Crying Wolf in the Lab

Arya Gaduh, Peter McGee, Alexander Ugarov*

September 14, 2023

Abstract

Abstract is here —-

Keywords: alarms, value of information, information economics, information design, —

1 Introduction

The 2010 gas blowout on Deep Horizon oil rig has killed 11 workers and caused one of the largest oil spills in history. The death toll was possibly aggravated by switching off a general safety alarm because its sirens interfered with workers' sleep.[1] This illustrates the trade-off between false-positive and false-negative test results with false-positive rates leading to higher false alarm costs and false-negative resulting in missed events.

Many real-life situations involve choosing binary tests to discover and prevent a negative outcome. Most binary tests transform continuous signals about the likelihood of an adverse state into simple yes/no prediction. This transformation relies on choosing a threshold for positive classification. Holding a continuous signal constant, a decrease in probability of no alarm in an adverse state (false-negative rate) corresponds to an increase in probability of alarm in a non-adverse state (false-positive rate). This trade-off motivates multiple discussions in medical diagnostics, alarm systems and extreme weather alerts. Despite ubiquity of binary alarms, there is little empirical evidence on how users evaluate alarms with different false-positive and false-negative rates.

In order to understand preferences over these trade-offs, we study the demand for information in the framework with a potential protection action. The subject, first, receives a signal about the probability of an adverse event. Then she decides to protect or not. This environment describes several practically important scenarios including extreme weather alerts, medical testing and safety alarms.

Some recent studies observe that many people put non-zero value on information about ego-relevant beliefs or future utility even if it has no apparent effect on subsequent decisions (all the citations). These preferences is not the focus of our study and hence we use relatively low stakes and ego-neutral information. As a result, our findings might not apply to settings with changing identity beliefs or to settings with delayed resolution of uncertainty and large potential payoffs.

We find that the value of information in our setup weakly correlates with the willingness-to-pay. First, subjects on average underreact to quality of the signal, resulting in overpaying for low-quality signal and underpaying for high-quaity signals. Second, subjects tend to overreact to false-negative rates when the prior probability is low and overreact to false-positive positive rates when priors are high. We show that this pattern is most consistent with failure to estimate the effect of frequencies of false-positive and false-negative outcomes on the costs of using the signal. Xu (2020) similarly finds that individuals(?) do not properly account for priors and often choose tests not affecting optimal decisions even then more instrumental tests are available.

Our work is one of a few experimental studies measuring demand for information used for decision-making (instrumental information). Previous experimental studies studies the demand for signals in the prediction game in which subjects have to choose an optimal state

[1] https://www.nytimes.com/2010/07/24/us/24hearings.html

under uncertainty. The field experiment conducted by (**?**) finds that the demand for information increases with initial uncertainty, but decreases with the signal's accuracy. However, the decrease in accuracy is more modest than expected for a Bayesian decision-maker resulting in subjects underpaying for high-quality signals. The laboratory experiment of **?** finds that subjects tend to underreact to the accuracy of the binary signal about state of the world, but put a premium on completely certain signals. The paper of **?** similarly employs a prediction game but varies priors on top of signal charasteristics. reducing prior uncertainty makes more signals non-instrumental in the sense that there should be no effect from a signal to optimal decisions. She find that many subjects choose non-instrumental over instrumental signals which is consistent with

Our setup differs in two important aspects from (**??**), because we study alerts and not prediction tasks. The subject faces a costly protection decision and not a prediction decision, resulting in three distinct payoffs: full payoff, full payoff minus protection costs and full payoff minus losses. It means that risk preferences affect the value of information and can change sensitivities to false-positive and false-negative rates. Our findings however are similar to prediction game findings. Consistent with **?** we also find that subjects undervalue accurate signals, but we do not find a premium for certain signals. And similar to **?** we find that subjects do not properly account for interaction between prior probabilities and signal characteristics.

Due to its applicability for studying preferences over expectations, there is a larger stream of literature on the demand for non-instrumental information. **?** find that subjects are willing to pay for signals even when these signals are excessive for making optimal choices. Their design involves subjects choosing between two boxes with one box containing a prize of $20. Most subjects pay just to know the probability of finding $20 in box A even if this box is more likely to contain a prize in all the possible states. This finding is inconsistent with expected utility maximization but indicates instead having preferences for certainty before making choices. Similar to this paper, **?** also study preferences over information structures differing which differ in false-positive and false-negative rates but in their setup allows for a larger role of expectations. They find that for a positive potential outcome, most subjects prefer facing high false-negative rates rather than high false-positive rates. In other words, they tolerate uncertainty after negative signals better than uncertainty after positive signals. These preferences are salient: subjects require an average payment of 18-35 cents to switch to their least preferred information structure.

There is some mixed evidence that people update beliefs differently when these beliefs are ego-relevant or concern future gains and losses. **?** find asymmetry in updating ego-relevant beliefs such as beauty and IQ. Subjects update more after receiving positive signals and do not update enough after negative signals. Additionally, subjects with high posterior ego-relevant beliefs are willing to pay to receive a more precise signals, but require a compensation for learning when their beliefs are low. In contrast, **?** does not find any updating asymmetry with respect to either ego-relevant beliefs or beliefs about future payoffs.

Our paper is the first to measure value of information in the experimental setting of diagnostic tests or alarms. Previous work studies the use of alarms in context of medical testing, medical monitoring, safety alarms and extreme weather. Early literature on decision-making of medical professionals finds that doctors suffer from multiple biases when ordering testing, including inaccurate posterior probability estimation due to availability heuristics, hindsight bias and regret (**?**). **?**find that very few mammologists understand mamogram results and tend to overestimate probability of cancer based on a positive result. Providing practitioners with natural frequencies instead of probabilities tends to reduce this bias.

