Communication, Perception and Strategic Obfuscation

Geoffroy de Clippel* Kareen Rozen*

Revised February 2021

Abstract

We study the empirical content of simple Sender-Receiver games in which disclosures are mandatory but may be obfuscated. We focus on the fungibility between strategic inference and costly perception, developing a stylized theoretical framework that highlights this channel. Our framework yields crisp testable implications for equilibrium play, and naturally lends itself to an experimental design. Our laboratory results show that a large majority of Senders strategically obfuscate; and an aggregate analysis of Receiver's stochastic-choice data suggests Receivers adjust their perception in response to strategic inference.

*Department of Economics, Brown University. This material is based upon work supported by the National Science Foundation under Grant No. 1559398 (*Communication, Perception and Strategic Obfuscation*, from June 2016 to September 2019). We are grateful to seminar audiences, Pedro dal Bó, Mark Dean, Brian Knight, Henrique Roscoe De Oliveira, and Ran Spiegler for valuable comments and suggestions, and to Tommaso Coen and Zeky Murra for research assistance.

1 Introduction

Communication is a key component of many interactions. Pharmaceutical companies advertise medications to consumers; job seekers describe their qualifications to employers; attorneys provide evidence to defend clients against litigants; researchers describe their studies to potential participants; food manufacturers report ingredients to shoppers. These examples share three common features. First, the goals of the informed party and uninformed decision-maker need not align. Second, regulations (or serious repercussions) force the informed party to truthfully disclose. Finally, without being dishonest, the informed party can attempt to mitigate detrimental information by making it harder to understand (e.g., job candidates dress up resumes, attorneys submit entire hard drives into evidence instead of just incriminating files). Concerns about strategic obfuscation have been raised in some of these settings, with the challenge that obfuscation is in the eye of the beholder. The Belmont report (1978), establishing ethical principles for human-subjects research, warns: "presenting information in a disorganized and rapid fashion, allowing too little time for consideration or curtailing opportunities for questioning, all may adversely affect a subject's ability to make an informed choice." The FDA considers presentational choices (e.g., typesetting) when deciding whether an advertisement provides a "fair balance" of risks and benefits.[1] Even the federal law mandating disclosure of GMOs spurred controversy by requiring food packaging to *either* clearly label GMOs *or* include a QR code linking to that information (the food industry backed the latter).[2]

If obfuscation occurs only when information is detrimental, then witnessing obfuscation is itself informative, and impacts a rational agent's efforts to decipher further. Hence strategic sophistication may substitute for costly perception. We introduce and analyze a stylized, easily-interpretable communication game highlighting this tradeoff, present experimental data to document and empirically assess the tradeoff's relevance, and provide welfare implications.

There are two possible actions and two equally-likely states. Receiver gets a fixed benefit for taking the action matching the state, while Sender benefits whenever one

[1] Does the law say anything about the design of ads for prescription drugs?

[2] Proponents touted Public Law 114-216 for mandating disclosure, but advocacy groups derided it as the "DARK Act," for Deny Americans the Right to Know. See, for instance, the WSJ article Consumer Advocates Wary of Digitally Coded Food Labels, and the Huffington Post entry Obama Expands Monsanto Doctrine By Signing DARK Act And Invalidating Vermont GMO Labeling Law.

particular action is taken. Messages can be *transparent* or *obscure*. Both fully reveal the state, but a transparent message does so immediately, while an obscure message requires costly effort to understand. Sender can condition his preferred message type on the state, but his communication goal may be imperfectly realized. Consider, for instance, an article under review. The authors may find their writing transparent, but referees may disagree; and conversely, referees may identify issues despite the authors' attempts at obfuscation. We encapsulate the potential for such disagreement in the *precision of communication* (p): the probability a Sender who aims to send an obfuscated message (or aims to send a transparent one) in a given state achieves his goal.

Following Caplin and Dean (2015), deciphering is modeled as a costly task whose cost is unknown to the modeler. Embedded in a game-theoretic framework, a novel feature is that Receiver's beliefs about the state after receiving an obscure message, *but before exerting effort to decipher it,* depend on expectations regarding Sender. In an undominated Bayesian-Nash equilibrium, a strategically-sophisticated Sender aims to obfuscate when his favored action is worst for Receiver. In equilibrium, as precision of communication increases, Receiver becomes more convinced that any observed obfuscation is intentional, and thus more skeptical of Sender's favored action. These beliefs inform his effort in deciphering the message.

Receiver's strategic inference and ensuing effort choice manifest in the probabilities, conditional on each state, that he chooses correctly despite obfuscated information. For instance, if he eschews effort entirely and simply takes Sender's worst action, then Receiver is always correct in the opposing-interest state, but always wrong in the common-interest state. If he uses some costly-perception strategy, his success in the common-interest state may increase at the expense of the opposing-interest state. We derive testable implications on Receiver's state-contingent stochastic-choice data for it to be consistent with a rational Receiver's equilibrium behavior. Collecting such data at the individual level is generally hard, and more so in our setting when p is large (obfuscated messages become rare in the common-interests state). It is important, then, to note our results extend to aggregate data in situations where Senders and Receivers are randomly matched and have potentially heterogenous perceptual costs, as in a laboratory experiment.

Our experiment has three treatments: low, medium and high precision of communication, with $p \in \{51\%, 70\%, 90\%\}$. To bring obscure and transparent messages to

life, we import the novel 'colored balls'-design of Dean and Nelighz (2019),[3] who experimentally test the rational inattention model through how success rates vary with incentives. They represent the state (Red/Blue) by a 10x10 matrix of randomly-arranged red and blue balls, where exactly 51 balls color-match the state. We use their construction for obfuscated messages. For transparent messages, we use those balls arranged neatly by color. Thus messages differ only in clarity, not substance.

