu[d_j(\omega)] = 0$. Because the prior has full support and the two payoff vectors differ, d_j is not constant under μ. Strict Jensen's inequality therefore gives, for every $\lambda > 0$,

$$\mathbb{E}_\mu[e^{d_j(\omega)/\lambda}] > e^{\mathbb{E}_\mu[d_j(\omega)]/\lambda} = 1.$$

Hence $\mu \notin I(\lambda; a_i)$ for any prior-optimal action. Lemma 5 rules out membership for the remaining actions as well. In particular, (ii) cannot hold. □

A.4 Existence of a Solution

Proof of Proposition 3. Write $a^*(\mu_0) = \{a_k\}$. By Proposition 2, a sufficiently large finite $\overline{\lambda}$ places μ_0 in $int(I(\overline{\lambda}; a_k))$. Proposition 1 and the expansion of no-learning regions then imply that every $\lambda \geqslant \overline{\lambda}$ induces no learning and action a_k. The Obfuscator's payoff is constant on this range, so we can restrict his choice to $[0, \overline{\lambda}]$.

For this proof, we will use the stochastic-choice representation of the Learner's problem instead of working with the distribution of posteriors.[32] Remember that $\pi(a|\omega)$ represents the probability that a Learner chooses action a given that the state is ω. Let

$$\Pi \coloneqq \{\pi : A \times \Omega \to [0, 1] : \sum_{a \in A} \pi(a|\omega) = 1\}.$$

[32]For a review, check Kamenica and Gentzkow (2011) and Matějka and McKay (2015).

Since A and Ω are finite, Π is compact. For $\pi \in \Pi$, let $P_\pi(a) = \sum_{\omega\in\Omega} \mu_0(\omega)\pi(a|\omega)$ be the unconditional action probabilities. Using H for entropy on any finite probability vector, the mutual information is $c(\pi) = H(P_\pi) - \sum_{\omega\in\Omega} \mu_0(\omega)H(\pi(\cdot|\omega))$. The Learner pays $\lambda c(\pi)$. Since entropy is continuous, $c : \Pi \to \mathbb{R}_+$ is continuous.

Now, we rewrite the Learner's objective function $f_{\mu_0} : \mathbb{R} \times \Pi \to \mathbb{R}$ as

$$f_{\mu_0}(\lambda, \pi) = \sum_{\omega\in\Omega} \sum_{a\in A} \mu_0(\omega)\pi(a \mid \omega)u(a, \omega) - \lambda c(\pi),$$

which is continuous in both λ and π. Therefore, by the maximum theorem, the solution correspondence

$$\mathcal{T}_{\mu_0}(\lambda) = \arg\max_{\pi\in\Pi} f_{\mu_0}(\lambda, \pi)$$

is nonempty, compact-valued, and upper hemicontinuous. The Obfuscator's expected payoff from a signal π is

$$g_{\mu_0}(\pi) = \sum_{\omega\in\Omega} \sum_{a\in A} \mu_0(\omega)\pi(a \mid \omega)v(a, \omega),$$

and as ties are broken in favor of the Obfuscator,

$$\tilde{V}_{\mu_0}(\lambda) = \max_{\pi\in\mathcal{T}_{\mu_0}(\lambda)} g_{\mu_0}(\pi).$$

Since the solution correspondence is nonempty and compact-valued and g_{μ_0} is continuous, $\tilde{V}_{\mu_0}$ is well defined. To show upper semicontinuity, take any $\lambda_n \to \lambda$. First select a subsequence along which $\tilde{V}_{\mu_0}(\lambda_n)$ converges to its limit superior. For each term, choose a maximizing experiment $\pi_n \in \mathcal{T}_{\mu_0}(\lambda_n)$. Compactness gives a further subsequence, indexed by n_k, with $\pi_{n_k} \to \pi^*$. Upper hemicontinuity implies $\pi^* \in \mathcal{T}_{\mu_0}(\lambda)$. Hence,

$$\limsup_{n\to\infty} \tilde{V}_{\mu_0}(\lambda_n) = \lim_{k\to\infty} g_{\mu_0}(\pi_{n_k}) = g_{\mu_0}(\pi^*) \leqslant \tilde{V}_{\mu_0}(\lambda).$$

Thus, $\tilde{V}_{\mu_0}$ is upper semicontinuous and attains a maximum on the compact interval $[0, \overline{\lambda}]$. $\square$

A.5 Equal Matching Rewards

Maintain the assumptions of Proposition 4. We use $m_k = \sum_{i=1}^{k} p_i$, $p_{N+1} = 0$, and $b(\lambda) = (e^{\Delta/\lambda} - 1)^{-1}$. The following lemma records the consideration-set boundaries used in the proof.

Define, for $1 \leqslant k < N$,

$$\beta_k = \frac{p_{k+1}}{m_k - k p_{k+1}}, \ \lambda_k = \frac{\Delta}{\ln(1 + 1/\beta_k)}, \tag{A.8}$$

and let $\beta_N = \lambda_N = 0$. The denominator is positive under (12).

Lemma 7 (Consideration sets and their boundaries)**.** *The cutoffs satisfy*

$$0 = \beta_N < \beta_{N-1} < \cdots < \beta_1, \ 0 = \lambda_N < \lambda_{N-1} < \cdots < \lambda_1 \leqslant \overline{\lambda}.$$

On the closed branch $b \in [\beta_k, \beta_{k-1}]$, for $2 \leqslant k \leqslant N$, the optimal probabilities are

$$P_i = \begin{cases} \dfrac{p_i}{m_k}(1 + kb) - b, & i \leqslant k, \\ 0, & i > k. \end{cases} \tag{A.9}$$

They have positive support $\{1, \ldots, k\}$ below the upper endpoint; action k has zero weight at that upper endpoint. At $b \geqslant \beta_1$, only action 1 is chosen. At $b = \beta_k$, the distribution is

$$P_i^{(k)} = \begin{cases} \dfrac{p_i - p_{k+1}}{m_k - k p_{k+1}}, & i \leqslant k, \\ 0, & i > k. \end{cases} \tag{A.10}$$

For $k = N$, this formula is the limiting transparency distribution $P^{(N)} = p$.

Proof. By (9), an action is used exactly when $p_i > \nu b$. Strict ordering implies that each positive support consists of the first k actions. Normalization gives $\nu = m_k/(1 + kb)$ and hence (A.9).

For that branch to satisfy the weak KKT conditions, the smallest included weight must be nonnegative and the largest excluded marginal value must be no greater than ν. For $k \geqslant 2$, the first condition is

$$p_k(1 + kb) \geqslant m_k b \Longleftrightarrow b \leqslant \frac{p_k}{m_k - k p_k} = \beta_{k-1}.$$

For $k < N$, the second is

$$p_{k+1}(1 + kb) \leqslant m_k b \iff b \geqslant \beta_k.$$

There is no excluded-action constraint for $k = N$, and the branch extends to $b = 0$ by continuity. For $k = 1$, the included weight is always one and the outside condition is $b \geqslant \beta_1$.

For $2 \leqslant k \leqslant N - 1$, use $m_k - kp_k = m_{k-1} - (k-1)p_k > 0$. The function $t \mapsto t/(m_k - kt)$ is strictly increasing on the relevant interval. Since $p_k > p_{k+1}$, it follows that $\beta_{k-1} > \beta_k$. Also $\beta_{N-1} > 0 = \beta_N$. The map $b(\lambda)$ is strictly increasing, with inverse $\Delta/\ln(1 + 1/b)$, so the cost cutoffs inherit this order. The first no-learning cost is $\lambda_1 = \Delta/\ln(p_1/p_2)$. Thus the sufficient upper bound satisfies $\overline{\lambda} \geqslant \lambda_1$.

At $b = \beta_k$ for $k < N$, the excluded-action constraint for a_{k+1} binds, so $\nu b = p_{k+1}$ and $P_{k+1} = 0$. Equation (9) then gives $P_i = (p_i - p_{k+1})/\nu$ for $i \leqslant k$. These are the only positive weights, so summing them to one gives $\nu = m_k - kp_{k+1}$ and hence (A.10). At a common branch boundary, both formulas satisfy KKT and therefore give the same unique distribution. Taking $b \downarrow 0$ in the full-support formula gives $P_i \to p_i$, which coincides with free learning at $\lambda = 0$. □

Proof of Proposition 4. On a branch with the first k actions, let $\bar{v}_k = \sum_{i=1}^{k} p_i v_i / m_k$. Substitution of (A.9) gives

$$\tilde{V}_{\mu_0}(\lambda) = \bar{v}_k + b(\lambda)\left(k\bar{v}_k - \sum_{i=1}^{k} v_i\right). \tag{A.11}$$

This is affine in b, so a maximum on each closed branch is attained at a branch endpoint. By Lemma 7, these branches cover $[0, \lambda_1]$. In particular, at the cost λ_k where a_{k+1} has just left consideration, Equation (4) gives

$$\tilde{V}_{\mu_0}(\lambda_k) = \sum_{i=1}^{k} P_i^{(k)} v_i = R_k, \qquad k < N.$$

Full transparency gives R_N. All costs at or above λ_1, including $\overline{\lambda}$, give R_1. This proves (14) and both endpoint statements.