Patients' willingness-to-pay for medical tests is large and largely responsive to test accuracy (**???**). But there are several apparent violations of rationality. First, users are willing to pay for tests having little or zero diagnostic value (**??**). For example, **?** find that 73% of Americans in their survey prefer a free full-body CT scan versus one thousand USD cash. However, medical professional do not recommend full-body CT scans for healthy people due to extreme likelihood of false-positive findings. Second, the framing of test accuracy seems to matter a lot. **?** conduct a discrete-choice experiment to measure willingness-to-pay for the colorectal cancer screening. Their subjects agree to get 23 unnecessary colonoscopies in order to find one additional true cancer, but only 10.4 for reducing the number of cancers missed by one even though these descriptions are equivalent. Surprisingly, the perceived risk of cancer (prior) did not significantly affect the WTP in their study though the effect may come from its relatively low variation in the population.

This work also relates to the vast literature on demand for insurance and protection. Similar to our findings, several studies observe that the demand for insurance goes up after the recent experience with low-probabiity events. Field evidence indicates that people underinsure with respect to rare natural disasters (Friedl et al, 2014). **?** find no under-insurance for low-probability events in the laboratory setting. One offered explanation (**?**) is that subjects overweight recent evidence leading to underinsurance when there were no negative events in the recent past and to overinsurance after the fact. It is consistent with underweighting prior probabilities relative to more recent signals.

The bias we are finding is similar to the base-rate and signal neglect phenomenons. Psychology researchers **?** and **?** first observed that subjects underweighted prior probabilities (base rates) when calculating posteriors. This phenomenon had received the name of *base-rate neglect*. Multiple studies in economics then confirmed (**??**) this phenomenon in incentivized laboratory experiments. Most of these studies find that subjects also underweight signals on top of priors. We observe both phenomenons in responses to our belief elicitation task, but the calculation of signals' values differs substantially from the calculation of posterior probabilities. While the calculation of posterior probabilities would require using a Bayes formula, signal's value depends only on products of prior probabilities. However, we observe that subjects underestimate the effect of priors compared to theoretical predictions for an expected-utility decision-maker.

2 Model

Environment. The model describes a decision-maker considering a purchase of threat-assessment information. Let $\omega \in \{0, 1\}$ denote the state of world, where 1 corresponds to some adverse event happening with probability π. The decision-maker has a lower utility in the adverse state, but only if she does not take the protective action. Denote actions by $a \in \{0, 1\}$ with 1 meaning taking the protective action. The protection technology is perfect: protected agents bear no losses but pay protection costs c regardless of the state ω. Decision-maker preferences are described by the utility function which depends on wealth Y, protective action a and potential damage in the adverse state $\omega(1 - a)$. Utility is separable in wealth, protection costs $c > 0$ and potential loss in the adverse state $L > c$ [2]:

$$U = U(Y, a, \omega(1 - a)) = u(Y - ac - \omega(1 - a)L) \tag{1}$$

The decision-maker can purchase a binary informative signal $s \in \{0, 1\}$ about the state of the world before making a decision. Let $P_{ij} \equiv P(s = i|\omega = j)$ be the probability of a signal taking value i conditional on the state of the world being j. After receiving the signal, the decision-maker updates her belief on the likelihood of the bad state to $\mu(s)$. Unless specified otherwise, we assume that the decision-maker forms her posterior beliefs by using the Bayes rule. Hence the posterior belief equals:

$$\mu(s) = \frac{\pi P_{s1}}{\pi P_{s1} + (1 - \pi)P_{s0}} \tag{2}$$

We also assume without loss of generality that a higher signal means a higher posterior probability of an adverse event $\mu(1) \geq \mu(0)$. Otherwise we can always re-label the signals.

Preferences. If there is no signal, the decision-maker protects if and only if it increases their expected utility:

$$EU_0 = \max[u(Y - c), \pi u(Y - L) + (1 - \pi)u(Y)] \tag{3}$$

The signal can increase expected utility if the decision-maker reacts differently to positive and negative signals. Under these assumptions, her expected utility with a signal is:

$$EU_s = \pi P_{11} u(Y - c) + \pi P_{01} u(Y - L) + (1 - \pi)P_{10} u(Y - c) + (1 - \pi)P_{00} u(Y) \tag{4}$$

[2] Separability condition does not impose additional restrictions on the utility function U as long as the variation in wealth has limited range. More specifically, if $Y \in [Y_{min}, Y_{max}]$ and $c < Y_{max} - Y_{min}, L < c + (Y_{max} - Y_{min})$, then the function $u(\cdot)$ can be constructed from segments of $U(\cdot, 0, 0), U(\cdot, 1, 0), U(\cdot, 0, 1)$. While the resulting function $u(\cdot)$ is not necessarily monotonic, it is likely to be monotonic if protective actions and potential damages are relatively high.

We consider the maximum amount b which the decision-maker is willing to pay for the signal. In our framework, it is a price paid with a signal such that a decision-maker is indifferent between having a signal and paying b and not having a signal. Because the decision-maker can always ignore a useless signal, the signal's value is bounded from below by zero. Hence it equals to the maximum between zero and the solution to the following equation:

$$\begin{aligned} P(s=1)u(Y-b-c)+\pi P_{01}u(Y-b-L)+(1-\pi)P_{00}u(Y-b)= \\ =\max[u(Y-c),\pi u(Y-L)+(1-\pi)u(Y)] \end{aligned} \tag{5}$$

The left-hand side expression of this equation is a strictly decreasing function of b. Additionally, for $b \to \infty$ the left-hand side is smaller than the right-hand side. It implies that the equation (5) above has a at most one positive solution.

