AI Accuracy & Human Trust in Human Decision Making Processes

Team Members:
Ruowen Xu & Yucen Zhong

Key Research Questions

1. (Paper 1) How some key properties affect humans' mental models of AI capabilities and the resulting team performance?
2. (Paper 2) Does laypeople's trust in a model vary depending on the model's stated accuracy on held-out data and on its observed accuracy in practice?
3. (Paper 3) How does showing AI's prediction versus not showing, affect trust, accuracy of AI-assisted predictions, and the effect of confidence score on trust calibration?
4. (Paper 3) How does knowing to have more domain knowledge than the AI affect humans' trust, accuracy of AI-assisted predictions, and the effect of confidence score on trust calibration?

Paper 1 - Beyond Accuracy: On the Role of Mental Models in Human-AI Teams

- Research Question: How do the following properties affect human's mental models of AI capabilities and the resulting team performance?
  - Two Key Properties of an AI's Error Boundary:
    - Parsimony
    - Stochasticity
  - One Property of Task:
    - Dimensionality

Paper 1 - AI-advised Human Decision Making

- Examples
  - Medical Diagnosis
  - Candidate screening for hiring
- Sequences
  - S1: The environment provides an input, x
  - S2: The AI (possibly mistaken) suggests an action, h(x)
  - S3: Based on this input, the human makes a decision, u
  - S4: The environment returns a reward, r
    - a function of the user's action
    - the (hidden) best action
    - other costs of the human's decision (e.g., time taken)

Paper 1 - Error Boundaries and Mental Models

f: (x, h(x)) ->{T,F} m: *x' -> {T, F}*

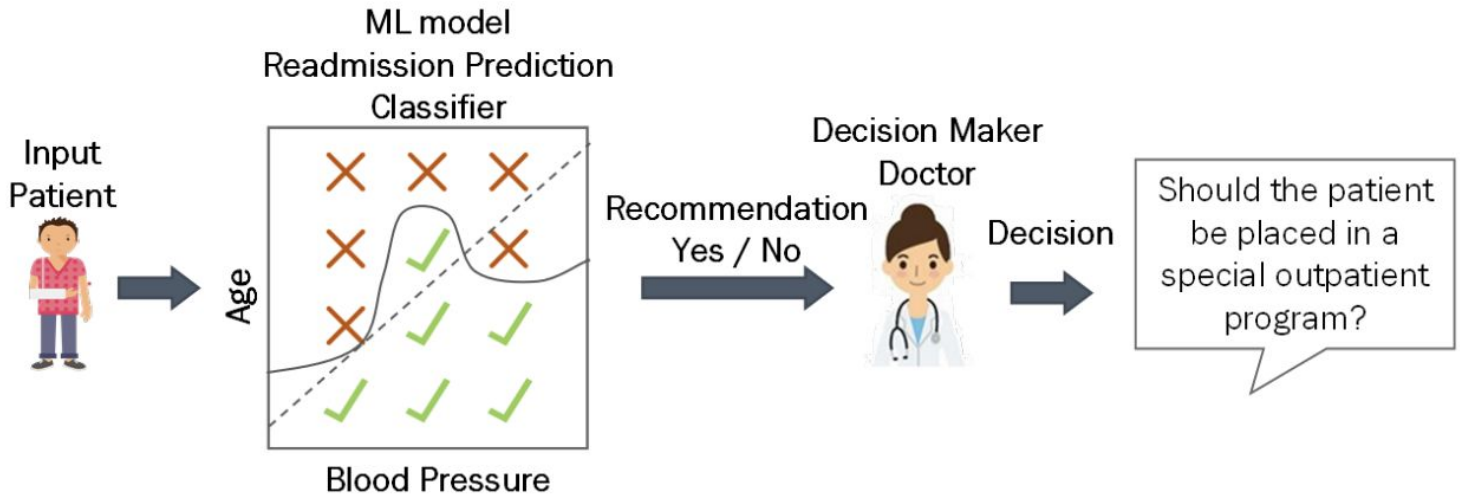

Figure 1: AI-advised human decision making for readmission prediction: The doctor makes final decisions using the classifier's recommendations.

Paper 1 - Characteristics - Parsimony

- Inversely related to the representational complexity of the error boundary
- Solid line

