Report on "YGAME-D-25-00181"

Title: Preferences for Warning Signal Quality: Experimental Evidence

Summary

This paper investigates how people value warning signals with different false-positive (FP) and false-negative (FN) error rates across varying prior probabilities of adverse events. Using a laboratory experiment with 105 university subjects, the authors implement a four-stage experimental design: Blind Protection (BP), Informed Protection (IP), Belief Elicitation (BE), and Willingness-to-Pay (WTP). The experimental parameters include four prior probabilities (0.1, 0.2, 0.3, 0.5) and seven combinations of FP and FN rates, creating 28 possible tasks with each subject completing 6 tasks. The paper's central finding is that subjects exhibit asymmetric sensitivity to signal errors depending on the prior probability: they underweight FP costs for low-probability events and underweight FN costs for high-probability events. The authors attribute this pattern to subjects failing to distinguish between the two error types.

Assessment

Overall I lean in favor of rejecting and resubmitting this paper. While the research question is practically relevant, I am concerned about fundamental theoretical errors, experimental design artifacts, and analytical limitations that undermine the paper's conclusions. The contribution may not be enough for a journal like GEB.

Major Comments

Theoretical Issues

First, the term "signal" used by the authors is somewhat confusing. When it is ex-post, the signal is the realization from an information source, while when it is ex-ante, it is the source itself. Given that the experimental design contains elicitations regarding both ex-post signals and ex-ante sources, I would like to see a clearer exposition of this terminology. In general in the literature, signal or news are often referred to as signal realizations, and ex-ante it is called the information source. I will follow this custom in this report.

About the theoretical sections, I have several questions regarding the model and propositions. First, focus on the analyses of Proposition 1:

- On page 6, the assumption that "access to a signal can increase expected utility if the signal affects her protection decisions" influences the calculation of the WTP b in equation (1), as well as the partial derivatives for sensitivity. I do not think this assumption is proper here. The assumption should be more fundamental and made

on the properties of the information source (P_{ij}), not on behaviors. In fact, in the set of parameters chosen for the experiment, four tasks did not satisfy this assumption (non-instrumental information sources). Another suggestion is to verify whether subjects' choices indeed follow this assumption. It provides a measurement at the individual level for contingent reasoning. If a significant number of subjects put a positive WTP for the non-instrumental information sources, then the theoretical analyses based on this assumption become questionable.

- Also in Appendix B.1 Proposition 1, the last equation is not correct. It should be $\leq$ not $=$. The statement "Because expected utility with a signal is a decreasing function of b_0 we obtain $b > b_0$" should be $b < b_0$."

- On page 7, I do not understand the prediction that "the model of a risk-neutral agent suggests that subjects' WTP should have equal sensitivity to costs from false-positive and false-negative signals (equation 2)." In equations (2)-(4), the partial derivatives of the WTP b with respect to FP and FN are different and only equal when the prior is 0.25. But the experiment tasks have varying priors from 0.1-0.5. As a result, Hypothesis 2 is not correct and the analyses for Result 2 should be restructured.

Propositions 2 and 3 rely on quite restrictive assumptions and simplifications. I would like to see discussion of the generalizability of these propositions, perhaps with simulations or numeric methods.

- For instance, Proposition 2 relies on the imprudence" assumption that might be inconsistent with standard utility specifications and the authors' own empirical findings. The condition $U'''(w) < 0$ fails to hold for the most commonly used utility functions in economics: CRRA utility has $U'''(w) > 0$ for reasonable risk aversion parameters $\gamma > 0$, while CARA utility has $U'''(w) > 0$ for any $\alpha > 0$. In the analyses for BP, the authors acknowledge that many of our subjects (24) start protecting at lower priors (0.05-0.15), indicating strict risk aversion. A smaller group of subjects makes choices consistent with risk loving by protecting at a probability of 0.3." There should be further discussion of the validity of the imprudence condition.