The data analysis in Section 3.2 substantiates the overall strategic sophistication and rationality of Senders and Receivers. We find 77% of Senders *strategically* obfuscate, aiming for clarity in the common-interests state and obfuscation in the opposing-interests state. Receivers' success varies with p, showing strategic inference can impact attention in games. We find evidence of optimal perception adjustment, as the testable implications on aggregate stochastic choice data are satisfied (or nearly satisfied, in one instance).

Given our evidence that average Receivers adjust perception with strategic inference, Section 4 concludes by highlighting welfare implications. First, a naive regulator (who does not recognize obfuscated messages carry information beyond immediate content) underestimates Receiver's welfare gain from mandating information disclosure. Second, and perhaps counterintuitively, greater alignment of preferences between Sender and Receiver does not guarantee greater Receiver success.

Related work on communication

Communication games are traditionally studied in one of two extreme settings. Information is soft in Crawford and Sobel (1982) cheap-talk setting: messages need not bear any verifiable relation to the truth, but could have meaning in equilibrium. At the other extreme is hard information, see Grossman (1981); Milgrom and Roberts (1986): messages are immediately verifiable, and their absence can lead to information unraveling.

Dewatripont and Tirole (2005) study intermediate situations, modeling communication as a moral-hazard-in-teams problem: Sender and Receiver have increasing, convex effort costs, and Receiver assimilates Sender's information with probability xy when Sender (Receiver) exerts effort $x \in [0, 1]$ (resp., y).[4] Effort choices are simul-

[3]Also appearing in an earlier working paper which Dean and Nelighz (2019) subsumes.

[4]Persson (2018) considers a multi-dimensional extension where the uninformed expert uses information overload as a manipulation device.

taneous in their main analysis, though strategic inferences are discussed in a couple dynamic extensions. Our works share the broad features that Receiver's understanding of messages is probabilistic, and Sender's choice impacts this distribution. But the focus and methodology are different. To crisply highlight the interplay between strategic inferences and costly perception, we abstract from Sender costs and study the case where Receiver freely distinguishes between transparent and non-transparent information. We derive testable implications, valid independently of the underlying cost function, and empirically evaluate our predictions. For that, we build on recent literature studying the empirical content of optimal attention in individual decision-making problems (Caplin and Dean, 2015; Caplin and Martin, 2015; de Oliveira et al., 2017; Ellis, 2018; Dean and Nelighz, 2019).

This paper contributes to the emerging theoretical literature on attention in games. Some works consider firms facing consumers whose perception is exogenously given, as in Gabaix and Laibson (2006) and Bordalo et al. (2015); or who can optimally allocate a fixed total effort among different dimensions, as in Spiegler (2006) and de Clippel et al. (2014). Others consider players who endogenously choose perceptual efforts at a cost, often modeled using the Shannon mutual-information function applied in Sims (2003); see Matějka (2015), among others. But strategic inference (and a fortiori its fungibility with optimal attention) is most often absent. An exception is Martin (2016b), who considers a firm's strategic pricing when facing a consumer who is rationally inattentive about product quality. He finds a mixed-strategy equilibrium in which the high-quality seller sets a high price, while the low-quality seller randomizes between low and high prices. The buyer's attention responds to the seller's equilibrium pricing strategy, using Sims' linear parametrization of attention costs.

In a companion paper, Martin (2016a) experimentally illustrates and calibrates the above model. A seller owns a product of low or high value to a randomly-matched buyer. Knowing the buyer's value, he chooses between a low-price and a high-price offer. Low-price offers are always profitable to the buyer, but high-price offers are only profitable for high-value products. Not knowing his value, the buyer can examine a string of twenty randomly generated numbers (between -100 and 100) whose sum is the true value. Using time responses and frequency of purchasing mistakes, Martin provides supporting evidence that selling price affects the attention buyers pay to learn product value. Focusing on a rationally-inattentive representative buyer, the best approximation of buyers' average behavior is obtained with a marginal-attention

cost of 11.9. Interestingly, explaining sellers' average behavior using the equilibrium in the paragraph above leads to a comparable estimate for the sellers' belief regarding buyers' marginal-attention cost.

Our paper differs from Martin (2016a,b) on multiple dimensions. First, our Sender decides whether he tries to obfuscate information, while in Martin's framework the seller chooses a price and information is always obfuscated. Second, we characterize testable implications: properties on observables remaining valid whatever participants' utilities and attention costs. Receiver is not required, for instance, to process information using Sims' model of rational inattention. Third, testable implications are derived while allowing the precision parameter to vary, permitting cross-observation tests of consistency with equilibrium play. In addition to the more stringent testable implications of equilibrium play, simply witnessing that success rates at guessing the state vary with the precision level provides evidence in a clean treatment-control design that Receivers adjust their attention based on strategic inference.

The experimental literature examines many aspects of communication. Blume et al. (2020) surveys a large literature on cheap-talk experiments. A smaller literature studies information unraveling with hard information; see Jin et al. (2016) and references therein. Fréchette et al. (2019) explores an umbrella framework nesting cheap talk, hard information, and Bayesian persuasion. They relax the commitment assumption in Bayesian persuasion through a probability senders can revise choices, but do not consider obfuscated information. Jin et al. (2019) studies obfuscation, albeit with a different goal. Their Senders know the state $s \in \{1, 2, \ldots, 10\}$ and send a string with $c \in \{1, 2, \ldots, 20\}$ numbers whose sum equals the state. The Receiver has 60 seconds to guess the sum (else a random guess is made), and is paid for accuracy. Sender's payoff increases in the guess. While information should unravel like with voluntary disclosure, Jin et al. (2019) provides experimental evidence that it doesn't and advances possible explanations. By contrast, unraveling does not occur at equilibrium in our model: quite realistically, perceiving a message as obfuscated is informative about Sender's intention, but does not reveal for sure that Sender aimed to obfuscate. This allows us to focus on our main question of interest, the testable implications for the substitutability between optimal perception and strategic inference.