Obviously, perfectly accurate signals always have positive value $b > 0$ because the payoff distribution with the signal first-order stochastically dominates the distribution without the signal. If the decision-maker protects without a signal, a perfect signal reduces the protection costs and if she takes chances, then it reduces losses in the adverse outcome from L to $c < L$. However, it is harder to determine the value of the imperfect signal without imposing more restrictions on preferences as it requires weighing $u(Y-L)$ against $u(Y-c)$.

Risk-neutral agent. If the decision-maker is risk-neutral, the expression above collapses to:

$$b + P(s=1)c + \pi P_{01}L = \min[c, \pi L]$$

The signal's value is just:

$$b = \max[0, \min[c, \pi L] - P(s=1)c - \pi P_{01}L] \tag{6}$$

We can express WTP b as a function of priors, false-positive and false-negative rates. This is the equation we use in our empirical work:

$$b = \max[0, \min[c, \pi L] - \pi(1-P_{01})c - (1-\pi)P_{10}c - \pi P_{01}L] \tag{7}$$

The sensitivity of (positive) value b with respect to false-positive and false-negative rates is given by:

$$\frac{db}{dP_{10}} = -(1-\pi)c \tag{8}$$

$$\frac{db}{dP_{10}} = -\pi(L-c) \tag{9}$$

Both false-positive and false-negative rates decrease the (positive) signal's value. The effect

is proportional to the adverse state probability for the false-negative rate and to the non-adverse state probability for the false-positive rates.

Risk Aversion Effects. In a more general expected utility framework, risk aversion can both increase and decrease the signal's value. More specifically, risk aversion decreases the value when the protection costs are low:

Proposition 1. *If protection costs are low $c < \pi L$, then the strictly risk-averse decision-maker pays less than a risk-neutral one.*

Proof. See the Appendix. □

It is harder to make definite statements for lower risks or higher protection costs. For example, risk aversion increases value of a perfect signal as long as risk-averse decision-maker still chooses to not protect without a signal. This follows from the standard argument of increasing demand for insurance with risk aversion and the fact that the protection problem with a perfect signal is isomorphic to the insurance problem with deductible c.

Next, we study the effect of false-positive and false-negative rates on the signal's value b. Assuming a differentiable utility function $u()$ we use implicit differentiation to derive sensitivities of WTP b to false-positive and false-negative rates:

$$\frac{db}{dP_{10}} = -\frac{(1-\pi)(u(Y-b) - u(Y-c-b))}{D(\pi, P_{01}, P_{10}, b)}$$

$$\frac{db}{dP_{01}} = -\frac{\pi(u(Y-c-b) - u(Y-L-b))}{D(\pi, P_{01}, P_{10}, b)}$$

With the denominator equal to the expected marginal utility:

$$D(\pi, P_{01}, P_{10}, b) \equiv P(S=1)u'(Y-c-b) + \pi P_{01}u'(Y-L-b) +$$

$$+(1-\pi)P_{00}u'(Y-b) = E[MU] > 0$$

It is clear that the signal's value decreases with false-positive and false-negative rates $\frac{db}{dP_{10}}, \frac{db}{dP_{01}} < 0$. We can also say a bit more about the sensitivity to false-negative rates:

Proposition 2. *Risk-averse and imprudent decision-maker has higher sensitivity to false-negative rates as compared to a risk-neutral one.*

Proof. Use the mean value theorem to rewrite the sensitivity as:

$$\frac{db}{dP_{01}} = -\frac{\pi u'(\zeta)(L-c)}{E[MU]}, \zeta \in (Y-c-b, Y-L-b)$$

Now let X denote a random payoff of the decision-maker with a signal. A risk-averse decision-maker puts a positive value on the signal only if its expected payoff is higher than the payoff

with full protection: $EX > Y - c - b$. If a decision-maker is imprudent ($u''' < 0$) then $E[MU] \equiv E[u'(X)] < u'(EX)$. Next, because u' is a strictly increasing function and $EX > Y - c - b$: $u'(\zeta) > u'(Y - c - b) > u'(EX)$. Hence $\frac{u'(\zeta)}{E[MU]} > 1$ and $\frac{db}{dP_{01}} < -\pi(L - c)$. □

However, risk aversion can both increase and decrease subject's sensitivity to false-positive rates depending on the utility function curvature and signal's characteristics. Intuitively, an expected marginal utility of a strongly risk-averse subject with a bad signal can be lower than the average slope of the utility function between $Y - c - b$ and $Y - b$ reducing sensitivity to false-positive rates. It can also be higher if either the signal is good or the curvature is small.

3 Experimental Design

In each session, subjects received a USD 5 show-up fee and were endowed with USD 25 that they might lose in the experiment. Subjects must then make a series of decisions in four sets of tasks: (i) Blind Protection; (ii) Informed Protection; (iii) Belief Elicitation; and (iv) Willingness to Pay Elicitation. To verify that subjects understand these tasks, they must answer a quiz before each task. If a subject gets any answer wrong, they read correct answers and explanations for each wrong answer. Additionally, subjects receive extra questions if they give wrong answers in a 5-question quiz given before the Informed Protection task. We do this because we consider Informed Protection as a first challenging task in the sequence which understanding is essential for the rest of the tasks. Each set of tasks has 6 rounds, for a total of 24 rounds. One of these rounds is selected at random as the payment round. A copy of the instruction is included in Appendix XX.

Subjects began with the Blind Protection (BP) task. In each round of the BP task, subjects must decide whether to insure (or "protect") against an adverse event (i.e., drawing a black ball from a box). Subjects were informed of the prior probability of drawing a black ball before making their decision. The cost to protect is USD 5. If a black ball is drawn, an unprotected subject will lose USD 20. Subjects then played six rounds, where the probability of drawing a black ball was varied between XX and XX percent in each round. During the BP task, subjects did not receive any feedback on how that round would have been realized were it chosen as the payment round.