- Proposition 3's derivation relies on a first-order Taylor approximation that drops higher-order terms without establishing error bounds or validity conditions. The proof provides no justification for why the neglected terms $o(b)$ and $o(b + c)$ are "small enough" relative to the retained terms. For meaningful values of b and c ($5protectioncost, upto5$ WTP), the neglected terms could be substantial. I would like to see error bounds demonstrating when the approximation is valid.

Experimental Design Concerns

The experiment consists of a four-task structure with WTP as the critical outcome measure. I calculated the correct posterior, optimal protection action after each signal, and WTP for each possible combination of priors and FP and FN values in Table 1 below:

I hope the authors can revise Tables 2 and 3 accordingly since the pooling of priors and FN and FP structures may be uninformative. Given that, it will be more straightforward to check how the elicited posteriors, protection actions and WTPs change for different priors and error types.

My main concern about the experiment is the parameter designs. As we observe from my calculations, many conditions produce extreme values. I am worried that some key findings appear to be driven by systematic measurement artifacts created by the experimental parameter choices rather than genuine behavioral patterns. For instance, on page 16, Result 1 shows that both-error conditions have systematically lowest WTP. This pattern might be suspicious since single-error conditions often produce extreme posteriors (0 or 1) while both-error conditions tend to produce intermediate posteriors. The complexity level is different. Likelihood insensitivity, rather than belief updating, might also explain the valuation. In addition, almost all the both-error conditions generate very low WTPs, thus the apparent overvaluation for them might "simply be due to reversion to the mean."

This may also drive the key conclusion about asymmetric sensitivity. For low prior conditions, the low base rate ($\pi = 0.1$) means theoretical WTP is already near zero for most signals. Adding false positive costs should theoretically drive WTP even lower toward zero, but there is a boundary effect where subjects cannot bid negative amounts, so any positive bid appears as insufficient reduction." This creates a measurement artifact where the failure to reduce WTP enough" is simply because observed WTP is bounded below at zero. For high prior conditions, the high base rate means theoretical WTP could be substantial ($3-5), and adding false negative costs should theoretically reduce WTP from these high levels. However, subjects may be reluctant to bid close to their $30 endowment or have budget constraints, creating a measurement artifact where the "underreaction" is because observed WTP is practically bounded above.

The analyses and conclusions of this paper might be stronger if the FN and FP combinations also induced some intermediate posteriors. Otherwise, I would suggest the authors have a discussion about how the parameter design may create artifactual results.

Data Analysis Issues

The data analyses are not well-structured for readers to grasp the takeaways. Most results use aggregate statistics that mask individual inconsistencies. They often pool subjects across tasks rather than tracking individual patterns and focus on population averages rather than individual coherence.

Table 1: Complete List of Experimental Tasks with Posteriors and Optimal Decisions