2 Theoretical benchmark

Consider a communication game with a prior π over two states, ω_1 and ω_2. Receiver has two possible actions, $A = \{a_1, a_2\}$, and prefers a_i in state ω_i. By contrast, Sender strictly prefers a_2 in all states, and wishes to persuade Receiver to pick it; hence we refer to their interaction as a persuasion game. Sender and Receiver are expected utility maximizers. For simplicity, we assume Receiver gets $\$m_R$ when his action matches the state, and nothing otherwise; while Sender gets $\$m_S$ when Receiver chooses a_2, and nothing otherwise. Thus players' interests are opposed in ω_1 and common in ω_2.

Knowing the state, Sender must communicate it to Receiver, but need not make this information easily understood. Sender can aim to *communicate clearly* or aim to *obfuscate*. The *precision level* $p \in (1/2, 1)$ calibrates how likely Sender's communication goal is achieved. If Sender aims to communicate clearly in state ω, then the resulting message to Receiver is *transparent* (denoted $T(\omega)$) with probability p, and *obscure* (denoted $O(\omega)$) otherwise. Oppositely, if Sender aims to obfuscate, then the message is obscure with probability p and transparent otherwise. The precision p is a parameter we vary across experimental treatments.

With two states, there are four message types. With a transparent message $T(\omega)$, ω is revealed at once. With an obscure message, Receiver's only way to distinguish $O(\omega_1)$ from $O(\omega_2)$ is to exert effort to decipher the message. In line with recent models of optimal attention in decision theory, he chooses a perception strategy $(\mathcal{S}, \mu)$, where $\mathcal{S}$ is a finite set whose elements $s \in \mathcal{S}$ are called *signals* and $\mu(s|\omega)$ is the probability of signal s in state ω. He also chooses a *decision rule* specifying an action for each signal, or equivalently the set $\mathcal{S}_1 \subseteq \mathcal{S}$ of signals resulting in action a_1 (with a_2 picked for any $s \in \mathcal{S}_2 = \mathcal{S} \setminus \mathcal{S}_1$). The choice of perception strategy and decision rule entails a cost $c_R(\mathcal{S}_1, \mathcal{S}, \mu)$, on which we impose no restriction aside from it being subtracted from the expected utility of earnings.

2.1 Equilibrium Conditions

Our equilibrium notion is *undominated Bayesian-Nash equilibrium*. Consider Sender's optimization problem first. Receiver's equilibrium perception strategy and decision rule define a *success probability* $\ell(\omega_i)$ for $i = 1, 2$, the probability of guessing state ω_i correctly following an obfuscated message. With $\tau(\omega)$ denoting the probability

Sender aims to communicate clearly in state ω, Sender's equilibrium conditions are

$$\begin{cases} \tau(\omega_1) \in \arg\max_{x\in[0,1]} \{u_S(m_S) - \nu(x, \ell(\omega_1))[u_S(m_S) - u_S(0)]\} \\ \tau(\omega_2) \in \arg\max_{x\in[0,1]} \{u_S(0) + \nu(x, \ell(\omega_2))[u_S(m_S) - u_S(0)]\}, \end{cases} \tag{1}$$

where $\nu(x, y) = [x(p + (1-p)y) + (1-x)(py + 1 - p)]$ is the unconditional probability Receiver correctly guesses the state, if Sender aims to communicate clearly with probability x and Receiver guesses correctly with probability y (1) following an obscure (respectively, transparent) message. Clearly, Sender's payoff is decreasing in ν in the opposing-interests state (top equation), and increasing in ν in the common-interests state (bottom equation).

Notice that Receiver's beliefs about the state after receiving an obscure message, *but before exerting effort to decipher it*, are:

$$\hat{\pi}(\omega) = \frac{\pi(\omega)\left(\tau(\omega)(1-p) + (1-\tau(\omega))p\right)}{\sum_{i=1,2} \pi(\omega_i)\left(\tau(\omega_i)(1-p) + (1-\tau(\omega_i))p\right)}. \tag{2}$$

Her perception strategy and decision rule maximize

$$u_R(0) + \sum_{i=1,2} \hat{\pi}(\omega_i)\mu(\mathcal{S}_i|\omega_i)(u_R(m_R) - u_R(0)) - c_R(\mathcal{S}_1, \mathcal{S}, \mu) \tag{3}$$

under the constraint that she prefers action a_i following $\sigma \in \mathcal{S}_i$. This means Receiver's posterior probability

$$\hat{\mu}(\omega|\sigma) = \frac{\mu(\sigma|\omega)\hat{\pi}(\omega)}{\sum_{\omega'\in\Omega} \mu(\sigma|\omega')\hat{\pi}(\omega')}$$

for state ω, conditional on getting signal σ from an obscure message, satisfies

$$\hat{\mu}(\omega_i|\sigma) \geq \hat{\mu}(\omega_{-i}|\sigma), \tag{4}$$

for all $\sigma \in \mathcal{S}_i$ and for all $i = 1, 2$.

2.2 Observables and Equilibrium Consistency

Consider repeated observations from several persuasion games differing only in the precision level p. Of course, perception strategies and decision rules are not observable. We provide testable implications on a *dataset* $\{(p^j; \tau^j, \ell^j)|j = 1, \ldots, J\}$,

where p^j is the precision level in persuasion game j, $\tau^j = (\tau^j(\omega_1), \tau^j(\omega_2))$ specifies the state-dependent probability Sender aims to communicate clearly in game j, and $\ell^j = (\ell^j(\omega_1), \ell^j(\omega_2))$ specifies the state-dependent probability Receiver chooses the correct action in game j.

The dataset is *consistent with equilibrium play* if there exist utility functions u_S, u_R, a perception-cost function c_R, and for each j a perception strategy and decision rule $(\mathcal{S}_1^j, \mathcal{S}^j, \mu^j)$ that combine with τ^j to form an equilibrium of the Sender-Receiver game given by p^j, such that $\ell^j(\omega_i) = \mu^j(\mathcal{S}_i|\omega_i)$ for $i = 1, 2$. This amounts to a revealed-preference exercise in a game with state-dependent stochastic-choice data. In other words, we are interested in finding all predictions of our communication games that remain valid whatever the utility and perception-cost functions. While we focused on an equilibrium notion, the testable implications we derive remain valid under alternative assumptions on participants' expectations.

