## Summary of the paper

The paper investigates how subjects' willingness to pay for a binary signal changes with the false-positive (FP) and false-negative (FN) rates before they take protective action (insurance). The main result is that, first, subjects' beliefs and protection actions change with FP and FN rates even if they should not, i.e., Table 2 and 3 shows that subjects' protection actions and beliefs change with FP when they receive a white (negative) signal and FN= 0. Second, the sensitivity of subjects' WTP for the binary signal w.r.t. FP and FN rate are remarkably similar even as priors vary, as shown in Figure 5. Combining both suggests that subjects do not distinguish (enough) FP and FN when they compute their WTP beliefs and decide their protection actions.

## Comments

I think the paper is well-written, and the experiment is well-executed. The results are novel and have implications for the literature and policies regarding the design of warning signal systems and the associated bias. The authors have done a good job of motivating the research questions. Here are my comments.

1. Empirical analysis of how future protection actions affect WTP:

   (a) I believe that, compared to existing studies on the demand for information, the most novel element of this paper is the inclusion of protection actions. As the authors mentioned, the existing literature employs prediction games; thus, states are usually symmetric, and guessing the underlying state is relatively trivial. Here, the protection action is asymmetric, and protection decisions become non-trivial, as we have known in the literature on choice under uncertainty. As future protection actions affect individuals' WTP for information, the presence of protection action would have a non-trivial effect on subjects' WTP.

   (b) However, the current version of the paper does not explore much of the relationship between protection actions and WTP. Therefore, I encourage the authors to investigate how the protection actions observed in the experiment affect WTP, which sets the paper apart from the existing literature. The analysis would also lead to new implications. Intuitively, WTP for signals is roughly affected by two factors: 1) the understanding/preference (etc.) over information 2) anticipated protection action taken by the subject's future self. As shown in Table 7, subjects exhibit failure to

distinguish FP and FN even when choosing their protection actions. Is the equal sensitivity of WTP w.r.t. FP and FN driven by rational anticipation of the "bias" in protective choice? Or Is it driven by the heuristic when subjects compute the expected benefit of getting the signal? Depending on the answer, the result will have different policy implications for encouraging the acquisition of warning signals.

2. Some re-ordering of the results, and exploring Figure 5:

   (a) Result 2 on page 18 is interesting. However, it is a comparison of subjects' WTP with the risk-neutral benchmark, and it is unclear what implications we can draw from it. One could also question whether risk neutrality is a meaningful benchmark. I understand the novelty and importance of the results only until the end of the paper, where the authors discuss the subjects' failure to distinguish FP and FN rates. Figure 5 on page 22 is particularly clear and striking. I suggest that the authors move the discussion of subjects' failure to distinguish FP and FN rates to right after result 2, and frame it as an analysis on the underlying mechanism. The rest of the discussion could be framed as a section discussing/ruling out alternative explanations.

   (b) Similarly, given the importance of Figure 5, it would be nice if the authors could include confidence interval of the regression coefficients, and present in more details the regression specification.

   (c) Figure 5 shows the sensitivity to FP and FN are similar and unchanged as prior changes. Does that mean fixing FP and FN, WTP is unchanged in prior? That would violate the intuition that the value of information decreases as the prior becomes more extreme. As I suggest the authors increase the spotlight on Figure 5, it would be a good opportunity to present some analysis on the reported WTP instead of the difference between the reported WTP and the risk-neutral benchmark.

3. Probability weighting/overweigh (resp. underweigh) small (resp. high) probability

   (a) Another potential explanation of the results is probability weighting. If somehow subjects exhibit probability weighting to the extent that probability compresses towards 0.25 for all priors, then sensitivity to FP and FN would be equal and is invariant in prior. I don't think this explanation is realistic, but some discussions (to rule it out) would be nice.

4. I find it unclear how subjects are allocated to different treatments. The authors write in the 2nd paragraph page 10: " Subjects go through two different priors and three types of signals". But on Table 1 page 10, there are total 4 types of priors and 7 types of signals. Are subjects randomly allocated to 2 out of the 4 types of prior and 3 out of the 7 types of signals? But the authors write in the 2nd paragraph, page 10: "Subjects go consecutively over all three signal types starting from the honest one for each prior.". That means one of the three signals is from the honest-only composition (3-0-0) for all subjects, and subjects are not randomly assigned to 3 out of the 7 types of signals? What are the sample sizes for each treatments?

5. I am also confused about Table 2 and 3, especially, column 4, the posterior. It seems that the authors have pooled the different 4*7 treatments into 8 different groups, based on the presence of FP and FN, and the hints received by the subjects. The authors write on page 12 that "Column 4 shows the posterior probability of a black ball averaged across all the treatment within a group" How are the posteriors calculated? Is it based on the empirical data, or is it some weighted average of Bayesian posteriors? Similarly, are the share optimal calculated based on the empirical data or the Bayesian posteriors?

6. Result 2 on page 18:

   The authors write: "subjects tend to overvalue false-negative costs for low probability events and overvalue false-positive costs for high probability events." Where do we see that in Table 5? The coefficients of FP costs, FN costs on column 4 and 5 are all positive. Should it be "subjects tend to overvalue more false-positive costs (coeff:0.800 vs 0.204) for low probability events and overvalue more false-negative (coeff: 0.407 vs 0.150) costs for high probability events."? When comparing coefficients, the authors should also report results in statistical tests.

7. Literature:

   The authors cited two laboratory experimental papers on the demand for information. I believe there are many more. I do not think the missing literature diminishes the contribution of this paper, but citing them would highlight the novelty of the paper's setup and better position the paper in the existing literature. Here are some examples:

   (a) Intrinsic Information Preferences and Skewness by Yusufcan Masatlioglu, Yeşim Orhun, and Collin Raymond

(b) How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment by Gary Charness, Ryan Oprea, Sevgi Yuksel

(c) Beyond Value: on the Role of Symmetry in Demand for Information by Aniol Llorente-Saguer, Santiago Oliveros, Ro'i Zultan

8. Discuss implications on the design of warning signals:

   I think one of the low-hanging fruits to improve the paper's contribution is to provide more discussion on policy implications related to warning signals. At the moment, the authors only discuss the papers' contribution in the introduction by saying that "currently there is no guidance on what this threshold might be beyond assuming that decision-makers weigh false-positive and false-negative costs equally." Throughout the rest of paper, the authors did not discuss any implications/guidance on the design of warning signals systems.

9. The framing of multiple gremlins:

   The use of multiple gremlins frames both FP and FN as "dishonest". Thus, one would imagine that subjects refrain from trusting/paying "dishonest" gremlins, leading to equal sensitivity to FP and FN. The authors could consider running new sessions that frame signals as generated from one imperfect gremlin/expert that has asymmetric tendencies to send false positive and false negative signals.

10. Exposition:

    (a) There are no "stars" in Table 5 onwards. Is that intentional? I find it impossible to calculate the p-values for every coefficient.

    (b) In various parts of the paper, the authors have used the terms "true posterior" and "accurate posteriors" (for example Table 2 and 3, the bottom of page 14, and many others.). Do they refer to Bayesian posterior or empirical posterior?

    (c) Page 18 after the regression equation: how exactly are FP cost and FN cost calculated? It is confusing to see FP and FN sometimes refer to rate and sometimes refer to cost.

11. Typo:

    (a) Extra "=" sign in Equation (1) page 5.

    (b) Missing a fullstop at the end of footnote 7 page 15.