Neither endpoint is optimal exactly when the maximum strictly exceeds both R_1 and R_N. A finite list has such a maximum exactly when an intermediate R_k exceeds both endpoints. Each intermediate λ_k is strictly between zero and λ_1, so it really induces partial learning.

This proves (15). □

Posterior interpretation of R_k. The candidate weights also have a posterior interpretation. When the first ℓ actions are considered, write $r = e^{\Delta/\lambda}$ and $q = \nu b = m_\ell/(r + \ell - 1)$. Formula (A.6) gives a baseline q in each considered state and a boost to rq in the state matching the recommendation. Excluded states retain their prior probabilities after every used recommendation.

Just before a_{k+1} leaves consideration, Bayes plausibility gives $p_{k+1} = q + (r-1)qP_{k+1}$. As that recommendation disappears, P_{k+1} goes to zero, so the baseline converges to p_{k+1}. At the cutoff, each remaining action is therefore chosen in proportion to how far the prior of its matching state lies above this baseline. Changing the common reward Δ changes the cost at which the action leaves consideration while preserving the choice probabilities at that cost.

A.6 Unequal Matching Rewards

Return to the maintained matching model with possibly unequal rewards. For a candidate K, abbreviate $B_K(\lambda) = \sum_{i\in K} b_i(\lambda)$ and $\nu_K(\lambda) = m_K/(1 + B_K(\lambda))$. The first step is to verify exactly when the candidate weights solve the Learner's problem.

Lemma 8 (Which candidate sets describe optimal learning?)**.** *For $\lambda > 0$, $P^K(\lambda)$ equals the Learner's optimal action distribution if the following inequalities hold:*

$$E_{Ki}(\lambda) \geqslant 0 \ (i \in K),\ E_{Kj}(\lambda) \leqslant 0 \ (j \notin K). \tag{A.12}$$

Conversely, every optimal action distribution is represented by a candidate satisfying these inequalities, by taking K to be its positive support. For a candidate satisfying (A.12)*,*

$$G_K(\lambda) = \tilde{V}_{\mu_0}(\lambda) - \overline{V}. \tag{A.13}$$

Proof. For $i \in K$, the formula for P_i^K is $p_i/\nu_K - b_i$. Also,

$$\sum_{i\in K} E_{Ki} = m_K(1 + B_K) - m_K B_K = m_K.$$

The inside inequalities therefore give nonnegative weights that sum to one. If an inside weight

is positive, its marginal value is ν_K. If an inside weight is zero, $p_i/b_i = \nu_K$, so its KKT inequality also holds. For an excluded action, the outside inequality is equivalent to

$$p_j(1 + B_K) \leqslant m_K b_j \iff p_j/b_j \leqslant \nu_K.$$

Thus all KKT conditions hold, and Lemma 2 gives optimality and uniqueness of the weights.

Conversely, take the actual positive support of an optimum. Summing $P_i = p_i/\nu - b_i$ over this support gives $\nu = \nu_K$. The inside inequalities are then strict, and the outside ones follow from the KKT conditions. Substituting the optimal weights into (19) proves (A.13). □

Proof of Proposition 5. Proposition 3 gives an attained maximum

$$M = \max_{\lambda \in [0,\overline{\lambda}]} \tilde{V}_{\mu_0}(\lambda) \geqslant \overline{V}.$$

The two endpoint payoffs are those in Equation (16). If the problem has a partial obfuscation solution, neither endpoint is optimal, so $M > \overline{V}$. Take an optimal cost and its actual positive support K. The optimal cost lies in $(0, \overline{\lambda})$, and Lemma 8 gives the E inequalities and $G_K = M - \overline{V} > 0$.

Conversely, suppose that a cost and candidate set satisfy Equation (20). Lemma 8 implies that the weights are optimal for the Learner and give $\tilde{V}_{\mu_0}(\lambda) = \overline{V} + G_K(\lambda) > \overline{V}$. Therefore, $M > \overline{V}$ and neither endpoint is optimal. This is a partial obfuscation solution under Definition 1. The candidate witnessing the strict gain need not itself maximize the payoff; the attained maximum is at least as large and must also rule out both endpoints.

The argument works with any sufficient upper bound $\overline{\lambda}$ from Section 2.4.3. Every cost that induces no learning gives $v(a_{i_0}) \leqslant \overline{V}$, so no such cost can satisfy the strict gain condition or maximize the payoff in a partial-obfuscation case. □

B Proofs for the 2×2 Special Case

B.1 Payoff Normalization

Lemma 9 (Normalization)**.** *In the 2×2 case with $v(a_H) > v(a_L)$, the Learner's problem and the Obfuscator's obfuscation choice are unchanged when we normalize payoffs to*

$$u'(a_H, \omega_H) = B, \qquad u'(a_L, \omega_H) = 0,$$
$$u'(a_H, \omega_L) = -C, \qquad u'(a_L, \omega_L) = 0,$$

where $B = u(a_H, \omega_H) - u(a_L, \omega_H)$ and $C = u(a_L, \omega_L) - u(a_H, \omega_L)$. The Learner's ex-ante welfare changes by a constant independent of λ. This constant is also independent of a policy change ε whenever the subtracted state-dependent payoffs are held fixed.

Proof of Lemma 9. Subtract $u(a_L, \omega_H)$ from both action payoffs in state ω_H and $u(a_L, \omega_L)$ from both action payoffs in state ω_L. By Bayes plausibility, every experiment's expected payoff falls by the same amount, $\mu_0 u(a_L, \omega_H) + (1 - \mu_0)u(a_L, \omega_L)$. Information costs do not change, so optimal learning and the Obfuscator's Problem are unchanged. Welfare comparisons across λ are preserved. Comparisons across a policy change ε are also preserved if these subtracted payoffs remain fixed, as they do in the policy exercise in Section 5. □

B.2 Obfuscation Characterization

Proof of Theorem 1. For the proof of part (i), we start by showing sufficiency. Our goal is to find a λ that yields a higher payoff for the Obfuscator than either $\lambda = 0$ or $\lambda = \overline{\lambda}$. Again, we assume $v(a_H) = V(1) > V(0) = v(a_L)$, as the other case follows a symmetric argument. Therefore, the goal of the Obfuscator is to maximize the probability that the Learner chooses a_H, which is given by

$$p_H(\lambda) = \frac{\mu_0 - \underline{\mu}(\lambda)}{\overline{\mu}(\lambda) - \underline{\mu}(\lambda)},$$

through Bayes plausibility. If we have a Skeptical Learner, then $\mu_0 < \mu_{ind} < \frac{1}{2}$, and the Learner would choose a_L if she acts according to the prior. This means that $\underline{\mu}(\overline{\lambda}) = \mu_0$ and $p_H(\overline{\lambda}) = 0$. We also have that $p_H(0) = \mu_0 > 0$, so we want to show that there exists $\lambda \in (0, \overline{\lambda})$

such that $p_H(\lambda) > \mu_0$. Using the expression for $p_H(\lambda)$, we have

$$\frac{\mu_0 - \underline{\mu}(\lambda)}{\overline{\mu}(\lambda) - \underline{\mu}(\lambda)} > \mu_0 \iff \underline{\mu}(\lambda) < \frac{\mu_0}{1-\mu_0}\Big(1 - \overline{\mu}(\lambda)\Big). \tag{B.1}$$

Define $q \coloneqq \frac{\mu_0}{1-\mu_0}$. Thus, it is enough to find $\lambda \in (0, \overline{\lambda})$ such that $\frac{\underline{\mu}(\lambda)}{1-\overline{\mu}(\lambda)} < q$. Since $\mu_{ind} = C/(B+C) < 1/2$, we have $C < B$. Using Equations (22) and (23), we get

$$\frac{\underline{\mu}(\lambda)}{1-\overline{\mu}(\lambda)} = \frac{e^{C/\lambda} - 1}{e^{B/\lambda} - 1}.$$

Since $C < B$, we have $\lim_{\lambda \to 0^+} \frac{\underline{\mu}(\lambda)}{1-\overline{\mu}(\lambda)} = 0$. Because $q > 0$, we deduce the existence of $\lambda^* \in (0, \overline{\lambda})$ small enough such that $\frac{\underline{\mu}(\lambda^*)}{1-\overline{\mu}(\lambda^*)} < q$. Hence $p_H(\lambda^*) > \mu_0$, and therefore $\tilde{V}_{\mu_0}(\lambda^*) > \tilde{V}_{\mu_0}(0) > \tilde{V}_{\mu_0}(\overline{\lambda})$. Thus neither boundary solves the Obfuscator's problem, and the solution is partial obfuscation.