2.3 Testable Implications

We are now ready to state the main theoretical result.

Proposition 1. *The dataset is consistent with equilibrium play if, and only if, all the following conditions hold:*

(i) For each persuasion game j, Sender aims to obfuscate in the opposing-interests state and aims to communicate clearly in the common-interests state: $\tau^j(\omega_1) = 0$ and $\tau^j(\omega_2) = 1$;

(ii) Receiver's belief upon receipt of an obfuscated message in game j is:

$$\hat{\pi}^j(\omega_1) = \frac{p^j \pi(\omega_1)}{p^j \pi(\omega_1) + (1 - p^j)\pi(\omega_2)};$$

(iii) In each persuasion game j, Receiver's expected success rate upon receipt of an obfuscated message is at least as high as the expected success rate from choosing action a_1:

$$\hat{\pi}^j(\omega_1)\ell^j(\omega_1) + (1 - \hat{\pi}^j(\omega_1))\ell^j(\omega_2) \geq \hat{\pi}^j(\omega_1);$$

(iv) Excess success rates are monotone: for any pair of persuasion games with $\hat{\pi}^j(\omega_1) > \hat{\pi}^k(\omega_1)$ (equivalently $p^j > p^k$, given (ii)) we have $\ell^j(\omega_1) - \ell^j(\omega_2) \geq \ell^k(\omega_1) - \ell^k(\omega_2)$.

The proof appears in the Appendix. We now provide some intuition and discuss the robustness of our testable implications against other belief assumptions. Receiver's strategy pins down success probabilities $\ell(\omega_i)$ for $i = 1, 2$. Whatever they are, Sender's payoff in (1) decreases (increases) in x in the opposing-interests (common-interests) state, and strictly so if Receiver fails to perfectly guess the state. Thus, beyond equilibrium play, property (i) is robust to alternative specifications of Sender's expectations. Aiming to obfuscate (communicate clearly) given ω_1 (ω_2) is his unique weakly dominant strategy, and unique best response with success rates strictly below one.

Updated beliefs in (ii) arise from Bayes' rule (see (2)), given Sender's equilibrium strategy in (i). We use this belief for our benchmark data analysis, but also explore alternate specifications. For instance, subjects may fail to update probabilities accurately, fail to recognize the strategy in (i) is dominant, or suspect Sender does not recognize that strategy is dominant.

Though part of a game, once Receiver's beliefs are fixed, her choices can be studied as an individual problem of optimal perception reminiscent of Caplin and Dean (2015). A difference is that they consider payoff changes, keeping states' probabilities unchanged. The very nature of our analysis entails (endogenous) changes in probabilities instead. This difference can be dealt with through mathematical transformation: probabilities premultiply utilities, so probability changes can be reinterpreted as specially-structured payoff changes. With this in mind, condition (iii) corresponds to Caplin and Dean's NIAS: picking a_1 cannot generate higher payoffs, as Receiver would be better off not exerting any effort. Condition (iv) relates to their NIAC condition, which says total utility cannot increase by reassigning attention across any cycle of decision problems (of any length). Given our structure–probabilities vary, not payoffs–we show it suffices to consider pairwise cycles, or that excess-success rates $\ell^p(\omega_1) - \ell^p(\omega_2)$ weakly increase in the precision p.

Conditions in (iv) apply to any Receiver beliefs $(\hat{\pi}^j)_{j=1}^J$, even if they violate (ii). When comparing excess-success rates in games j and k, what determines the direction of inequality to check is whether $\hat{\pi}^j(\omega_1)$ or $\hat{\pi}^k(\omega_1)$ is larger, not their cardinal values. Thus (iv) is robust to other belief specifications provided that $\hat{\pi}^j(\omega_1)$ increases in the precision p^j, which one generally expects since Sender has no incentive to obfuscate in the common-interests state ω_2.

2.4 A population of Senders and Receivers

The above presumes a single Sender and Receiver. Suppose Senders and Receivers are drawn uniformly at random from a population. All players strictly prefer more money, but may have different utility functions and perception costs.

This is a setting of practical relevance for at least two reasons. First, a Sender may expect to face a population of Receivers with different perceptual costs (e.g., heterogenous consumers). Second, it can be time-consuming for subjects in experiments to generate individual-level, state-dependent stochastic-choice data. This is especially so in our experiment, where messages are generated endogenously, and we would expect to see few obfuscated ones in the common-interests state when p is high. Fortunately, Proposition 1 extends to characterize testable implications in this setting. Regarding Senders, the argument for (i) immediately shows the same strategy remains weakly dominant (strictly so if there is any chance of mistake following obfuscated messages). Hence Receivers share the same belief (ii) as before, upon seeing an obfuscated message but before attempting to decipher it.

Proposition 1 holds for *every* Receiver. Hence they each satisfy $(iii) - (iv)$, which are linear in individual success rates. Let $\bar{\ell}^j(\omega)$ be the average success rate of Receivers in game j and state ω. In our between-subject analysis, different sets of subjects play our communication game with different precisions levels. We can reasonably assume, as is standard, that different treatments draw from the same population of characteristics. Summing (iii) and (iv) over all Receivers in each treatment implies:

$$\hat{\pi}^j(\omega_1)\bar{\ell}_i^j(\omega_1) + (1 - \hat{\pi}^j(\omega_1))\bar{\ell}_i^j(\omega_2) \geq \hat{\pi}^j(\omega_1), \text{ for all } j.$$
$$\bar{\ell}_i^j(\omega_1) - \bar{\ell}_i^j(\omega_2) \geq \bar{\ell}_i^k(\omega_1) - \bar{\ell}_i^k(\omega_2), \text{ when } \hat{\pi}^j(\omega_1) > \hat{\pi}^k(\omega_1).$$

These necessary conditions are the same as Proposition $1(iii) - (iv)$, but using *average* success rates. Conversely, if average success rates satisfy these conditions, then Proposition 1 implies the data can be explained by a hypothetical population of Receivers identical in utility functions and perception costs, even if they truly are heterogenous. Thus Proposition 1, applied to population averages, remains necessary and sufficient for consistency with equilibrium play in this more general setting.