Second, subjects played 6 rounds of the Informed Protection (IP) task. As in the BP task, subjects must make a protection decision in each round after receiving information about the prior probability of the adverse event (or a black ball). However, subjects are now given a signal which, following Coutts (2019), was represented by a gremlin that gave a hint of whether the randomly drawn ball is black. There are three types of gremlins, each representing a signal type: accurate (an honest gremlin), false positive (a black-swamp gremlin that always announces that the ball is black), and false negative (a white-swamp gremlin that always announces that the ball is white). Figure XX illustrates how the different gremlin types were presented to the

subjects. Subjects knew the composition of the group of gremlins from which the hint would come from, but did not know the type of randomly drawn gremlin that provided the hint. After receiving the hint, subjects decide whether or not to protect. We vary the proportion of black balls in the box (prior probability of a black ball) and the composition of gremlins (the signal) between rounds.

The third task is the Belief Elicitation (BE) task, where we elicited subjects' beliefs about the likelihood of an adverse event and an adverse signal conditional on prior and signal characteristics in an incentive-compatible way. Similar to the IP task, subjects were informed of the prior probability of a black ball and the composition of the group of gremlins that would provide an additional signal. However, instead of asking subjects to make a protection decision, we asked them to estimate the probability that: (i) the ball is black ball when a randomly drawn gremlin says that it is white; (ii) the ball is black when a randomly drawn gremlin says that it is black.

We follow the stochastic version of the Becker-DeGroot-Marshak mechanism developed by **?** and **?** to make subjects' belief elicitation responses incentive compatible. In this mechanism, a subjects submits his belief of the probability of the event $\mu \in [0, 1]$. If this belief is above some uniform random number $r \in [0, 1]$ then the subject receives x only if the event E happens. Otherwise the subject's payoff comes from an independent lottery which pays x with probability r and 0 otherwise. The benefit of this mechanism versus other probability elicitation mechanism (for example, quadratic scoring) is that reporting truthfully is a dominant strategy regardless of risk preferences (**?**). The only requirements a subject needs to satisfy are probabilistic sophistication and dominance: they rank lotteries based on their probabilities only and prefer higher probabilities of higher payoffs. Given the difficulty of the mechanism, our instructions start with stressing the fact that reporting beliefs truthfully always pays off better than lying, and only afterwards explain the algorithm used to calculate payoffs.

Finally, in the Willingness to Pay Elicitation (WTPE) task, subjects were asked for their willingness to pay (WTP) for signals. Subjects know the prior probability of a black ball and the group composition of the gremlins from which one will be randomly drawn to give a hint. We then ask subjects for their WTP to receive a hint from a randomly drawn gremlin. Subjects can choose a value from USD 0 to 5 with USD 50-cent increments. Their decisions are incentive compatible: if a WTPE round is selected as the payment round, a random price of a hint will be drawn. If that price exceeded the subjects' WTP, they will play another round of BP. Otherwise, the subject would pay for the hint and play another round of IP. After completing the WTPE task, subjects were asked a few demographic questions. The session concluded with the random selection and realization of the payment round, after which subjects were paid and dismissed.

The first three tasks were designed to provide measures of the different components of WTP described in Section XX and use them to examine the extent to which they explain subjects' WTP measured in the WTPE task. We use the BP task both to measure subjects' responses

to the prior and their risk aversion. Next, we use the IP task to examine how signals affect protection decisions. Finally, we use the BE task as a measure of subjects' ability to estimate the probability of a signal for a given quality and to perform Bayesian updating. To construct these measures, we presented our subjects with 6 different priors for the BP task, and 3 priors and 2 gremlin groupings for the IP, BE, and WTPE tasks. Table XX shows the values of the different priors in our treatments, as well as the gremlin groupings (along with the associated false positive and false positive rates) that we used for the different tasks.

We conducted this experiment in the Behavioral Business Research Lab (BBRL) at the University of Arkansas between October and November 2021. The experiment was implemented using Qualtrics. There were a total of 105 subjects. 84 percent of the subjects were university students and 41 percent were male. About 60 percent of the subjects had taken at least one statistics course. On average, including the show-up fee, subjects received around USD 26 for a session lasting around 45 minutes.

4 Results

We begin with a descriptive analysis of subject behavior in the different tasks. We follow that discussion with a regression analysis to explain subjects' WTP for signals of different qualities. Our regression results suggest that subjects' WTP deviated from those of a risk-neutral utility maximizing subject which was driven by their failure to fully account for signal quality when calculating their WTP. Furthermore, we find that these deviations remained after controlling for risk aversion or subjects' ability to perform Bayesian updating. The final subsection explored theories that are consistent with these observed behaviors.

4.1 Risk-Aversion, Updating, and Willingness to Pay

Risk Aversion. Subjects' responses to our first blind protection tasks are generally consistent with the expected utility framework. Figure XX shows the probability of protection decision given prior probability of a black ball from the BP task. Subjects protect more when the probability of the negative outcome is higher: only 13% subjects protect when the probability of a black ball is 10% in contrast to 70% protecting when the probability is 30%.

Also consistent with expected utility theory, the majority of subjects (54%) have a threshold level of probability above which they always protect. About 30% of subjects though make at least one switch in which they protect for probability P_1 and do not protect for at least one probability $P_2 > P_1$. Most of these subjects (24 out of 39) make only one switch in 6 rounds, and their responses differ from a rational EU maximizer responses by only one element indicating a likely inattention error. Following –, we adjust these responses by making that one necessary correction when calculating subjects' risk aversion.