Prior π	Gremlins (BE, WE)	FP P_{10}	FN P_{01}	Post. (white)	Post. (black)	Opt. (white)	Opt. (black)	WTP ($)
0.1	2 (0, 0)	0	0	0.000	1.000	not	protect	1.50
0.1	3 (1, 0)	0.33	0	0.000	0.250	not	protect*	0.00
0.1	3 (0, 1)	0	0.33	0.036	1.000	not	protect	1.00
0.1	3 (1, 1)	0.33	0.33	0.053	0.182	not	not	0.00
0.1	5 (1, 0)	0.2	0	0.000	0.357	not	protect	0.60
0.1	5 (0, 1)	0	0.2	0.022	1.000	not	protect	1.20
0.1	5 (1, 1)	0.2	0.2	0.027	0.308	not	protect	0.30
0.2	2 (0, 0)	0	0	0.000	1.000	not	protect	3.00
0.2	3 (1, 0)	0.33	0	0.000	0.429	not	protect	1.67
0.2	3 (0, 1)	0	0.33	0.077	1.000	not	protect	2.00
0.2	3 (1, 1)	0.33	0.33	0.111	0.333	not	protect	0.67
0.2	5 (1, 0)	0.2	0	0.000	0.556	not	protect	2.20
0.2	5 (0, 1)	0	0.2	0.048	1.000	not	protect	2.40
0.2	5 (1, 1)	0.2	0.2	0.059	0.500	not	protect	1.60
0.3	2 (0, 0)	0	0	0.000	1.000	not	protect	3.50
0.3	3 (1, 0)	0.33	0	0.000	0.562	not	protect	2.33
0.3	3 (0, 1)	0	0.33	0.125	1.000	not	protect	2.00
0.3	3 (1, 1)	0.33	0.33	0.176	0.462	not	protect	0.83
0.3	5 (1, 0)	0.2	0	0.000	0.682	not	protect	2.80
0.3	5 (0, 1)	0	0.2	0.079	1.000	not	protect	2.60
0.3	5 (1, 1)	0.2	0.2	0.097	0.632	not	protect	1.90
0.5	2 (0, 0)	0	0	0.000	1.000	not	protect	2.50
0.5	3 (1, 0)	0.33	0	0.000	0.750	not	protect	1.67
0.5	3 (0, 1)	0	0.33	0.250	1.000	protect*	protect	0.00
0.5	3 (1, 1)	0.33	0.33	0.333	0.667	protect	protect	0.00
0.5	5 (1, 0)	0.2	0	0.000	0.833	not	protect	2.00
0.5	5 (0, 1)	0	0.2	0.167	1.000	not	protect	1.00
0.5	5 (1, 1)	0.2	0.2	0.200	0.800	not	protect	0.50

Notes: BE = black-eyed gremlins (false positives); WE = white-eyed gremlins (false negatives).

* indicates indifference between protect and not protect (posterior = 0.25).

Parameters: $c = \$5$, $L = \$20$, $Y = \$30$. WTP calculated using risk-neutral formula.

A missing piece of the data analyses is cross-task consistency checks. The authors' experimental design contains four parts, yet they analyze each component in isolation without examining individual-level consistency across tasks. This represents a significant missed opportunity for validation, as their theoretical framework explicitly requires subjects to integrate these components coherently.

- For instance, a BE-IP consistency check could examine whether, given a threshold posterior of 0.25, a consistent decision maker who reports posterior > 0.25 in the BE task also chooses Protect" in the IP task. Across the six tasks generating 12 posteriors, the authors could identify decision thresholds for different subjects and inconsistency measures at the individual level.

- Similarly, an IP-WTP consistency check could examine whether participants who chose to protect or not protect regardless of possible signals assign zero WTP to information structures, as predicted by the assumption that access to a signal can increase expected utility if the signal affects her protection decisions." These analyses would likely reveal that high-consistency subjects show clearer false positive/false negative sensitivity patterns, while low-consistency subjects drive the null aggregate results, explaining why their current pooled analysis fails to detect the theoretically predicted effects.

In the key section "WTP and Signal Characteristics," the authors use the difference between individual WTP and risk-neutral WTP for regressions. With the decisions in other tasks, I think it makes more sense to use the difference between observed WTP and predicted WTP, which should be a function of individual risk attitude from BP, individual beliefs from BE, and individual action plans from IP. The function could be a simple OLS but could also incorporate behavioral elements that account for different biases in each component of the tasks.

For section 6, the evidence provided for the key conclusion is not convincing enough. The claim that people cannot distinguish false positive from false negative rates relies mainly on failing to find a significant difference in how people respond to FP versus FN costs, but failing to find a difference does not prove that no difference exists. It could simply mean their test lacks sufficient power or that their data is too noisy. More importantly, their approach contradicts their own theory since they average responses across all subjects and conditions, but their theory predicts that different types of people (risk-averse versus risk-neutral) should show different patterns of FP/FN sensitivity. By pooling everyone together, they make it impossible to detect the very differences their theory expects to find. The authors also never directly test whether people actually use a "total error rate" strategy but simply assume this explains their results without considering other possibilities like measurement problems or task confusion.