Necessity is proved case by case. We show that whenever the Learner is not Skeptical, the solution is a boundary solution. If $\mu_0 > \mu_{ind}$, then the Learner chooses a_H at the prior. Since a_H is the Obfuscator's preferred action, this is Agreement at the Prior. Full obfuscation gives payoff $V(1)$, which is the maximal payoff the Obfuscator can obtain. Hence $\overline{\lambda}$ solves the Obfuscator's problem, and we have a boundary solution (full obfuscation). It remains to consider the case when $\mu_0 < \mu_{ind}$, but the Learner is not Skeptical, so $\mu_{ind} \geqslant \frac{1}{2}$. Equivalently, $C \geqslant B$, and so

$$\frac{e^{C/\lambda} - 1}{e^{B/\lambda} - 1} \geqslant \frac{C}{B} \text{ for all } \lambda > 0,$$

as $x \mapsto \frac{e^{x/\lambda}-1}{x}$ is increasing for $x > 0$. For any $\lambda \in (0, \overline{\lambda})$, we want to show that $p_H(\lambda) < \mu_0$, which holds if, and only if, $\frac{\underline{\mu}(\lambda)}{1-\overline{\mu}(\lambda)} > \frac{\mu_0}{1-\mu_0}$, as in Equation (B.1). But

$$\mu_0 < \mu_{ind} \implies \frac{\mu_0}{1-\mu_0} < \frac{\mu_{ind}}{1-\mu_{ind}} = \frac{C}{B} \leqslant \frac{e^{C/\lambda} - 1}{e^{B/\lambda} - 1} = \frac{\underline{\mu}(\lambda)}{1-\overline{\mu}(\lambda)}.$$

Therefore, $p_H(\lambda) < \mu_0$ for any $\lambda \in (0, \overline{\lambda})$, and this is not a problem of partial obfuscation. We now prove part (ii). Assume we have agreement at the prior. Let $a^*(\mu_0) = \{\overline{a}\}$, where $\overline{a}$ is optimal for the Obfuscator, and we have $\tilde{V}_{\mu_0}(\lambda) = P_\lambda(a_H)v(a_H) + P_\lambda(a_L)v(a_L) \leqslant v(\overline{a})$ for every λ. Since we have agreement at the prior, $\tilde{V}_{\mu_0}(\overline{\lambda}) = v(\overline{a})$, and $\overline{\lambda}$ solves the Obfuscator's problem. Now, assume we have full obfuscation, i.e., $\overline{\lambda}$ is a solution to the Obfuscator's

problem. As before, assume $v(a_H) > v(a_L)$, or equivalently, that $V(1) > V(0)$. Then, $\{a_H\} = \arg\max_{a \in A} v(a)$. We need to show that $\{a_H\} = a^*(\mu_0) = \arg\max_{a \in A} \mathbb{E}_{\mu_0}[u(a, \omega)]$. Since we have full obfuscation,

$$\begin{aligned} &\tilde{V}_{\mu_0}(\overline{\lambda}) \geqslant \tilde{V}_{\mu_0}(\lambda)\ \forall \lambda \in \mathbb{R}_+ \implies V(\mu_0) \geqslant \tilde{V}_{\mu_0}(\lambda)\ \forall \lambda \in \mathbb{R}_+ \\ \implies &V(\mu_0) \geqslant \tilde{V}_{\mu_0}(0) = \mu_0 V(1) + (1 - \mu_0) V(0) \end{aligned} \tag{B.2}$$

With state-independent preferences, $V(\mu) \in \{V(0), V(1)\}$ for all μ, so $V(\mu_0) \in \{V(0), V(1)\}$. As μ_0 is interior and $V(1) > V(0)$, Equation (B.2) implies that $V(\mu_0) = V(1) = v(a_H)$, which in turn gives $a^*(\mu_0) = \{a_H\}$ and we have agreement at the prior. The only other possible case is of full transparency, which gives us (iii). □

B.3 The Optimal Information Cost

Proof of Proposition 6. For $\lambda \in (0, \overline{\lambda})$, substituting the two posterior thresholds into Equation (24) gives

$$\tilde{V}_{\mu_0}(\lambda) = P_\lambda(a_H) = \mu_0 + \frac{\mu_0}{e^{C/\lambda} - 1} - \frac{1 - \mu_0}{e^{B/\lambda} - 1}.$$

Differentiating with respect to λ, we get

$$\frac{d\tilde{V}_{\mu_0}(\lambda)}{d\lambda} = \frac{1}{4\lambda^2} \left[C\mu_0 \operatorname{csch}^2\left(\frac{C}{2\lambda}\right) - B(1 - \mu_0) \operatorname{csch}^2\left(\frac{B}{2\lambda}\right) \right],$$

where csch denotes the hyperbolic cosecant. Rearranging gives

$$\frac{d\tilde{V}_{\mu_0}(\lambda)}{d\lambda} \gtrless 0 \iff \operatorname{csch}^2\left(\frac{C}{2\lambda}\right) \gtrless \frac{1 - \mu_0}{\mu_0} \frac{B}{C} \operatorname{csch}^2\left(\frac{B}{2\lambda}\right).$$

In what follows, let

$$G(\lambda) \coloneqq \frac{\operatorname{csch}^2\left(\frac{C}{2\lambda}\right)}{\operatorname{csch}^2\left(\frac{B}{2\lambda}\right)} = \left[\frac{\sinh\left(\frac{B}{2\lambda}\right)}{\sinh\left(\frac{C}{2\lambda}\right)} \right]^2,$$

and $K \coloneqq \frac{1-\mu_0}{\mu_0} \frac{B}{C}$, so that

$$\operatorname{sgn}\left(\frac{d\tilde{V}_{\mu_0}(\lambda)}{d\lambda}\right) = \operatorname{sgn}\left(G(\lambda) - K\right).$$

Moreover,

$$\frac{d}{d\lambda}\ln G(\lambda) = \frac{1}{\lambda^2}\left[C\coth\left(\frac{C}{2\lambda}\right) - B\coth\left(\frac{B}{2\lambda}\right)\right].$$

The function $x \mapsto x\coth x$ is strictly increasing on $(0,\infty)$. Therefore, G is strictly decreasing when $B > C$ and strictly increasing when $B < C$. If $B = C$, then $G(\lambda) = 1$ for every $\lambda > 0$. In this last case, $K = \frac{1-\mu_0}{\mu_0} \neq 1$, because $\mu_0 \neq \mu_{ind} = 1/2$ by assumption. Consequently, in every case the equation $G(\lambda) = K$ has at most one solution. It follows that there exists at most one $\lambda^* \in (0, \overline{\lambda})$ satisfying

$$\frac{d\tilde{V}_{\mu_0}(\lambda^*)}{d\lambda} = 0.$$

□

B.4 Symmetric Payoffs and Concavity

Proof of Lemma 3. Because the Learner's preferences are symmetric, $B = C$ and $\overline{\mu}(\lambda) = 1 - \underline{\mu}(\lambda)$. By Equation (24), for $0 \leqslant \lambda \leqslant \overline{\lambda}$ the probability of choosing a_H is

$$P_\lambda(a_H) = \frac{\mu_0 - \underline{\mu}(\lambda)}{1 - 2\underline{\mu}(\lambda)} = \frac{1}{2} + \frac{\mu_0 - 1/2}{1 - 2\underline{\mu}(\lambda)}.$$

The denominator decreases as λ increases. Thus the Obfuscator's payoff increases up to the no-learning cutoff if $\mu_0 > 1/2$ and decreases up to that cutoff if $\mu_0 < 1/2$. Beyond the cutoff, the payoff is constant. □

Proof of Proposition 7. Suppose V is concave and take $0 \leqslant \lambda_1 < \lambda_2 \leqslant \overline{\lambda}$. The posterior interval for λ_2 lies inside that for λ_1 and both contain μ_0. Let L_1 and L_2 be the corresponding chords of V. By concavity, L_1 is weakly below V at both endpoints of the smaller interval, where L_2 equals V. Since both chords are affine, $L_1(\mu_0) \leqslant L_2(\mu_0)$. This gives the desired payoff comparison. The inequality reverses for convex V, and the payoff is constant beyond $\overline{\lambda}$. □

B.5 Policy Design for Transparency

Proof of Theorem 2. Assume that we have a Skeptical Learner. Then $C < B$, so $\underline{\varepsilon} = B - C > 0$. Lowering $u(a_H, \omega_L)$ by any $\varepsilon \geqslant \underline{\varepsilon}$ raises the cost of buying a bad product to $C + \varepsilon \geqslant B$. The new belief of indifference is therefore at least $1/2$, so Theorem 1 rules out partial obfuscation. Since the prior is below the new belief of indifference, full obfuscation induces a_L and gives the Obfuscator zero. Full transparency gives him $\mu_0 > 0$, so he chooses full transparency and the Learner obtains the first-best payoff u_0^*.