Blind protection responses indicate significant heterogeneity in terms of risk aversion. Risk-neutral subjects maximize their expected utility by protecting whenever the prior probability exceeds 0.25, which is the ratio of the protection cost (USD 5) to the potential loss (USD 20). In contrast, many subjects start protecting for lower probabilities of 0.1 or 0.2 indicating risk aversion. A smaller group of subjects makes choices consistent with risk loving by never protecting or protecting for the probability of 0.3.

Protection Response to Signal. Next, we examine how subjects responded to signals for given priors. Because subjects were only presented with the signals and had to calculate the true posterior probabilities, their protection decision in the IP task would be based on the beliefs of these posteriors.

Figure XX presents the subjects' responses in the IP task plotted against the true posteriors for the signals. We find a pattern similar to subjects' responses to the priors, with a strong positive correlation between the share who protected and posterior probabilities. Examining subject responses between the first two tasks, we observe that conditional on the true probabilities, subjects protected more in the IP task compared to the BP task at low probabilities. This suggests that at low probabilities, subjects' updated beliefs overshot the true posterior. It can also indicate some ambiguity aversion in the IP task. However, we cannot use these tasks to make general inferences about subjects' ability to update since the BP task did not cover the full range of probabilities.

Subjects' Ability to Bayesian Update. Figure XX shows how subjects update their beliefs given a signal of a certain quality based on their responses to the BE task. To construct Figure XXa , we use subjects' responses in the BE task to calculate the difference between the posterior and subjects' elicited belief on the posterior probability of a black ball for a given signal. In Figure XXa, we plot the distribution of these differences, which captures the errors in their Bayesian updating process. The distribution of updating errors is centered at 0. Figure XXb presents the scatter plot of the elicited belief and the true posterior probability for a given signal. The correlation between elicited belief and the true posterior was XXXX .

However, Table XX shows that some of our treatments included signals that are perfect (i.e., where the group of gremlins comprises only honest gremlins) as well as treatments where the ball color could have been deduced with certainty (e.g., a group with honest gremlins and white-eyed gremlins with a hint that the ball is black — or vice versa). In Figure XXa, we plot the distribution of updating errors when without cases where the ball could have been deduced with certainty (including the case of all honest gremlins). The errors tend to be negative, suggesting that subjects are more likely to overestimate the likelihood of adverse events (given priors and signals) when they update their beliefs. The correlation between their belief and the true posterior in this subset of observations is XXX.

At the same time, we observe incorrect belief reports even when the ball color can be deduced logically with certainty. Figure XXa plots the distribution of updating errors in such a case. While most beliefs (80%) elicited with completely honest groups of gremlins are completely

accurate, about 20% are incorrect. About half of these errors involve reporting uncertainty about the color and about half reports probability 1 when the probability should be zero. The ball color can be also certain even when some gremlins in the group are dishonest. For example, if the group of gremlins includes only honest and black-eyed gremlins (always reporting black) then the hint of a white ball means that the ball is certainly white. However, only 51% of responses in these cases are absolutely correct and the rest indicate uncertain beliefs.

4.2 Regression Analysis: Willingness-to-Pay

Figure XX shows how subjects deviate substantially from the theoretical value for a risk-neutral subject. Here we explore these deviations more systematically by signal characteristics and risk preferences.

In Table XX, we use a regression analysis to investigate how and why subjects deviate from the risk-neutral subjects' theoretical WTP for a signal of a given quality (conditional on prior). We first estimate how individual deviations from the theoretical benchmark b_s^* for a given signal s are correlated with signal's charactaristics:

$$b_{is} - b_s^* = \beta_0 + \beta_1 FalsePositive + \beta_2 FalseNegative + \epsilon_{is}$$

where b_{is} is the reported WTP of individual i for treatment s and b_s^* is the signal's value for a risk-neutral subject, and FalsePositive (FalseNegative) is the false positive (false positive) costs variables that captures signal quality. We calculate false positive costs as the product of prior probability of a white ball multiplied by the conditional probability of getting a black signal ("The ball is black!") from a randomly chosen gremlin: $FalsePositive = (1-\pi)P_{10}c$. Similarly we calculate false negative costs as the probability of an adverse state multiplied by a conditional probability of getting a white signal conditional on the ball being black and multiplied by potential loss $FalseNegative = \pi P_{01}L$. Note that these costs already account for expected frequency of receiving different incorrect signals as consistent with their base rate. If our subjects were risk neutral expected utility maximizers, we expect β_1 and β_2 to be zero.

Column 1 of Table XX confirms that subjects' WTP deviated from the theoretical benchmark. Subjects did not fully account for signal quality, resulting in overpaying for signals with either high false-positive and false-negative costs. Naturally, two potential sources of deviations from this theoretical benchmark based on a risk-neutral Bayesian updater are subjects' risk-preferences and their ability to perform Bayesian updating. To test for these mechanisms, we interacted the false positive (FP) and false negative (FN) variables with individual risk aversion, whether individuals have accurate belief (as measured by our BE task), and the different priors.

Column 2 shows the results of the regression where the signal quality variables were interacted with the subject's risk preference.[3] This premium doesn't seem to come from risk

[3]Our risk preference estimates come from blind protection choice: subjects switching from no protection to

aversion, as the coefficient on the interaction of risk aversion with FP and FN rates is relatively small and insignificant. Belief accuracy measured in the belief elicitation task apparently explains away the excess sensitivity to the FP rate but this finding should be taken with caution because the coefficient is not statistically significant despite its large absolute magnitude.

These results suggest that, on average, subjects failed to fully account for signal quality, resulting in overpaying for signals with high false-positive and false-negative costs. FP/FN significantly impact the deviation from the theoretical value no matter what else is included.