Contribution and Literature

The contribution is not clearly pointed out within the field of experimental studies for the demand for information sources. For instance, the failure of distinguishing the value of FP and FN is also implied in Xu (2022) while also accounting for the effects of risk attitudes and belief updating biases. I want to see a clearer comparison and innovation of this paper. Some other key references are missing, for instance, Masatlioglu, Orhun, and Raymond (2023), Charness, Oprea, and Yuksel (2021), Montanari and Nunnari (2023), which all study skewed information sources, and some of their skewness conditions have the same structure as the FP and FN rates studied in this paper.

Another strand of the literature missing is ROC (Receiver Operating Characteristic) analysis, which is a statistical method used to evaluate the performance of binary classification models, particularly in radiology, diagnostic testing and machine learning. The structure is relevant. I hope there would be more exploration of how the results in this paper relate to findings in this literature. Please see Fawcett (2006) for a review.

Minor Comments

Please indicate significance of the estimated parameters in your regression tables for better reading experience.

On page 1, the sentence “While the optimal threshold depends on user preference over the costs of these probabilistic errors, currently there is no guidance on what this threshold might be beyond assuming that decision-makers weigh false-positive and false-negative costs equally” is not consistent and not true. The optimal threshold depends on the costs of the false-positive and false-negative probabilistic errors, which in most cases are not weighted equally, such as in medical tests where the consequence of missed detection is weighted more than the consequence of false alarm.

On page 8, the IP task description is not clear. It states “subjects must make a protection decision given the prior probability of drawing a black ball. Subjects learn a prior and signal’s accuracy,” but subjects should make two decisions about protection, each for a possible hint.

On page 9, there is a concerning inconsistency between the belief elicitation mechanism described in the paper and the actual experimental instructions given to subjects. The paper claims to follow "the stochastic version of the Becker-DeGroot-Marshak mechanism developed by Grether (1992) and Holt and Smith (2009)" with a specific mathematical formulation involving uniform random draws and conditional losses. However, the experimental instructions describe a completely different mechanism that simply tells subjects "you make more money if your guess is closer to the actual probability of the event" with examples using different probability values and payoff calculations that bear no resemblance to the theoretical mechanism described in the paper.

This discrepancy raises serious concerns about the validity of the belief elicitation data. Moreover, the instructions themselves appear potentially confusing to subjects, particularly the statement about "the actual probability of the ball being black is 10 percent" when each individual ball is deterministically either black or white. If the goal is incentive-compatible belief elicitation where accuracy is rewarded, it would be much clearer to implement a proper scoring rule that directly penalizes prediction errors rather than using this convoluted mechanism that appears to differ between the paper's description and the actual implementation.

Given the central finding on asymmetric sensitivity depending on the prior probability, I think the comparative statics of the sensitivity for different priors is missing in this paper. It is easy for the risk-neutral benchmark but is not straightforward with risk attitudes.

On page 14, "We define updating errors as the difference between the subjects' elicited belief and the actual posterior probability of drawing a black ball for a given signal" will be understood by most people as Updating Errors = Elicited Belief - Actual Posterior Probability, which is not consistent with your figures and follow-up descriptions.

There are also some ambiguous pronoun references throughout the paper. On page 15, This suggests that subjects understand the signal structure but have difficulty implementing the optimal strategy" leaves unclear what This" refers to, since the preceding sentence discusses multiple results.

Similarly, on page 17, These results are consistent with our hypothesis but contradict previous findings" does not specify which results or which previous findings. Page 18 references low WTP conditions" and "high WTP conditions" without clearly defining the threshold or providing descriptive statistics.

References

Charness, G., Oprea, R., & Yuksel, S. (2021). How do people choose between biased information sources? evidence from a laboratory experiment. *Journal of the European Economic Association*, *19*(3), 1656–1691.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Masatlioglu, Y., Orhun, Y., & Raymond, C. (2023). Intrinsic information preferences and skewness. *American Economic Review*, *113*(10), 2615–2644.

Montanari, G., & Nunnari, S. (2023). Audi alteram partem: An experiment on selective exposure to information.

Xu, Y. (2022). Revealed preferences over experts and quacks and failures of contingent reasoning. *Working Paper*.