Now suppose we have Agreement at the Prior, so $\mu_{ind} < \mu_0$. Let

$$\underline{\varepsilon} = \max\left\{\frac{\mu_0}{1-\mu_0}B - C,\ B - C\right\}.$$

Reducing $u(a_H, \omega_L)$ by any $\varepsilon > \underline{\varepsilon}$ raises the mistake cost to $C + \varepsilon$. The new belief of indifference is strictly above μ_0 and at least $1/2$. We therefore have neither Agreement at the Prior nor a Skeptical Learner. By Theorem 1, the Obfuscator chooses full transparency. Since correct-action payoffs have not changed, the Learner obtains the same first-best payoff u_0^*. □

B.6 Welfare and Mistake Costs

The next two lemmas are used to prove Proposition 8.

Lemma 10. *Fix an Obfuscator's Problem and assume we have partial obfuscation, so $\lambda^*(C) \in (0, \overline{\lambda})$. Then, $\lambda^*(C)$ is continuously differentiable, and*

$$\frac{d\lambda^*(C)}{dC} < 0.$$

Proof of Lemma 10. By Proposition 6, the equilibrium $\lambda^*(C)$ is unique, and since we have a Skeptical Learner, $\lambda^*(C) > 0$. Using our normalization and the fact that $\operatorname{csch}^2(x) = \operatorname{csch}^2(-x)$, we can simplify the Obfuscator's FOC to

$$\frac{d\tilde{V}_{\mu_0}(\lambda)}{d\lambda} = \frac{v(a_H) - v(a_L)}{4\lambda^2}\left[-B(1-\mu_0)\operatorname{csch}^2\left(\frac{B}{2\lambda}\right) + C\mu_0\operatorname{csch}^2\left(\frac{C}{2\lambda}\right)\right].$$

Since, by assumption, $v(a_H) - v(a_L) > 0$, at the optimum $\lambda^*(C)$, the FOC implies

$$\begin{aligned} -B(1-\mu_0)\operatorname{csch}^2\left(\frac{B}{2\lambda^*(C)}\right) + C\mu_0\operatorname{csch}^2\left(\frac{C}{2\lambda^*(C)}\right) &= 0 \\ \implies \operatorname{csch}^2\left(\frac{C}{2\lambda^*(C)}\right) &= \frac{1-\mu_0}{\mu_0}\frac{B}{C}\operatorname{csch}^2\left(\frac{B}{2\lambda^*(C)}\right). \end{aligned}$$

Using the fact that $\operatorname{csch}(x) = 1/\sinh(x)$ and simplifying,

$$\frac{\sinh\left(\frac{B}{2\lambda^*(C)}\right)}{\sinh\left(\frac{C}{2\lambda^*(C)}\right)} = \sqrt{\frac{1-\mu_0}{\mu_0}\frac{B}{C}}. \tag{B.3}$$

Now, let

$$h(\lambda, C) \coloneqq 2\ln\left(\sinh\left(\frac{B}{2\lambda}\right)\right) - 2\ln\left(\sinh\left(\frac{C}{2\lambda}\right)\right) - \ln\left(\left(\frac{1-\mu_0}{\mu_0}\frac{B}{C}\right)\right),$$

and the FOC is reduced to $h(\lambda^*(C), C) = 0$. If we differentiate in each case, we get

$$\begin{aligned} h_\lambda(\lambda, C) &= \frac{1}{\lambda^2}\left[C\coth\left(\frac{C}{2\lambda}\right) - B\coth\left(\frac{B}{2\lambda}\right)\right], \text{ and} \\ h_C(\lambda, C) &= \frac{1}{C} - \frac{1}{\lambda}\coth\left(\frac{C}{2\lambda}\right). \end{aligned}$$

Note that $x \mapsto x\coth x$ is strictly increasing on $(0, \infty)$, and since $B > C$, $h_\lambda < 0$. Moreover, letting $x = C/(2\lambda) > 0$, we obtain

$$h_C(\lambda, C) = \frac{1 - 2x\coth x}{C} < 0$$

as $x\coth x \geqslant 1$. Therefore, the Implicit Function Theorem implies that $\lambda^*(C)$ is continuously differentiable and

$$\frac{d\lambda^*(C)}{dC} = -\frac{h_C(\lambda^*(C), C)}{h_\lambda(\lambda^*(C), C)} < 0.$$

□

Lemma 11. *Under the assumptions of Lemma 10, let $W^*(C)$ denote the Learner's ex-ante equilibrium welfare. If we have a Skeptical Learner, then*

$$\frac{dW^*(C)}{dC} = -(1-\mu_0)\mathbb{P}_{\lambda,C}(a_H \mid \omega_L) + MI^*(C)\left|\frac{d\lambda^*(C)}{dC}\right|, \tag{B.4}$$

where $\mathbb{P}_{\lambda,C}(a_H|\omega_L)$ is the equilibrium probability that the Learner chooses a_H given ω_L, and

$MI^*(C) = H(\mu_0) - \mathbb{E}_{\mu\sim\tau^*_{\lambda,C}}[H(\mu)]$ *is the mutual information acquired by the Learner. All quantities are evaluated at* $\lambda = \lambda^*(C)$.

Proof of Lemma 11. Under our normalization and tie-breaking rule, the Learner chooses a_H when $\mu \geqslant C/(B+C)$ and a_L otherwise. Her indirect material utility is $\max\{0, (B+C)\mu - C\}$, so for a fixed pair (λ, C) her welfare is

$$W(\lambda, C) \coloneqq \max_{\tau} \mathbb{E}_{\mu\sim\tau}[\max\{0, (B+C)\mu - C\}] - \lambda(H(\mu_0) - \mathbb{E}_{\mu\sim\tau}[H(\mu)]),$$
$$\text{s.t. } \mathbb{E}_{\mu\sim\tau}[\mu] = \mu_0.$$

Let $\tau^*_{\lambda,C}$ be an optimal experiment. We again denote the supported elements by $\underline{\mu}$ ($\overline{\mu}$) when the recommended action is a_L (a_H). We then get

$$W(\lambda, C) = \tau^*_{\lambda,C}(\overline{\mu})[(B+C)\overline{\mu} - C] - \lambda(H(\mu_0) - (\tau^*_{\lambda,C}(\overline{\mu})H(\overline{\mu}) + \tau^*_{\lambda,C}(\underline{\mu})H(\underline{\mu})))$$

By the envelope theorem[33], $W_C(\lambda, C) = \tau^*_{\lambda,C}(\overline{\mu})(\overline{\mu} - 1)$. But $\overline{\mu} = \mathbb{P}_{\lambda,C}(\omega_H \mid a_H)$. Therefore, $1 - \overline{\mu} = \mathbb{P}_{\lambda,C}(\omega_L \mid a_H)$, and since $\tau^*_{\lambda,C}(\overline{\mu}) = \mathbb{P}_{\lambda,C}(a_H)$, the unconditional probability that a_H is chosen given (λ, C), we get

$$\tau^*_{\lambda,C}(\overline{\mu})(1 - \overline{\mu}) = \mathbb{P}_{\lambda,C}(a_H)\mathbb{P}_{\lambda,C}(\omega_L \mid a_H) = \mathbb{P}_{\lambda,C}(a_H, \omega_L).$$

Therefore,

$$W_C(\lambda, C) = -\mathbb{P}_{\lambda,C}(a_H, \omega_L) = -(1 - \mu_0)\mathbb{P}_{\lambda,C}(a_H \mid \omega_L).$$

Using the envelope theorem again, now for λ, gives

$$W_\lambda(\lambda, C) = -(H(\mu_0) - \mathbb{E}_{\mu\sim\tau^*_{\lambda,C}}[H(\mu)]).$$

Now let $W^*(C) \coloneqq W(\lambda^*(C), C)$ denote the Learner's equilibrium welfare as a function of

[33]Note that the maximizing function is smooth at the posteriors when we have partial obfuscation.