Another plausible explanation would be that subjects are engaging in some sort of probability weighting depending on the priors along the line of XXXX. In Column 5, we find that .... Probability weighting involves making decisions not based on actual probabilities but on transformed probabilities. Typical transformation is mapping low probability to higher probabilities and very high probabilities to lower, resulting in compressing the distribution more towards 0.5. For example, a subject can react to an event with 0.05 probability as if it is an event with 0.15 probability. Finally, we consider the signal characteristics, i.e., do subjects deviate from the theoretical value of a signal because they imbue false positives or negatives with undue importance.

Digging into these determinants of the discrepancy, Table 2 looks at the effect of FP/FN by each possible prior. When separated this way a distinct pattern emerges: when priors are low, subjects underreact to FP rates and overreact to FN rates compared to a risk-neutral decision-maker. When priors are high, subjects do the opposite by overreacting to FN rates and underreacting to FP rates.

One potential explanation for the observed pattern is that subjects do not correctly estimate the frequency of false-positive and false-negative outcomes. These frequencies are sufficient statistics for determining the signal's value in any expected utility framework (see x). Subjects are given only false-positive rates and priors, but what matters for the signal's value in most common frameworks including the expected utility framework, is the frequency of observing false-positive and false-negative signals. We call this bias a probability estimation bias. This bias is similar to other biases in calculating posterior probabilities such as base-rate neglect or signal neglect, but instead of failing to calculate a posterior probabilities here subjects fail to estimate the probability of a compound event given priors and conditional probabilities.

5 Explaining Underutilization of Signal Quality

We also need to make sure that any more accepted theories of decision-making under uncertainty do not explain this pattern. We analyze three alternative theories, including risk aversion, loss aversion and probability weighting. Both risk aversion with prudence and loss aversion can

protection at exactly the cost-loss ratio $\pi = 0.2$ are considered risk-neutral, while switching at lower (higher) levels indicates risk aversion (risk-loving).

produce choices consistent with the observed pattern. Probability weighting in theory produces the pattern opposite to the observed one.

WTP regressions show that subjects overpay for signals with high FP and FN rates, but is this consistent with the way they use signals? In Table 3, we look at what influences a subject's choice to protect in the IP treatment where they have been given a signal. For any rational subject (following one of the theories above), posterior probability provides all the information needed for making protection decisions. But we observe that even controlling for the posterior probability (with splines of the posterior prob), subjects tend to protect more when faced with signals with high false-positive or false-negative rates.[4]

WTP regressions show that subjects overpay for signals with high FP and FN rates. Informed protection responses also demonstrate that subjects overuse bad signals. Controlling for the posterior probability (with splines of the posterior probability), subjects tend to protect more when faced with signals with high false-positive or false-negative rates.

Their behavior is consistent with underweighting prior and/or signals when calculating the posterior probability. Even controlling for the posterior, subjects protect more when the prior probability is high (¿0.2) (columns 3-4). For high priors, they also react less to false-positive rates but more to false-negative rates. Practically all the extra-protection in response to false-negative rates comes from white signals (as expected), but the effect of false positive rates exists both for white and black signals.

Subjects obviously do not behave as risk-neutral decision-makers when paying for signals. Which theory explains their behavior the best? We study this problem in context of different decision theories and find that several of them are consistent with the pattern (of underreacting to FP rates and overreacting to FN rates for low priors and vice versa) for certain ranges and parametrizations. Strictly risk-averse and prudent decision-makers become relatively more sensitive to false-negative rates with increasing priors as compared to risk-neutral subjects. For loss-averse subjects, the sensitivity to false-positive rates should decrease with priors due to shifting baselines. Finally, the probability estimation bias means that subjects do not correctly map reported false-positive and false-negative rates and priors into probabilities of false-positive and false-negative outcomes.

Here we specifically test for the probability estimation bias versus other theories. Probability estimation bias predicts that subjects consistently react to other factors such as FP and FN rates conditional on priors and probabilities of FP and FN outcomes. Indeed we observe that even with flexible controls for priors and total probabilities, subjects reduce their WTP with increasing FP and FN rates.

[4]To account for the non-linearity of potential response to posterior probability, we split the domain of this variable in 5 equal segments. Then we add 5 linear splines, so that each variable is equal to the posterior probability when the posterior probability is inside the corresponding segment and zero otherwise.