3 Experiment

3.1 Design

We first describe the implementation of the communication game, and then describe how subjects are matched.

There are two equally-likely keywords, Blue and Red. Given the keyword, Receiver gets one of two possible message types. Messages always have 100 balls of blue and red color arranged in a 10x10 matrix, with exactly 51 balls whose color matches the keyword and 49 balls whose color mismatches it. However, messages differ in how balls are arranged. In a transparent message, the balls are arranged by color; such a message immediately reveals the majority color. In an obfuscated message, the same balls are randomly placed into the 10x10 matrix, and it takes effort to garner information on the majority color.[5] Both message types reveal the true keyword through the majority color. Hence Sender cannot lie, but can obfuscate.

Receiver's message is always one of these two types, with the probability of each type chosen by Sender. For each possible keyword (Blue and Red), Sender is asked to choose whether they prefer it to be more likely (with probability $p > 1/2$) that the balls will be arranged by color or more likely (with the same probability p) that the balls will be in random order. Thus Sender makes a contingent messaging plan, tailored to each keyword. We consider three treatments, corresponding to $p \in \{51\%, 70\%, 90\%\}$. The value of p is constant within each session, to avoid pollution across precision levels and ensure sessions are not unreasonably long.

In each Sender-Receiver matchup, the computer draws the keyword (Blue or Red) uniformly at random. Receiver's message is drawn, given p, based on Sender's preference for that keyword. Receiver is asked to guess the keyword, and permitted unlimited time to examine their message before making this choice. Sender's payoff from the matchup is \$15 if Receiver guesses Red, and \$0 otherwise. The Receiver's payoff from the matchup is \$15 for guessing the keyword correctly, and \$0 otherwise.

The timing of decisions and matching process are as follows. Each session has two phases. In the Sending phase, each subject acts as Sender and selects a contingent messaging plan. Decisions from the Sending phase determine messages in the Receiving phase. In this second phase, each subject acts as Receiver and is matched

[5]For an example of these messages, see the instructions in the Online Appendix.

forty times. Each match is with an independently and uniformly drawn Sender other than themselves. In each match, the computer independently and uniformly draws the keyword and implements the matched Sender's decision for that keyword. That is, the computer uses the relative likelihood the Sender chose for that keyword, to display a message with balls arranged either by color or randomly. The Receiver has an unlimited amount of time to examine this message before guessing the keyword.

The experiment is designed so that if Senders make communication decisions rationally, the probability Receiver gets an obfuscated screen in a match is independent of p (indeed $\pi(\text{Red})(1-p) + \pi(\text{Blue})p = \frac{1}{2}(1-p+p) = \frac{1}{2}$). Hence on average, there is an equal burden for Receivers across treatments. What changes across treatments is the distribution of keywords *conditional* on seeing an obfuscated screen.

No feedback is provided at any point. The experiment concludes after the Receiving phase. Subjects are given an optional exit survey. The computer determines each subject's payoff by randomly picking a role (Sender or Receiver) and a match in which the subject played that role. Each subject receives their payoff from that match plus the \$10 show-up fee. Subjects are not told choices others made or payoffs others received.

There were six sessions, two for each $p \in \{51\%, 70\%, 90\%\}$. 131 subjects participated, with 42 in the 51% treatment, 44 in the 70% treatment, and 45 in the 90% treatment. Sessions were conducted at BUSSEL, the Brown University Social Sciences Experimental Laboratory, in April and May 2018. Subjects were paid their earnings in cash before leaving the laboratory.

3.2 Results

A rational, self-interested Sender would only aim to obfuscate in the opposing-interests state (Proposition 1(i)). The vast majority of Senders are in line with this prediction: 71.4%, 86.4%, and 73.3%, for the $51\%, 70\%$, and 90% treatments, respectively. For each treatment, a binomial test rejects (at all significance levels) equality with the 25% random-choice benchmark. The most common deviation from Proposition 1(i) is to always communicate clearly (21.4%, 9.1% and 17.8%, respectively), which is consistent, for instance, with altruistically easing Receivers' perceptual burden.

Consider now subjects' choices as Receivers. With equally-likely states, Receivers' equilibrium belief on ω_1 is precisely the treatment's p. To test Proposition 1(*iii*) –

(iv), we estimate Receivers' aggregate success probability $\ell^p(\omega)$ following obfuscated messages, for each state ω and treatment p. There are 2,446 observations of Receivers' guesses for obfuscated messages, with 805 observations from the 51% treatment, 835 observations from the 70% treatment, and 806 observations from the 90% treatment.

To estimate success probabilities, define the following dummy variables: $Correct_i$ indicates whether the Receiver in observation i guessed correctly, Red_i^p indicates whether the observation is from treatment p *and* the keyword was Red, and $Blue_i^p$ indicates whether the observation is from treatment p *and* the keyword was Blue. Let the vector of explanatory variables be $X = (Blue^{51}, Red^{70}, Blue^{70}, Red^{90}, Blue^{90})$. Interacting the treatment and keyword, we estimate success probabilities through the logistic regression:

$$\ln\left(\frac{\mathrm{P}(Correct=1|X)}{\mathrm{P}(Correct=0|X)}\right) = \alpha_0 + \alpha_B^{51} Blue^{51} + \alpha_R^{70} Red^{70} + \alpha_B^{70} Blue^{70} + \alpha_R^{90} Red^{90} + \alpha_B^{90} Blue^{90},$$

and use heteroscedasticity-robust errors clustered at the individual level. Recalling that the opposing-interests state ω_1 is the Blue keyword, we find:

Precision p	$\ell^p(\omega_1)$	$\ell^p(\omega_2)$	$\ell^p(\omega_1) - \ell^p(\omega_2)$
51%	0.837	0.783	0.054
70%	0.878	0.669	0.208
90%	0.927	0.590	0.337

Table 1: Estimated state-dependent success probabilities of Receivers per treatment, and excess-success probabilities, rounded to three decimal places.