C. Evaluating the following expressions at $\lambda = \lambda^*(C)$ and using Lemma 10, we get

$$\begin{aligned}
\frac{dW^*(C)}{dC} &= W_C(\lambda^*(C), C) + W_\lambda(\lambda^*(C), C)\frac{d\lambda^*(C)}{dC} \\
&= -(1-\mu_0)\mathbb{P}_{\lambda,C}(a_H \mid \omega_L) - (H(\mu_0) - \mathbb{E}_{\mu\sim\tau^*_{\lambda,C}}[H(\mu)])\frac{d\lambda^*(C)}{dC} \\
&= -(1-\mu_0)\mathbb{P}_{\lambda,C}(a_H \mid \omega_L) + (H(\mu_0) - \mathbb{E}_{\mu\sim\tau^*_{\lambda,C}}[H(\mu)])\left|\frac{d\lambda^*(C)}{dC}\right| \\
&= -(1-\mu_0)\mathbb{P}_{\lambda,C}(a_H \mid \omega_L) + MI^*(C)\left|\frac{d\lambda^*(C)}{dC}\right|
\end{aligned}$$

□

Proof of Proposition 8. Let $A = \frac{1-\mu_0}{\mu_0}$ and $C_0 = \frac{\mu_0}{1-\mu_0}B$. Note that $C \in (C_0,\ C_0+\eta) \cup (B-\eta, B)$ means that, even though we have a Skeptical Learner, we approach indifference at the prior as $C \downarrow C_0$, or full transparency as $C \uparrow B$. We first show that there exist $\eta_1, \eta_2 > 0$ such that $\frac{dW^*(C)}{dC} > 0$ whenever $C \in (C_0, C_0+\eta_2)$ or $C \in (B-\eta_1, B)$. By Lemma 11, it is sufficient to check the values of terms on the right-hand side of Equation (B.4) near the boundaries. Consider the case where $C \to B^-$. We know from Equation (B.3) that as $C \to B^-$, the right-hand side converges to $\sqrt{A} > 1$ (as $\mu_0 < \frac{1}{2}$). This implies $\lambda^*(C) \to 0$. Since $\sinh(x) \simeq \frac{1}{2}e^x$ as $x \to \infty$, we get

$$\lambda^*(C) \sim \frac{B-C}{\ln A}.$$

To obtain the derivative, use the implicit formula in Lemma 10. Writing $\lambda^* = \lambda^*(C)$, we have $h_C(\lambda^*, C) \sim -1/\lambda^*$ and

$$h_\lambda(\lambda^*, C) = \frac{-(B-C) + O(e^{-C/\lambda^*})}{(\lambda^*)^2} \sim -\frac{B-C}{(\lambda^*)^2}.$$

The remainder follows from $\coth(x) = 1 + O(e^{-2x})$ and is $o(B-C)$ since $B - C \sim \lambda^* \ln A$. Thus $d\lambda^*(C)/dC = -h_C/h_\lambda \to -1/\ln A$. But as $\lambda^*(C) \to 0$, the Learner gets full information and the mistake probability goes to zero, i.e., $\mathbb{P}(a_H \mid \omega_L) \to 0$. Moreover, $MI^*(C) \to H(\mu_0) > 0$. Therefore,

$$\lim_{C\to B^-} \frac{dW^*(C)}{dC} = \frac{H(\mu_0)}{\ln A} > 0.$$

By continuity, there exists $\eta_1 > 0$ such that $\frac{dW^*(C)}{dC} > 0$ for all $C \in (B-\eta_1, B)$.

Now consider the case when $C \to C_0$. Let $r \coloneqq \frac{B}{C} \in (1, A)$ and $z \coloneqq \frac{C}{2\lambda^*(C)}$, so that $\lambda^*(C) = \frac{B}{2rz}$.[34] The Obfuscator's first-order condition becomes $\frac{\sinh(rz)}{\sinh z} = \sqrt{rA}$. As $C \to C_0$, we have $r \to A$. We simplify the FOC using a Taylor expansion around $z = 0$, which yields

$$A - r = \frac{A(A^2-1)}{3} z^2 + O(z^4).$$

Moreover, differentiating the first-order condition in (r, z) coordinates gives

$$\frac{d\lambda^*(C)}{dC} = \frac{\frac{1}{2} - z\coth z}{2z^2\Big(r\coth(rz) - \coth z\Big)}.$$

Further known simplifications $z\coth z = 1 + \frac{z^2}{3} + O(z^4)$ and $r\coth(rz) - \coth z = \frac{r^2-1}{3} z + O(z^3)$ give

$$\left|\frac{d\lambda^*(C)}{dC}\right| = \frac{3}{4(r^2-1)} z^{-3} + O(z^{-1}). \tag{B.5}$$

We can rewrite $\underline{\mu}, \overline{\mu}$ using r, z as

$$\underline{\mu} = \frac{1 - e^{-2z}}{e^{2rz} - e^{-2z}}, \text{ and } \overline{\mu} = \frac{e^{2rz}(1 - e^{-2z})}{e^{2rz} - e^{-2z}}.$$

A Taylor expansion then gives

$$\overline{\mu} - \underline{\mu} = \frac{2r}{r+1} z + O(z^2) = \left(\frac{2A}{A+1} + o(1)\right) z \text{ as } C \downarrow C_0. \tag{B.6}$$

Hence, there exists $\iota > 0$ such that $\overline{\mu} - \underline{\mu} \geqslant \iota z$.

Let $p \coloneqq \mathbb{P}(a_H)$ be the unconditional probability of choosing a_H at the optimum. Since $r = A + O(z^2)$ and $\mu_0 = 1/(A+1)$, the posterior expansions give

$$\overline{\mu} - \mu_0 = \frac{A}{A+1} z + O(z^2), \qquad \mu_0 - \underline{\mu} = \frac{A}{A+1} z + O(z^2).$$

Bayes plausibility gives $p = (\mu_0 - \underline{\mu})/(\overline{\mu} - \underline{\mu}) \to 1/2$. Hence $p(1-p) \geqslant \kappa$ for some $\kappa > 0$ near C_0. Let $D(p\|q)$ denote the Kullback-Leibler divergence between distributions p and q. Then

$$MI^*(C) = pD(\overline{\mu}\|\mu_0) + (1-p)D(\underline{\mu}\|\mu_0),$$

[34] Note that as $C \to C_0$, $r \to A$ and $z \to 0$.

Pinsker's inequality gives $D(q\|p) \geqslant 2(q-p)^2$, and, together with Bayes plausibility, we get

$$\begin{aligned} MI^*(C) &\geqslant 2\left[p(\overline{\mu}-\mu_0)^2 + (1-p)(\underline{\mu}-\mu_0)^2\right] \\ &= 2p(1-p)(\overline{\mu}-\underline{\mu})^2 \geqslant 2\kappa(\iota z)^2 \geqslant cz^2 \end{aligned}$$

by Equation (B.6) for some $c > 0$ if C is close enough to C_0. Together with Equation (B.5), this yields

$$MI^*(C)\left|\frac{d\lambda^*(C)}{dC}\right| \geqslant c' z^{-1} \to \infty \text{ as } C \downarrow C_0.$$

Since $0 \leqslant (1-\mu_0)\mathbb{P}_{\lambda^*(C),C}(a_H \mid \omega_L) \leqslant 1$, we get that $\frac{dW^*(C)}{dC} > 0$ for all C sufficiently close to C_0. Thus there exists $\eta_2 > 0$ such that $\frac{dW^*(C)}{dC} > 0$ for all $C \in (C_0, C_0+\eta_2)$. Finally, setting $\eta \coloneqq \min\{\eta_1, \eta_2\}$ yields the first part of the proof.

Within the Skeptical region, the equilibrium remains interior and W^* is differentiable. A strictly positive derivative at C therefore implies that there exists $\delta > 0$ such that $W^*(C+\varepsilon) > W^*(C)$ for every $\varepsilon \in (0, \delta)$. □

C Proofs for the Comparison to Communication Games

Proof of Proposition 9. We start by proving (i). Note that in Bayesian Persuasion, the Sender maximizes his utility by choosing a distribution of posteriors in the set

$$BP(\mu_0) \coloneqq \{\tau \in \triangle(\triangle(\Omega)) : \sum_{\mu \in \text{supp}(\tau)} \mu\tau(\mu) = \mu_0\}.$$

So, given commitment, the Sender is only restricted by Bayes plausibility. Under Strategic Obfuscation, the Obfuscator maximizes his utility in the set

$$SO(\mu_0) \coloneqq \{\tau \in \triangle(\triangle(\Omega)) : \exists \lambda \in \mathcal{O} \text{ with } \text{supp}(\tau) \subseteq \{\underline{\mu}(\lambda), \overline{\mu}(\lambda)\} \wedge \sum_{\mu \in \text{supp}(\tau)} \mu\tau(\mu) = \mu_0\}.$$

We can clearly see that $SO(\mu_0) \subseteq BP(\mu_0)$, and so the Obfuscator in Strategic Obfuscation cannot do better than the Sender in Bayesian Persuasion.