A Figures

Figure 1: Average Protection Response

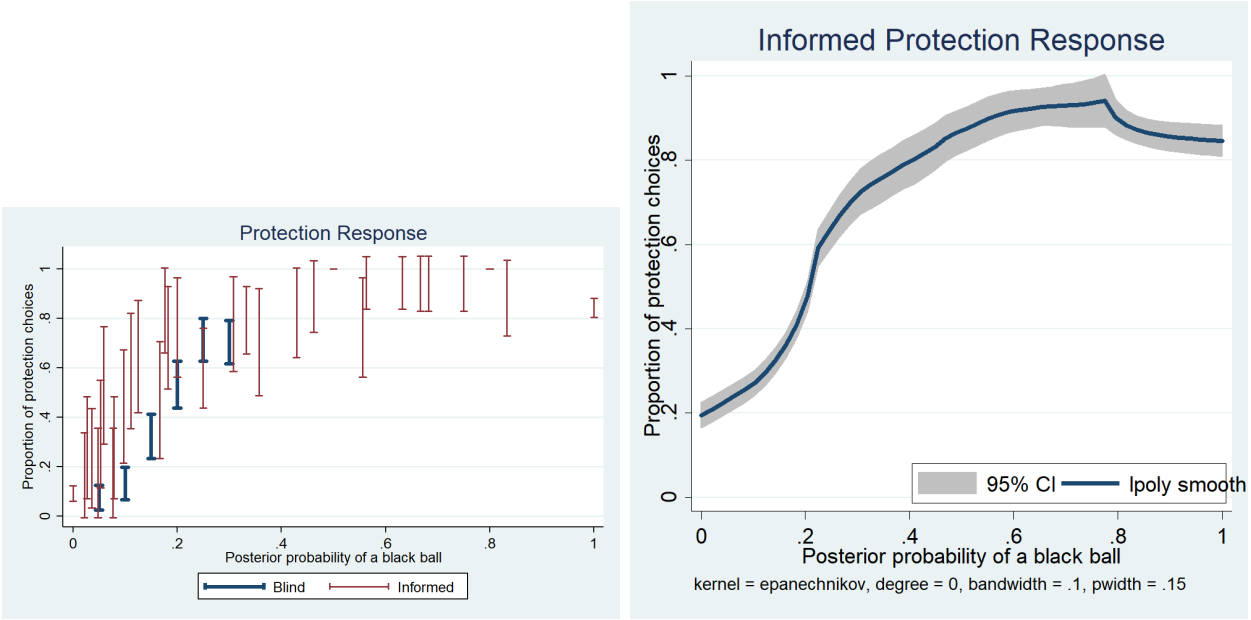


Figure 2: Belief Updating

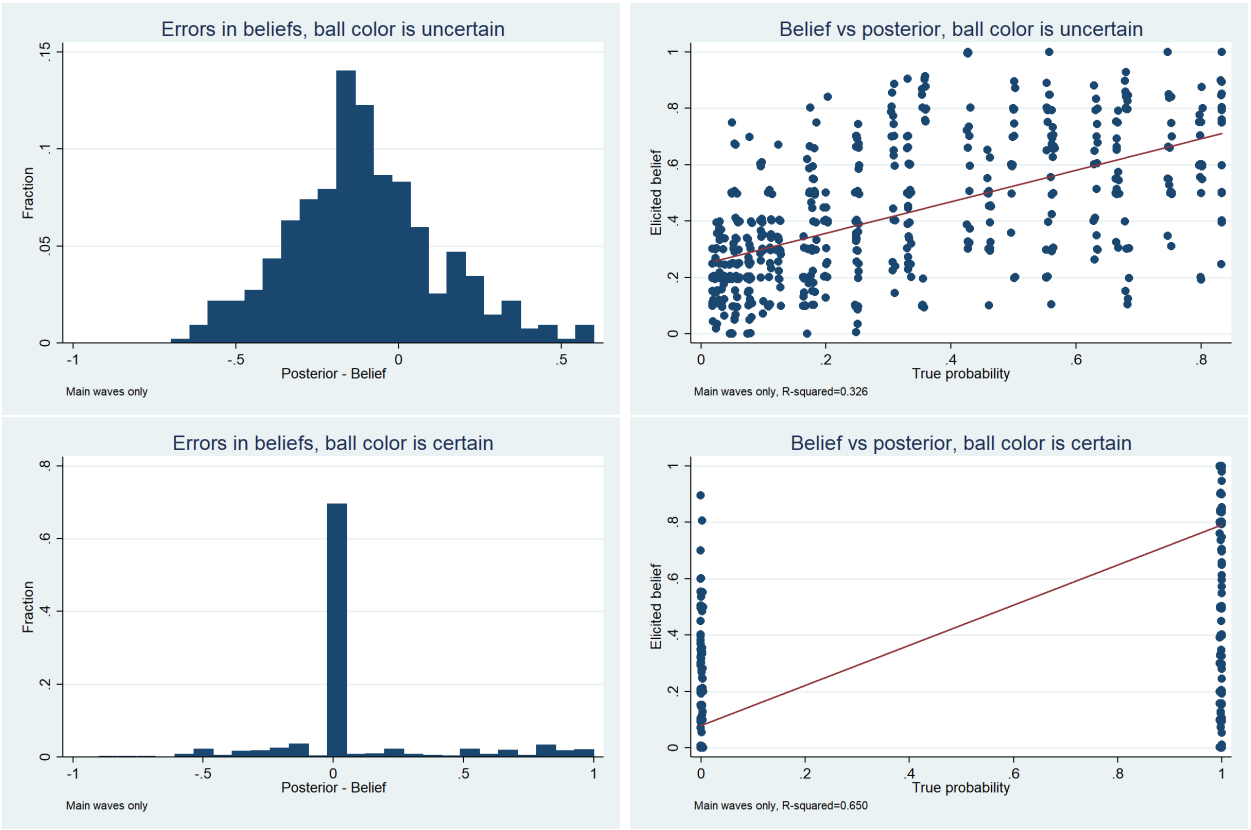


Figure 3: Theoretical vs actual WTP

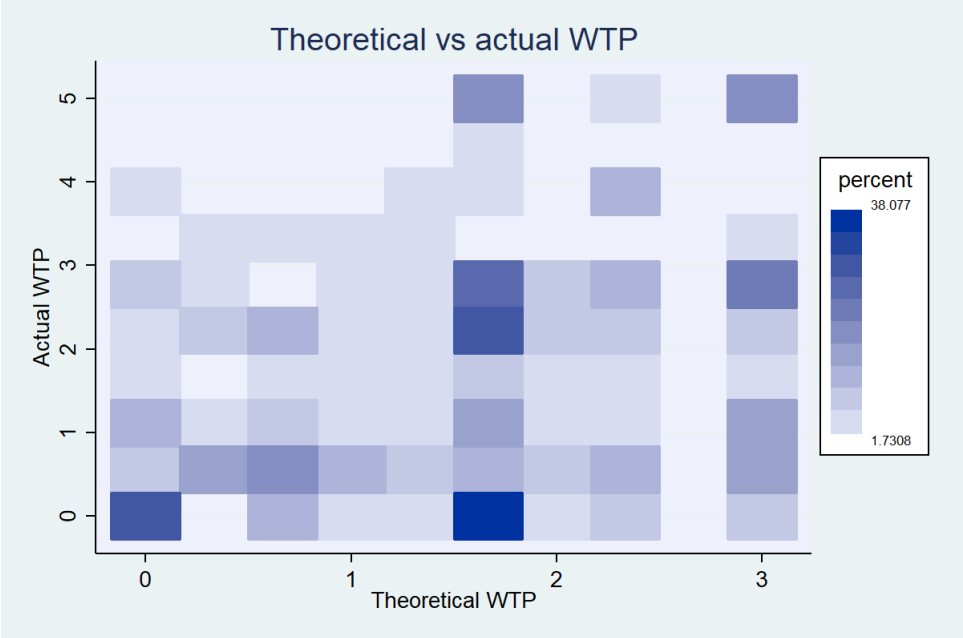
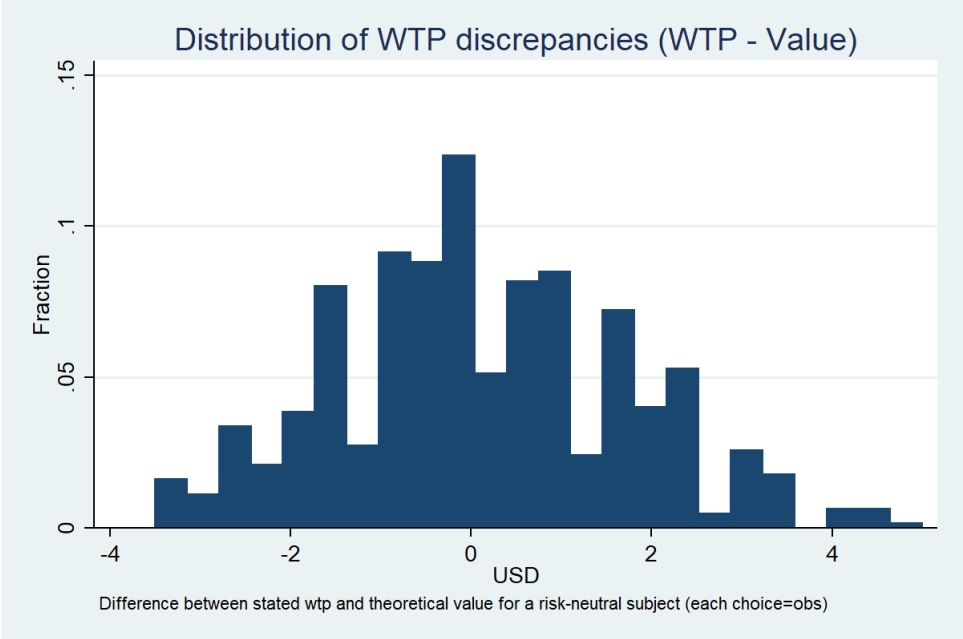


Figure 4: WTP discrepancy



B Proofs

Proof of Proposition 1:

Proof. If protection costs are low enough $c < \pi L$ than the risk-neutral decision-maker should always protect without a signal:

$$U = \max[\pi(Y - L) + (1 - \pi)Y, Y - c] = Y - c$$

It means that a strictly risk-averse decision-maker with a utility function $u()$ should also protect:

$$\pi u(Y - L) + (1 - \pi)u(Y) < u(\pi(Y - L) + (1 - \pi)Y) = u(Y - c)$$

Then denote stochastic payoff with a signal as X so that expected utility with a signal is $Eu(X-b)$ where b is the willingness-to-pay solving:

$$Eu(X - b) = u(Y - c)$$

Let b_0 be the willingness-to-pay for a risk-neutral decision-maker. By Jensen's inequality:

$$Eu(X - b_0) < u(EX - b_0) = u(Y - c) = Eu(X - b)$$

Because expected utility with a signal is a decreasing function of b_0 we obtain $b > b_0$. □

C Alternative Explanations

We show that the observed pattern of switching sensitivities to false-positive and false-negative rates with priors is not consistent with the probability-weighting and preferences for certainty.

Loss-averse decision-maker. A loss-averse decision-makers have extra-sensitivity to losses or deviations of incomes below a certain baseline. As a result, utility function becomes convex in the domain of losses. There are different functional specifications of loss aversion, but —