These success probabilities pertain to Receiver's choices following obfuscated messages. Our framework presumes Receivers choose the correct action following transparent messages. Out of the 2,794 transparent messages our Receivers faced over three treatments, there were only 12 instances of a Receiver clicking on the wrong keyword, a 0.0043 probability of failure.

As for obfuscated messages, are Receivers sophisticated about their meaning? Notice the only difference between treatments is the precision level associated with Sender's choice. If Receivers were strategically unsophisticated, then receiving an obfuscated message would have no impact on beliefs and success rates for ω_1 and ω_2 would each be independent of p. To the contrary, the joint null hypothesis $\ell^{51}(\omega_1) =$

$\ell^{70}(\omega_1) = \ell^{90}(\omega_1)$ and $\ell^{51}(\omega_2) = \ell^{70}(\omega_2) = \ell^{90}(\omega_2)$ is rejected (p-value 0.0247). This demonstrates Receivers exhibit some strategic sophistication, but does not yet imply consistency with equilibrium play.

With equally-likely states, condition (ii) in Proposition 1 means $\hat{\pi}^p(\omega_1) = p$. Conditions (iii) and (iv) are tested through weak linear inequalities over estimated success probabilities. Condition (iv) requires estimated excess-success probabilities to weakly increase in p. This is confirmed by Table 1. Going beyond the testable implications, we can ask whether the inequality's slack is also significantly different from zero. A two-sided test of the null LHS-RHS=0 find the increase in excess-success rate between the 51% and 90% treatments is statistically significant (p-value 0.0028). The stepwise increases, from 51% to 70%, and from 70% to 90%, do not reach significance at the 5% level (p-values 0.0885 and 0.2007, respectively).

Condition (iii) requires $\ell^p(\omega_2) + \hat{\pi}^p(\omega_1)(\ell^p(\omega_1) - \ell^p(\omega_2)) - \hat{\pi}^p(\omega_1) \geq 0$, which represents how much additional success probability is attained beyond guessing Blue after each obfuscated message, given $\hat{\pi}^p(\omega_1) = p$. The point estimates strictly satisfy these inequalities for the 51% and 70% treatments. Again going beyond the testable implications, the slack of 0.300 for the 51% treatment and 0.115 for the 70% treatment are significantly different from zero (both p-values 0.0000). The inequality is violated for the 90% treatment by 0.007, which is not significantly different from zero (p-value 0.6827). Of course, for fixed success rates, condition (iii) is more demanding as the belief on ω_1 increases: at some point, it may be optimal to simply pick a_1. The precision p is the equilibrium belief, but not all subjects obfuscated in ω_1. In fact, condition (iii) would hold strictly (and with statistical significance) under rational-expectations beliefs, and condition (iv) would be unchanged, since the rational-expectations beliefs increase in p.[6]

4 Discussion

We presented evidence that strategic inferences from obfuscation are used to adjust perceptual choices. To conclude, we point to some broader welfare implications.

Suppose Senders would not voluntarily disclose information, as Receiver takes

[6]Isolating p, condition (iv) amounts to $\hat{\pi}^p(\omega_1) \leq \ell^p(\omega_2)/(1 + \ell^p(\omega_2) - \ell^p(\omega_2))$. The 95% CI for this ratio in the 90% treatment is $(0.838, 0.941)$. The point estimates strictly satisfy (iv) if $\hat{\pi}^{90}(\omega_1) \leq 0.889$, and the null of equality is rejected for $\hat{\pi}^{90}(\omega_1) \leq 0.838$. With rational expectations, $\hat{\pi}^{90}_{RE}(\omega_1) = 0.8122$ ($\hat{\pi}^{51}_{RE}(\omega_1) = 0.5069$ and $\hat{\pi}^{70}_{RE}(\omega_1) = 0.6729$).

Sender's preferred action absent communication ($\pi(\omega_2) > 1/2$). What are the welfare implications of mandating disclosure, if obfuscation cannot be prevented? A naive regulator might only consider the immediate informational content of obfuscated messages. Suppose he assesses a consumer facing a complex product label incurs perception cost c for a $\ell(\omega)$-chance in state ω of correctly guessing whether the product is worth purchasing.[7] Mandating disclosure, he believes, yields the following ex-ante gain for Receiver (setting $u_R(m_R) = 1$ without loss):

$$[\pi(\omega_1)p\left(\ell(\omega_1) - c\right)) + \pi(\omega_2)(1-p)\left(\ell(\omega_2) - c\right)) + \pi(\omega_1)(1-p) + \pi(\omega_2)p - \pi(\omega_2), \quad (5)$$

where the subtracted $\pi(\omega_2)$ represents the success probability in the absence of disclosure (i.e., from choosing a_2). But, as our analysis highlights, obfuscated messages reveal information beyond their immediate content, and average Receivers factor this in. A rational Receiver, using Bayesian-updated beliefs $\hat{\pi}$ following obfuscation, would instead incur perception cost $\hat{c}$ to have a $\hat{\ell}(\omega)$ chance in state ω of correctly guessing whether the product is worth buying. Receiver's true ex-ante welfare gain from mandated disclosure is

$$[\pi(\omega_1)p\left(\hat{\ell}(\omega_1) - \hat{c}\right)) + \pi(\omega_2)(1-p)\left(\hat{\ell}(\omega_2) - \hat{c}\right)) + \pi(\omega_1)(1-p) + \pi(\omega_2)p - \pi(\omega_2), \quad (6)$$

Subtracting (6) from (5) gives:

$$\pi(\omega_1)p\left((\ell(\omega_1) - c) - (\hat{\ell}(\omega_1) - \hat{c})\right) + \pi(\omega_2)(1-p)\left((\ell(\omega_2) - c) - (\hat{\ell}(\omega_1) - \hat{c})\right). \quad (7)$$