For the second statement of part (i), suppose we have Agreement at the Prior, with $v(a_H) > v(a_L)$. Then, $s^O = \tilde{V}_{\mu_0}(\overline{\lambda}) = V(\mu_0) = v(a_H)$. A fully uninformative signal is optimal

under BP, so $s^P = V(\mu_0) = v(a_H)$. Now, assume we do not have Agreement at the Prior. Then, $V(\mu_0) = v(a_L)$. Under BP, the optimal split of posteriors would put positive weight on the belief of indifference, which, as we have seen, is not reached under SO for any $\lambda \in \mathbb{R}_+$. Therefore, the concavification result is not reached.

For part (ii), state-independent preferences imply that every sender type ranks the receiver's action lotteries in the same way. In a Cheap Talk equilibrium, all messages used with positive probability must therefore induce the same probability of a_H; otherwise, every type would prefer a message with a higher probability. If the receiver mixes after every on-path message, each corresponding posterior must equal μ_{ind}, which would imply $\mu_0 = \mu_{ind}$ by Bayes plausibility, contrary to our maintained assumption. Thus the common action probability is either zero or one. Bayes plausibility and the receiver's optimality then imply that the chosen action is the one optimal at the prior. The sender's equilibrium payoff is consequently $V(\mu_0)$, which is attained by a babbling equilibrium. Under Strategic Obfuscation, choosing $\lambda = \overline{\lambda}$ gives this same payoff, so $s^O \geqslant s^C$. □

D Proofs for Other Information Cost Functions

D.1 Indifference at the Prior

Proof of Proposition 11. Fix $\lambda \in \mathbb{R}_+$ and $\mu_0 = \mu_{ind}$. With our normalization, we can write the Learner's problem as $\max_\tau \mathbb{E}_\tau[\max\{0, (B+C)\mu - C\}] - \lambda(\mathbb{E}_\tau[S_{\operatorname{supp}\mu_0}(\mu)] - S_{\operatorname{supp}\mu_0}(\mu_0))$. If the Learner acquires an uninformative signal δ_{μ_0} there is no cost of information, and she gets $\mathbb{E}_{\delta_{\mu_0}}[U(\mu)] = 0$ (given the normalization). To show that δ_{μ_0} is not optimal, we will construct an attention strategy τ_ε that yields a strictly positive payoff. Consider a symmetric signal that induces posteriors $\mu_H = \mu_{ind} + \varepsilon$ and $\mu_L = \mu_{ind} - \varepsilon$, each with probability $\frac{1}{2}$, for some $\varepsilon > 0$ small enough such that $\mu_H, \mu_L \in int(\Delta(\Omega))$. This strategy clearly satisfies Bayes plausibility. The expected utility under τ_ε is

$$\begin{aligned}\mathbb{E}_{\tau_\varepsilon}[U(\mu)] &= \frac{1}{2}\max\{0, (B+C)(\mu_{ind}+\varepsilon) - C\} + \frac{1}{2}\max\{0, (B+C)(\mu_{ind}-\varepsilon) - C\} \\ &= \frac{1}{2}\max\{0, (B+C)\varepsilon\} + \frac{1}{2}\max\{0, -(B+C)\varepsilon\} = \frac{1}{2}(B+C)\varepsilon.\end{aligned}$$

Next, we look at the cost of the signal (where we suppress the index on $S_{\operatorname{supp}\mu_0}$), which gives

$$c_\lambda(\mu_{ind}, \tau_\varepsilon) = \lambda \left[\frac{1}{2} S(\mu_{ind} + \varepsilon) + \frac{1}{2} S(\mu_{ind} - \varepsilon) - S(\mu_{ind}) \right].$$

Because S is differentiable at μ_{ind}, we can take the Taylor expansion of S around μ_{ind}:

$$S(\mu_{ind} \pm \varepsilon) = S(\mu_{ind}) \pm S'(\mu_{ind})\varepsilon + o(\varepsilon).$$

Substituting this into the cost function yields $c_\lambda(\mu_{ind}, \tau_\varepsilon) = \lambda \cdot o(\varepsilon)$. The net expected payoff from strategy τ_ε is therefore

$$L_\lambda(\tau_\varepsilon) = \frac{1}{2}(B + C)\varepsilon - \lambda \cdot o(\varepsilon).$$

Because $\frac{1}{2}(B + C) > 0$, for any $\lambda \in \mathbb{R}_+$, there exists an $\varepsilon > 0$ sufficiently small such that $\frac{1}{2}(B + C)\varepsilon > \lambda \cdot o(\varepsilon)$. Thus, $\delta_{\mu_0} \notin T_{\mu_0}(\lambda)$ for any $\lambda \in \mathbb{R}_+$. □

D.2 Robustness to Cost Functions

Proof of Proposition 12. Normalize $v(a_H) = 1$ and $v(a_L) = 0$. Under Shannon costs, in the Skeptical region the Obfuscator's expected payoff under full transparency is μ_0. Under full obfuscation, the Learner acquires no information and relies on the prior. Because $\mu_0 < \mu_{ind} = \frac{C}{B+C}$, the Learner's optimal action under the prior is a_L, yielding a payoff of 0 for the Obfuscator. Therefore, $\tilde{V}_{\mu_0}(0) > \tilde{V}_{\mu_0}(\overline{\lambda})$.

Let $\lambda_0 \in (0, \overline{\lambda})$ be the optimal cost under Shannon. Since the Learner is Skeptical, $\tilde{V}_{\mu_0}(\lambda_0) > \tilde{V}_{\mu_0}(0)$. Its two posteriors lie in the interior of J by hypothesis. Hence, the probability the Learner chooses a_H under Shannon, denoted $p_E(\lambda_0)$, satisfies $p_E(\lambda_0) > \mu_0$.

For any strictly convex and C^2 posterior-separable cost function S, the Learner's optimal interior posteriors $\underline{\mu}$ and $\overline{\mu}$ for a given λ are characterized by the tangency conditions (equal slopes and intercepts for the net utility functions) in Caplin et al. (2022). This can be

rewritten as a system of equations $F(\underline{\mu}, \overline{\mu}; S) = 0$, where

$$F(\underline{\mu}, \overline{\mu}; S) := \begin{pmatrix} S'(\overline{\mu}) - S'(\underline{\mu}) - \frac{B+C}{\lambda_0} \\ S(\overline{\mu}) - S(\underline{\mu}) - S'(\underline{\mu})(\overline{\mu} - \underline{\mu}) - \frac{(B+C)\overline{\mu} - C}{\lambda_0} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

At $S = S_E$, let the unique roots be $(\underline{\mu}_E, \overline{\mu}_E)$. The Jacobian matrix of F with respect to $(\underline{\mu}, \overline{\mu})$ evaluated at S_E is given by

$$\mathcal{J} = \begin{pmatrix} -S_E''(\underline{\mu}_E) & S_E''(\overline{\mu}_E) \\ -S_E''(\underline{\mu}_E)(\overline{\mu}_E - \underline{\mu}_E) & S_E'(\overline{\mu}_E) - S_E'(\underline{\mu}_E) - \frac{(B+C)}{\lambda_0} \end{pmatrix}$$

But note that at a solution of the first tangency condition, the bottom-right element equals zero. Therefore, $\det \mathcal{J} = S_E''(\underline{\mu}_E) S_E''(\overline{\mu}_E)(\overline{\mu}_E - \underline{\mu}_E)$. For the Shannon case,

$$S_E(\mu) = -H(\mu) = \mu \ln(\mu) + (1 - \mu) \ln(1 - \mu),$$

and at any interior belief,

$$S_E''(\mu) = \frac{1}{\mu(1 - \mu)} > 0 \implies \det \mathcal{J} = \frac{\overline{\mu}_E - \underline{\mu}_E}{\overline{\mu}_E(1 - \overline{\mu}_E)\underline{\mu}_E(1 - \underline{\mu}_E)} > 0$$

The Jacobian is nonsingular, so the Implicit Function Theorem gives nearby roots $\underline{\mu}(S)$ and $\overline{\mu}(S)$ that depend continuously on S in the $C^2(J)$ topology. Shrinking the neighborhood if necessary, these roots remain in the interior of J and satisfy $\underline{\mu}(S) < \mu_0 < \overline{\mu}(S)$. For every admissible strictly convex generator in this neighborhood, the tangency conditions give a common supporting line to the two net action payoffs. Convexity of S ensures that this line dominates both payoffs on the entire belief interval. The two posteriors therefore solve the Learner's problem, even though closeness to Shannon is required only on J.