Probability-weighting decision-maker. A decision-maker does probability weighting if it reacts to overreacts to very low probabilities and underreacts to very high probabilities. Their behavior can be described as a standard expected utility maximization but transforming all the probabilities to number closer to the middle of the support (1/2). We can show that probability weighting implies a reverse pattern of responses to false-positive and false-negative rates with priors and hence cannot explain our observations. Willingness-to-pay solves the following equation, which is equivalent to eqation () but with probabilities x replaced by its monotonic transform $f(x)$:

$$\begin{aligned} f(P(s=1))u(Y-b-c) + f(\pi P_{01})u(Y-b-L) + f((1-\pi)P_{00})u(Y-b) = \\ = \max[u(Y-c), f(\pi)u(Y-L) + f((1-\pi))u(Y)] \end{aligned} \tag{10}$$

Taking derivatives from both sides we obtain:

$$\frac{db}{dP_{10}} = -\frac{(1-\pi)(u(Y-b) - u(Y-c-b))}{D(\pi, P_{01}, P_{10}, b)}$$

$$\frac{db}{dP_{01}} = -\frac{\pi(u(Y-c-b) - u(Y-L-b))}{D(\pi, P_{01}, P_{10}, b)}$$

Preferences for certainty. ? observe subjects paying for signals which have no potential effect on their decisions. Their theoretical explanation assumes that decision-maker's certainty in a chosen action directly affects their decision. In our setup it is equivalent to adding a strictly increasing function of the posterior belief μ to the consumption utility. Math —