By Bayesian updating,

$$\hat{\pi}(\omega_1) = \frac{p\pi(\omega_1)}{p\pi(\omega_1) + (1-p)\pi(\omega_2)}.$$

Dividing (7) by $p\pi(\omega_1) + (1-p)\pi(\omega_2)$, the sign of (7) equals the sign of

$$\hat{\pi}(\omega_1)\left((\ell(\omega_1) - c) - (\hat{\ell}(\omega_1) - \hat{c})\right) + \hat{\pi}(\omega_2)\left((\ell(\omega_2) - c) - (\hat{\ell}(\omega_1) - \hat{c})\right). \quad (8)$$

Since a Receiver with updated beliefs $\hat{\pi}$ prefers the perception strategy yielding

[7] In our context, this amounts to guessing or eliciting success probabilities and perception costs in an decision-making experiment à la Caplin and Dean (2015), with no Sender.

$(\hat{\ell}(\omega_1), \hat{\ell}(\omega_2))$ over one yielding $(\ell(\omega_1), \ell(\omega_2))$,

$$\hat{\pi}(\omega_1)\hat{\ell}(\omega_1) + \hat{\pi}(\omega_2)\hat{\ell}(\omega_2) - \hat{c} \geq \hat{\pi}(\omega_1)\ell(\omega_1) + \hat{\pi}(\omega_2)\ell(\omega_2) - c.$$

Hence (8) is at most zero. We conclude *a naive regulator underestimates Receivers' welfare gain from mandating disclosure; hence overlooking strategic sophistication can lead to misguided policy decisions* when weighing Receiver benefits against welfare implications for Sender and costs of mandating disclosure.

Our analysis focused on a persuasion payoff structure. Keeping Receiver's payoffs unchanged, we could vary the alignment of Sender's and Receiver's interests. They are opposed (common) if Sender's payoff in state ω_2 (resp., ω_1) were instead m_S if Receiver chooses a_1 and 0 otherwise. Sender's dominant strategy is to obfuscate (clarify) when interests are opposed (resp., common). Hence, under each of those payoff structures, Receiver's updated belief conditional on receiving an obfuscated message coincides with her prior π. Because Receiver's choice of perception strategy for an obfuscated message depends only on her beliefs, she chooses the same perception strategy, with success probabilities ℓ, under both opposed- and common-interests payoff structures. Our paper highlights that Receiver's success probabilities $\hat{\ell}$ under persuasion-payoff structures are typically different from ℓ: rational Receivers can optimally adjust their perception strategy given strategic inference. Not recognizing this, one might suspect success probabilities increase as Sender's and Receiver's preferences get more aligned (from opposed interests to persuasion, and from persuasion to common interests).

That intuition turns out to be wrong: *greater alignment of preferences does not guarantee greater Receiver success*. Imagine Receiver's perceptual costs are such that in an individual decision-making setting with the prior $\pi(\omega_2) = 51\%$, she optimally uses a symmetric strategy with a 0.9-success probability in each state. This would then be her optimal perception strategy in games with common or opposed interests, including when precision is $p = 0.8$. In a persuasion game with $p = 0.8$, she prefers to simply choose action a_1. Her success probability is 92% (98%) for ω_1 and ω_2 when interests are opposed (resp. common). For the persuasion-payoff structure, her success probability is 1 in ω_1 and 80% in ω_2. Hence actual success can decrease in ω_2 (ω_1) when moving from opposed interests to persuasion (resp., persuasion to common interests). In fact, even the expected success probability can decrease: $0.51*0.8+0.49$ for the persuasion case is strictly inferior to 0.92 in the case of opposed interests.

Appendix: Proof of Proposition 1

It remains to show (iii) and (iv) capture the empirical content of Receiver's problem given his belief from (ii).

Step 1: Receiver's choices are consistent with costly information acquisition under rational-expectations beliefs if, and only if, they are consistent with costly-information acquisition in a transformed *individual* decision-making problem with *uniform* beliefs about states and state-dependent payoffs. Letting $\Delta u_i^j = 2\hat{\pi}^j(\omega_i)(u_R(m_S) - u_R(0))$, Receiver's objective (3) for persuasion game j is:

$$u_R(0) + \frac{1}{2}\mu(\mathcal{S}_1|\omega_1)\Delta u_1^j + \frac{1}{2}\mu(\mathcal{S}_2|\omega_2)\Delta u_2^j - c_R(\mathcal{S}_1, \mathcal{S}, \mu),$$

and the constraint Receiver prefers action a_i following $\sigma \in \mathcal{S}_i$ is $\mu(\sigma|\omega_i)\Delta u_i^j \geq \mu(\sigma|\omega_{-i})\Delta u_{-i}^j,\ \forall \sigma \in \mathcal{S}_i,\ \forall i = 1, 2$. Thus, the problem in game j is equivalent to one with uniform beliefs over states, after rescaling the payoff gain in state i from choosing correctly to Δu_i^j.

Step 2: Using Caplin and Dean (2015, Theorem 1), consistency in the transformed problem is equivalent to Receiver's data satisfying their NIAC and NIAS conditions. Translated to our setting and notation, and using $p > 1/2$, NIAS corresponds to condition (iii) in Proposition 1, while NIAC corresponds to the condition that for any integer $J \geq 2$ and any J-length cycle $(p_1, p_2, \ldots, p_J, p_1)$ of persuasion games,

$$\sum_{j=1}^{J} \left(\ell^j(\omega_1) - \ell^{j+1}(\omega_1)\right) \Delta u_1^j \geq \sum_{j=1}^{J} \left(\ell^j(\omega_2) - \ell^{j+1}(\omega_2)\right) \Delta u_2^j, \tag{9}$$

where $\ell^{J+1} = \ell^1$.