By Bayes plausibility, the unconditional probability of selling under the cost function S is

$$p_S(\lambda_0) = \frac{\mu_0 - \underline{\mu}(S)}{\overline{\mu}(S) - \underline{\mu}(S)}$$

By continuity of the two posteriors, $p_S(\lambda_0)$ is continuous in S and equals $p_E(\lambda_0) > \mu_0$ at $S = S_E$. Thus, in a sufficiently small neighborhood, this fixed cost still gives the Obfuscator a payoff strictly above full transparency. Whenever no learning is induced, his payoff is zero because a_L is uniquely optimal at the prior.

It remains to check that a finite optimum exists for these costs. Let ℓ be the tangent to S at μ_0 and let $d(\mu) = S(\mu) - \ell(\mu)$. Strict convexity gives $d(\mu) \geqslant 0$, with equality only at μ_0, so continuity gives $\delta = \min_{\mu \in [\mu_{ind}, 1]} d(\mu) > 0$. For $\lambda > B/\delta$, the affine function $-\lambda\ell$ dominates both net action payoffs: the a_L payoff is $-\lambda S$, while the material payoff from a_H is nonpositive below μ_{ind} and at most B above it. At μ_0, the bound is attained by a_L. Thus no learning is uniquely optimal at all sufficiently high costs.

Since S is continuous on the compact belief interval, its expected information cost is continuous in the direct-experiment representation, including at zero-probability recommendations. The compactness and upper semicontinuity argument in the proof of Proposition 3 therefore applies. The Obfuscator attains a maximum, and the strict gain at λ_0 rules out both endpoints. Hence the solution is partial obfuscation. □

E Costly Obfuscation

In this section, we extend the model to allow for costly obfuscation and briefly discuss our conclusions. We consider that it is costly to hide or provide information relative to a status quo, which depends, for instance, on the technology available. Think of a couple planning where to go out for dinner at the cheapest restaurant in the neighborhood. In the past, finding such a restaurant required making a list of local places, then calling each in turn; now, all it takes is a quick search on Google Maps.

To allow for this status quo, we add a parameter $\lambda_0 \in \mathbb{R}_+$, which we call the *natural* marginal cost of information. Obfuscation (or revealing) becomes costlier the farther the chosen λ is from λ_0. In the restaurant example, before Google Maps was invented, we had a high λ_0. After Google Maps, we have a low λ_0.

Definition 5. *The Obfuscator's Problem with Costly Obfuscation is*

$$\max_{\lambda \in \mathbb{R}_+} \tilde{V}_{\mu_0}(\lambda) - \kappa(\lambda - \lambda_0)^2. \tag{E.1}$$

Here, $\kappa \geqslant 0$ measures the cost of departing from the natural information cost λ_0. When $\kappa = 0$, this is the original problem, for which Proposition 3 guarantees a solution when the Learner has a unique optimal action at the prior. When $\kappa > 0$, the Obfuscator's gross

payoff is bounded and the quadratic cost grows without bound as $\lambda \to \infty$. Together with upper semicontinuity, this guarantees a finite optimum even when the Learner has several prior-optimal actions.

With a unique prior-optimal action, the gross payoff $\tilde{V}_{\mu_0}$ is constant above $\overline{\lambda}$. We can then restrict the costly problem to $[0, \max\{\overline{\lambda}, \lambda_0\}]$. Keeping λ_0 in this interval matters: if $\lambda_0 > \overline{\lambda}$, choosing the natural cost induces no learning without paying to move away from it.

If the natural information cost λ_0 already maximizes the Obfuscator's gross payoff $\tilde{V}_{\mu_0}$, it remains optimal with the adjustment cost. Moving away from λ_0 then incurs a cost without increasing the gross payoff.

For a simple example, take $B = C = 2$ and $\mu_0 = 0.25$ in the binary model, with $v(a_H) = 2$ and $v(a_L) = 0$. Without an adjustment cost, Theorem 1 gives full transparency. Now set $\kappa = 0.2$ and $\lambda_0 = 1$. Moving all the way to full transparency or full obfuscation is costly. Figure E.1 shows that the Obfuscator now chooses an intermediate cost.

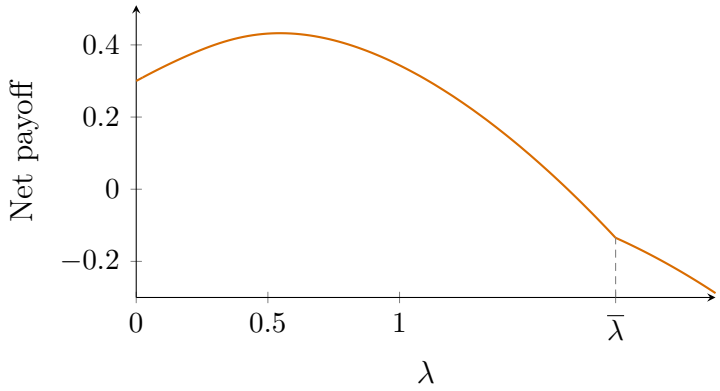


Figure E.1: The Obfuscator's payoff net of adjustment costs when $\lambda_0 = 1$ and $\kappa = 0.2$. The kink is at $\overline{\lambda} = 2/\ln 3 \approx 1.82$, where the Learner stops acquiring information and the gross payoff becomes constant.

In this example, reducing κ or λ_0 can bring the chosen cost closer to full transparency. The effect more generally depends on where the natural information cost lies relative to the gross-payoff maximizer.

F A Broader Payoff Environment

In this section, we relax the matching-payoff assumption and allow the Obfuscator's preferences to depend on the state. The number of states and actions may also differ. We present the broader setup below; the information acquisition technology, timing, and tie-breaking rule are the same as in the main model.

There are two players, the Obfuscator (*he*) and the Learner (*she*). The Learner chooses an action from a finite set $A = \{a_1, \ldots, a_N\}$. We denote by $\Omega = \{\omega_1, \ldots, \omega_M\}$ the finite set of payoff-relevant states of nature. The state is drawn from a common prior $\mu_0 \in int(\triangle(\Omega))$, known by both players, but neither player observes its realization before the information acquisition stage.[35]

The Learner and the Obfuscator have material utility functions $u : A \times \Omega \to \mathbb{R}$ and $v : A \times \Omega \to \mathbb{R}$, respectively. We let $a^*(\mu) \coloneqq \arg\max_{a \in A} \mathbb{E}_\mu[u(a, \omega)]$ represent the set of actions that are optimal for the Learner given a belief $\mu \in \triangle(\Omega)$. We break ties in favor of the Obfuscator, both when the Learner chooses an action and when she chooses how much to learn.[36] Our solution concept is the Obfuscator-preferred subgame perfect equilibrium, which we simply refer to as "equilibrium" throughout the paper.

We restrict A to contain only actions that are optimal for the Learner for at least one belief.[37] We further assume that no two actions give the Learner the same payoffs in every state. Finally, we assume that there is only one optimal action at the prior, so that $a^*(\mu_0)$ is a singleton. By Proposition 2, this ensures that full obfuscation is possible at a finite information cost.[38] This property holds almost everywhere in the space of beliefs.[39]

The Learner may purchase a signal structure (or Blackwell experiment) $\pi = \{\pi(\cdot|\omega)\}_{\omega \in \Omega}$

[35]We denote by $\triangle(\Omega)$ the space of all probability distributions on Ω, and by $int(X)$ the interior of set X, taken relative to the belief simplex for sets of beliefs. Measure-zero statements below refer to Lebesgue measure on the affine hull of the simplex.

[36]While this is a widely used assumption in the Bayesian Persuasion literature (Kamenica and Gentzkow, 2011), its implications in our model are two-fold: when the Learner is indifferent among different distributions of posteriors, she chooses the one favored by the Obfuscator; in addition, if the Learner is indifferent between any two actions at an induced posterior, she again chooses the one favored by the Obfuscator.

[37]The Learner would never choose any other actions, and they can be removed without loss of generality.

[38]We could remove this assumption if we allow for a cost of obfuscation, as in the extension in Appendix E.

