Conference Submission Review – Ashley Herman

This work presents a study on how users might be persuaded against over-relying on a model that performs customer classification. The authors highlight the effectiveness of nudging users to evaluate their decisions against the model's while uncovering some interesting discrepancies between users' self-perception and actual behaviour. They found that their system, which prompts users to think critically about model results, improved decision-making and reduces over-reliance except in the case where there is low confidence in the system. Given the prevalence of AI, the researchers have chosen to address an important and timely question: how might designers encourage critical thinking and prevent over-reliance on AI? Additionally, they have judiciously chosen to focus on ecological validity in their experimental design. The writing style is precise, concise, and objective. The contribution –a technique for improving critical thinking while using AI – is very clear. However, there are gaps in the methodology and subsequent results that will need to be addressed for the paper to be accepted. The authors do not mention ethics approval and have not included supplemental materials such as the questionnaire that was used in the study, which in turn leads to questions about what exactly is being measured in the reported results. There are some places where the metrics being reported are unclear or unexplained, and common limitations, such as a small sample size, are never acknowledged. This work could be accepted with necessary revisions, as outlined below.

The strengths of the submission:

- Given the prevalance of AI for business use, the question of overreliance the authors address is important and timely.
- The researchers are deliberate in their attempt to address gaps in prior research by choosing a specific domain and testing with real users (while simulated users / crowd workers have been relied upon in the past).
- Their study is highly transferrable to a real-world scenario and likely has high ecological validity.
- The methodology includes both quantitative (variables measured during the experiment) and qualitative measures (questionnaire).

The weaknesses of the submission, which need to be addressed, are:

- When explaining the Random Forests model that was used in the study, the authors mention it has 98% accuracy. Barring questions of overfitting, this would mean that far more examples of model correctness were used than model incorrectness. Was this, in fact, the case? Did some scenarios have a much smaller sample size than others?

- When describing the study, the authors mention a questionnaire however they do not go into any details about it. Is it a standardized questionnaire, or one that they designed, and if so, how did they design it? Perhaps the questionnaire could be included in an appendix or in supplementary materials. The authors mention that they are measuring users' self-reliance, but they do not mention how exactly it is being measured - I am assuming including the questionnaire would shed some light on this.
- In the "User Study" section, the authors explain that "Participants had above-average expertise in context (Median=4) and an average AI understanding (Median=3)." However, they decline to give any scale or frame of reference for the numbers "3" and "4". This makes the claims difficult to evaluate.
- When describing the study, the authors do not mention whether they have ethics approval.
- The study suffers from a small sample size and does not account for common factors like user fatigue in repetitive tasks. These are very common concerns in HCI studies and not at all a reason to reject the results – however, they issues should be pointed out as limitations somewhere in the discussion or the conclusion.

Below are some additional, but lower-priority, concerns:

- The graphs in figure 4 do not have a label on the y-axis.
- The authors point out that the study participants have diverse experience with AI. Are there any other relevant demographics that could be pointed out here (how many men/women, age groups, occupations, etc)?
- For reader clarity, a very brief description of units (such as minutes or the scale being used) could be added to the means and standard deviations wherever they are reported.

Recommendation: +1 Accept, with revisions as outlined above.

Confidence in the recommendation: 2

My expertise in the area: 2