Step 3: In our setting, (9) reduces to the pairwise condition of Proposition 1(iv). To see this, rewrite (9) as:

$$\sum_{j=1}^{J} \left(\ell^j(\omega_1) - \ell^{j+1}(\omega_1) - \ell^j(\omega_2) + \ell^{j+1}(\omega_2)\right) \hat{\pi}^j(\omega_1) \geq \sum_{j=1}^{J} \left(\ell^j(\omega_2) - \ell^{j+1}(\omega_2)\right) = 0,$$

using $\hat{\pi}^j(\omega_2) = 1 - \hat{\pi}^j(\omega_1)$ and cancelling the factor $2(u_R(m_S) - u_R(0))$ in the Δu_i^j's.[8]

[8]Our result and proof extend when Receiver's benefit from a correct guess is state dependent, by

Letting $\Delta\ell^j = \ell^j(\omega_1) - \ell^j(\omega_2)$, (9) is then equivalent to:

$$\sum_{j=1}^{J} \left(\Delta\ell^j - \Delta\ell^{j+1}\right) \hat{\pi}^j(\omega_1) \geq 0. \tag{10}$$

For $J = 2$ (a cycle (p^j, p^k, p^j)), condition (10) reduces to condition (iv):

$$(\hat{\pi}^j(\omega_1) - \hat{\pi}^k(\omega_1))(\Delta\ell^j - \Delta\ell^k) \geq 0. \tag{11}$$

We prove by induction that if (11) holds for all pairs of persuasion games, then (10) holds for any cycle length $J > 2$. Suppose (10) holds for cycles of length $J-1$, and consider one of length J. We may translate the elements of the cycle so the J-th element corresponds to the highest $\hat{\pi}^j$ (the sum in (10) is invariant to where the cycle begins). Notice $\sum_{j=1}^{J} \left(\Delta\ell^j - \Delta\ell^{j+1}\right) \hat{\pi}^j(\omega_1)$ can be decomposed into:

$$\begin{aligned} &\sum_{j=1}^{J-1} \left(\Delta\ell^j - \Delta\ell^{j(mod(J-1))+1}\right) \hat{\pi}^j(\omega_1) - \left(\Delta\ell^{J-1} - \Delta\ell^1\right) \hat{\pi}^{J-1}(\omega_1) \\ &+ \left(\Delta\ell^{J-1} - \Delta\ell^J\right) \hat{\pi}^{J-1}(\omega_1) + \left(\Delta\ell^J - \Delta\ell^1\right) \hat{\pi}^J(\omega_1). \end{aligned} \tag{12}$$

The first term in (12) corresponds to (10) for the $(J-1)$-length cycle $(p_1, p_2, \ldots, p_{J-1}, p_1)$ omitting p_J; this is nonnegative by the inductive hypothesis. To reconstruct the sum in (10) for the original cycle, the second term removes the link from p^{J-1} to p^1, and the next two terms add back links from p^{J-1} to p^J, and from p^J to p^1. These final three terms in (12) sum to $\left(\hat{\pi}^J(\omega_1) - \hat{\pi}^{J-1}(\omega_1)\right)\left(\Delta\ell^J - \Delta\ell^1\right)$. Given our numbering scheme, $\hat{\pi}^J(\omega_1)$ is maximal among all $\hat{\pi}^j(\omega_1)$, so the first factor in this product is nonnegative. Similarly, the pairwise condition (11) applied to $(1, J)$ ensures $\Delta\ell^J \geq \Delta\ell^1$, so the second factor is nonnegative. Thus (12) is nonnegative, implying (10) holds for the J-length cycle. □

defining $\Delta u_i^j = 2\hat{\pi}^j(\omega_i)(u_R(m_{S,i}) - u_R(0))$ and $\Delta\ell^j = \ell^j(\omega_1)(u_R(m_{S,1}) - u_R(0)) - \ell^j(\omega_2)(u_R(m_{S,2}) - u_R(0))$.

References

Blume, A., E. Lai, and W. Lim (2020). Strategic information transmission: A survey of experiments and theoretical foundations. In C. M. Capra, R. Croson, M. Rigdon, and T. Rosenblat (Eds.), *Handbook of Experimental Game Theory*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.

Bordalo, P., N. Gennaioli, and A. Shleifer (2015). Competition for attention. *The Review of Economic Studies 83*(2), 481–513.

Caplin, A. and M. Dean (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review 105*(7), 2183–2203.

Caplin, A. and D. Martin (2015). A testable theory of imperfect perception. *The Economic Journal 125*(582), 184–202.

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

de Clippel, G., K. Eliaz, and K. Rozen (2014). Competing for consumer inattention. *Journal of Political Economy 122*(6), 1203–1234.

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

Dean, M. and N. Nelighz (2019). Experimental tests of rational inattention. *Working Paper*.

Dewatripont, M. and J. Tirole (2005). Modes of communication. *Journal of Political Economy 113*(6), 1217–1238.

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

Fréchette, G. R., A. Lizzeri, and J. Perego (2019). Rules and commitment in communication: An experimental analysis. *Working paper*.

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.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law & Economics 24*(3), 461–483.

Jin, G. Z., M. Luca, and D. Martin (2016). Is no news (perceived as) bad news? an experimental investigation of information disclosure. *Working Paper*.

Jin, G. Z., M. Luca, and D. Martin (2019). Complex disclosure. *Working Paper*.

Martin, D. (2016a). Rational inattention in games: Experimental evidence. *Working paper*.

Martin, D. (2016b). Strategic pricing with rational inattention to quality. *Games and Economic Behavior 104*, 131–145.

Matějka, F. (2015). Pricing and rationally inattentive consumer. *Journal of Economic Theory 158*.

Milgrom, P. and J. Roberts (1986). Relying on the information of interested parties. *Rand Journal of Economics 17*(1), 18–32.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare (DHEW) (1978). The Belmont Report. *Washington, DC: United States Government Printing Office*.

Persson, P. (2018). Attention manipulation and information overload. *Behavioural Public Policy 2*(1), 78–106.

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

Spiegler, R. (2006). Competition over agents with boundedly rational expectations. *Theoretical Economics 1*(2), 207–231.