[39]For any two actions, indifference is defined by a linear equation in beliefs. Since their payoff vectors differ, this equation describes a set of lower dimension in the simplex. With finitely many actions, the set of beliefs at which the optimal action is not unique has measure zero.

over the set A of signals, interpreted as recommended actions.[40] It will be convenient to work directly with the induced distribution of posteriors $\tau \in \triangle(\triangle(\Omega))$ (Kamenica and Gentzkow, 2011). If recommendation a is realized with positive probability, the posterior μ^a is calculated by Bayes' rule:

$$\mu^a(\omega) = \frac{\pi(a|\omega)\mu_0(\omega)}{\sum_{m=1}^{M} \pi(a|\omega_m)\mu_0(\omega_m)}. \tag{F.1}$$

The Learner is rationally inattentive. Upon choosing a distribution of posteriors τ, she incurs an information cost given by

$$c_\lambda(\tau) \coloneqq \lambda(H(\mu_0) - \mathbb{E}_\tau[H(\mu)]), \tag{F.2}$$

where $\lambda \geqslant 0$ is the marginal cost of information and $H : \triangle(\Omega) \to \mathbb{R}_+$ is the Shannon entropy (Shannon, 1948),

$$H(\mu) = -\sum_{m=1}^{M} \mu(\omega_m) \ln(\mu(\omega_m)).$$

Entropy can be interpreted as a measure of uncertainty, so the cost function charges for how much uncertainty the signal reduces.[41] There are two ways a signal can have zero cost: (i) the Obfuscator sets $\lambda = 0$ and learning is free, or (ii) the Learner chooses an uninformative signal, so the posterior equals the prior and there is no change in entropy.

F.1 The Learner's Rational Inattention Problem

Let $U(\mu) \coloneqq \max_{a \in A} \mathbb{E}_\mu[u(a, \omega)]$ denote her indirect utility at belief μ.

Given μ_0 and $\lambda \in \mathbb{R}_+$, the Learner solves

$$\begin{aligned} &\max_{\tau \in \triangle(\triangle(\Omega))} \mathbb{E}_\tau\left[U(\mu)\right] - \lambda(H(\mu_0) - \mathbb{E}_\tau[H(\mu)]), \\ &\text{s.t. } \mathbb{E}_\tau[\mu] = \mu_0. \end{aligned} \tag{F.3}$$

The restriction is Bayes plausibility: averaging the possible posteriors using their prob-

[40] It is without loss to treat signal realizations as recommended actions. Signals that induce the same action can be combined without changing behavior and while weakly reducing information costs; see Matějka and McKay (2015).

[41] For beliefs on the boundary of $\triangle(\Omega)$, we let $0 \cdot \ln(0) \coloneqq \lim_{x \to 0^+} x \ln x = 0$, as usual.

abilities must recover the prior. It ensures that the learning outcomes are consistent with the initial beliefs. We denote by $T_{\mu_0}(\lambda)$ the set of solutions to (F.3) for a given λ, and by τ_λ a solution selected under our tie-breaking rule. To study the Obfuscator's problem, it will also be useful to think of this solution in terms of the induced action probabilities. Indeed, Matějka and McKay (2015) show that we can alternatively write the Learner's problem as a choice of conditional action probabilities $P_\lambda(a|\omega)$ (Corollary 1). For an experiment π_λ that induces τ_λ, these conditional probabilities are $P_\lambda(a|\omega) = \pi_\lambda(a|\omega)$. We write $P_\lambda(a) \coloneqq \sum_{m=1}^{M} \mu_0(\omega_m) P_\lambda(a|\omega_m)$ for the unconditional probability of choosing a. As in Section 2.1, the direct representation uses one selected action per posterior, with a fixed choice among any remaining ties. Thus, $P_\lambda(a) = \tau_\lambda(\mu^a)$ for any action chosen with positive probability. We let $A(\tau_\lambda) = \{a \in A : P_\lambda(a) > 0\}$ denote the set of such actions.

F.2 The Obfuscator's Problem

With arbitrary payoff matrices, the Obfuscator's indirect expected utility at a belief μ is

$$V(\mu) \coloneqq \max_{a \in a^*(\mu)} \mathbb{E}_\mu[v(a, \omega)].$$

The maximization over $a^*(\mu)$ reflects the tie-breaking rule. If there are multiple optimal distributions of posteriors, the Learner also chooses the one favored by the Obfuscator. Thus, the Obfuscator's expected utility from choosing λ is

$$\tilde{V}_{\mu_0}(\lambda) \coloneqq \max_{\tau \in T_{\mu_0}(\lambda)} \mathbb{E}_\tau[V(\mu)].$$

We use the prior as an index to highlight that, because of Bayes plausibility, it affects the Obfuscator's choice. For a selected solution τ_λ, we can write this payoff as

$$\tilde{V}_{\mu_0}(\lambda) = \sum_{a_i \in A(\tau_\lambda)} P_\lambda(a_i) \sum_{m=1}^{M} \mu^{a_i}(\omega_m) v(a_i, \omega_m). \tag{F.4}$$

When the Obfuscator has state-independent preferences, this expression simplifies to $\tilde{V}_{\mu_0}(\lambda) = \sum_{i=1}^{N} P_\lambda(a_i) v(a_i)$. Intuitively, the Obfuscator chooses the cost that induces the most favorable distribution over actions. In the binary analysis in Section 4, this amounts to maximizing the probability that his favorite action is played.

At this point, the assumption that $a^*(\mu_0)$ is a singleton comes into play. By Proposition 2, there exists a finite $\overline{\lambda}$ such that the Learner's unique optimal choice is to acquire no information for every $\lambda \geqslant \overline{\lambda}$. The Obfuscator's payoff is then constant and equal to $V(\mu_0)$. We can therefore restrict his choice set, without loss of generality, to $\mathcal{O} \coloneqq [0, \overline{\lambda}]$.

The Obfuscator solves

$$\max_{\lambda \in [0,\overline{\lambda}]} \tilde{V}_{\mu_0}(\lambda). \tag{F.5}$$

The conditions ensuring that the Obfuscator's Problem has a solution are unchanged. With a unique optimal action at the prior, Proposition 2 provides a finite bound on the cost needed to induce no learning. Breaking ties in the Obfuscator's favor then gives the upper semicontinuity used in Proposition 3. Its proof applies to this broader environment: it does not rely on matching payoffs, equal numbers of states and actions, or state-independent preferences for the Obfuscator.

F.3 Transparency and Obfuscation Benchmarks

This broader environment admits three useful benchmark cases for transparency and obfuscation. Full transparency corresponds to $\lambda = 0$, while full obfuscation corresponds to no learning.

Aligned Obfuscator: $\arg\max_{a_i \in A} u(a_i, \omega_m) \subseteq \arg\max_{a_i \in A} v(a_i, \omega_m)$ for all $\omega_m \in \Omega$. With all information available, the Learner chooses an Obfuscator-optimal action in every state. Hence, $\lambda = 0$ maximizes the Obfuscator's expected utility and we have *full transparency*. Strictly profitable obfuscation relative to transparency therefore requires a state in which some Learner-optimal action is suboptimal for the Obfuscator.

Totally misaligned Obfuscator: $u(a_i, \omega_m) + v(a_i, \omega_m) = c$ for all a_i, ω_m, where c is a constant. Comparing the Learner's optimal experiments at two costs shows that the experiment chosen at the higher cost uses weakly less information. The experiment chosen at the lower cost must therefore yield at least as high a material payoff, since the Learner prefers it even when information is cheaper. The Obfuscator's payoff, being c minus the Learner's expected material payoff, is weakly increasing. Any cost that induces no learning is therefore optimal whenever such a cost exists.

Agreement at the Prior: the Obfuscator has state-independent preferences ($v(a_i, \omega_m) = v(a_i)$ for all $m = 1, \ldots, M$) and $a^*(\mu_0) \subseteq \arg\max_{a_i \in A} v(a_i)$. Then $\lambda = \overline{\lambda}$ is necessarily optimal, and we have a full obfuscation problem. This formalizes the intuition from the product example in the Introduction: the Obfuscator does not want to talk the Learner out of his preferred action. In the binary case, Theorem 1 shows that agreement at the prior is also necessary for full obfuscation.

As a final remark, consider a paternalistic Obfuscator with $v = u$. Lipnowski et al. (2020) show that full disclosure is always optimal with two states when the receiver faces a given cost of processing the disclosed information. With more states, restricting disclosure can help. In our model, choosing $\lambda = 0$ removes the cost of learning, so the Learner can take an ex post optimal action in every state, regardless of the number of states.