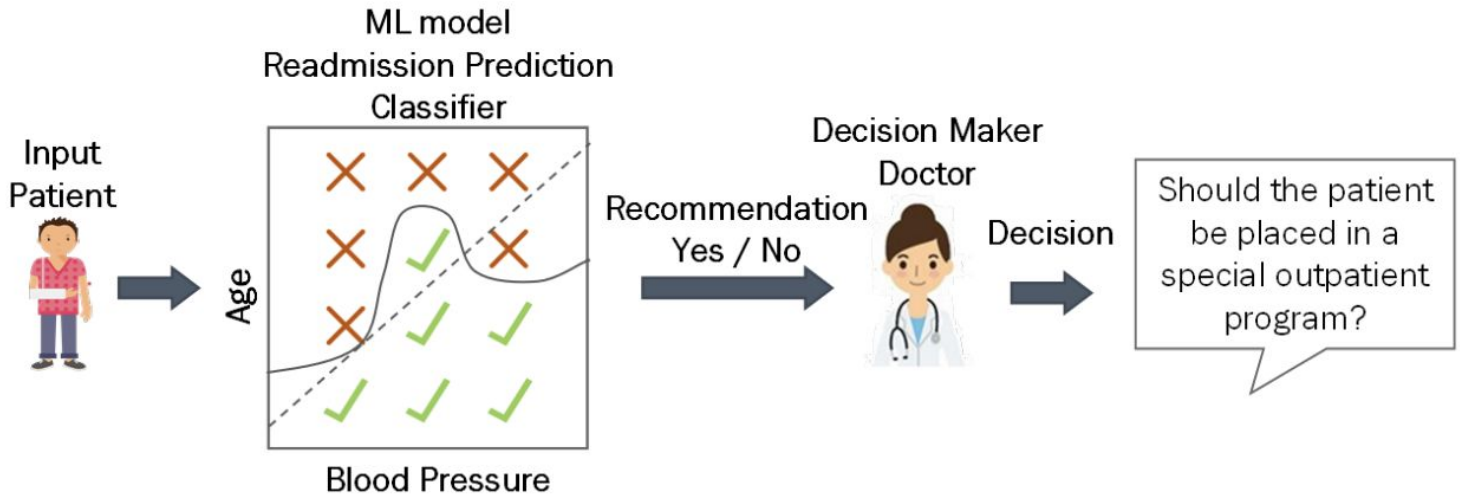

Figure 1: AI-advised human decision making for readmission prediction: The doctor makes final decisions using the classifier's recommendations.

Paper 1 - Characteristics - Non-Stochasticity

- Separates all mistakes from correct predictions
- Example: f1: {age = young ∧ blood pressure = low}

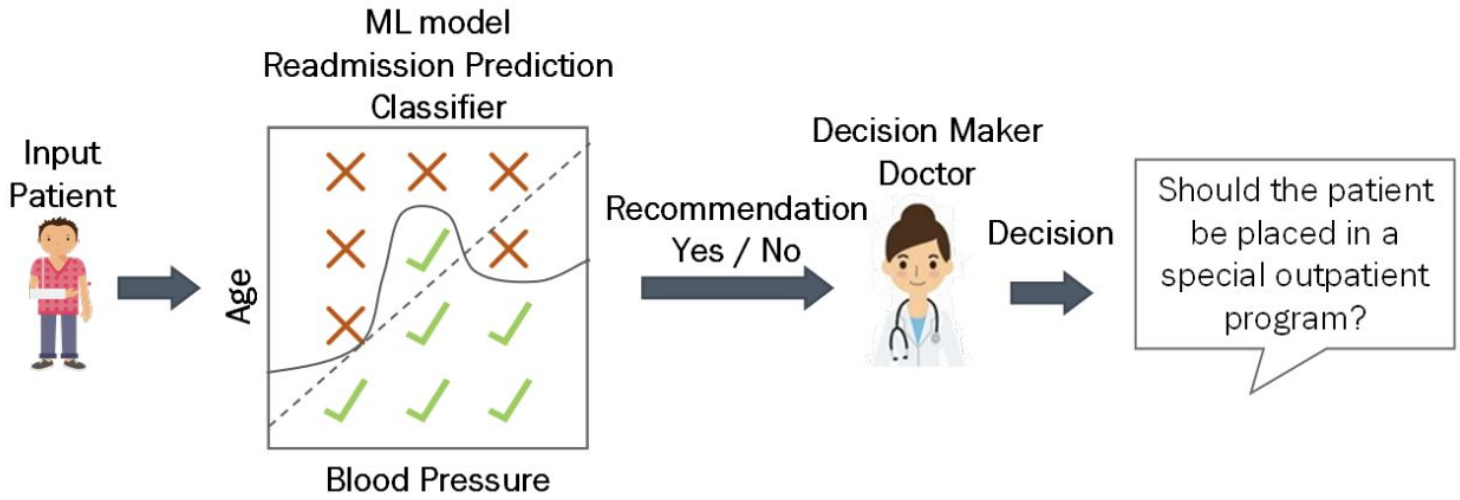

Figure 1: AI-advised human decision making for readmission prediction: The doctor makes final decisions using the classifier's recommendations.

Paper 1 - Characteristics - Task Dimensionality

- Number of features defining each instance
- Larger the better?

Paper 1 - Experiment

- CAJA
  - An open-source, game-like platform that mimics AI-advised human decision making
- Make decision
  - Whether or not the objects going over the pipeline are defective
  - High stake
- Get reward

Figure 2: For each object, a subject can either choose to use Marvin's recommendation or perform the task independently.

Paper 1 - Discussion

1. Can you think of any examples of AI applications in which you decided to accept or override the AI’s recommendation? How did you make the judgment call on whether to override the AI’s recommendation?
2. The paper mentions that the highest team performance is often reached when they both know how and when to complement one another. It focuses on one factor that is crucial to supporting such complementary, which is the human's mental model of AI’s capacity. Can you think of other factors?

Paper 1 - Experiment Continued

- Parameters relevant to AI-advised human decision making
  - Task dimensionality
  - AI performance
  - Length of interaction
  - Parsimony and stochasticity of the error boundary
  - Cost of mistakes

Results - Mental Models

- Do people create mental models of the error boundary?
- How do mental models evolve with interaction?

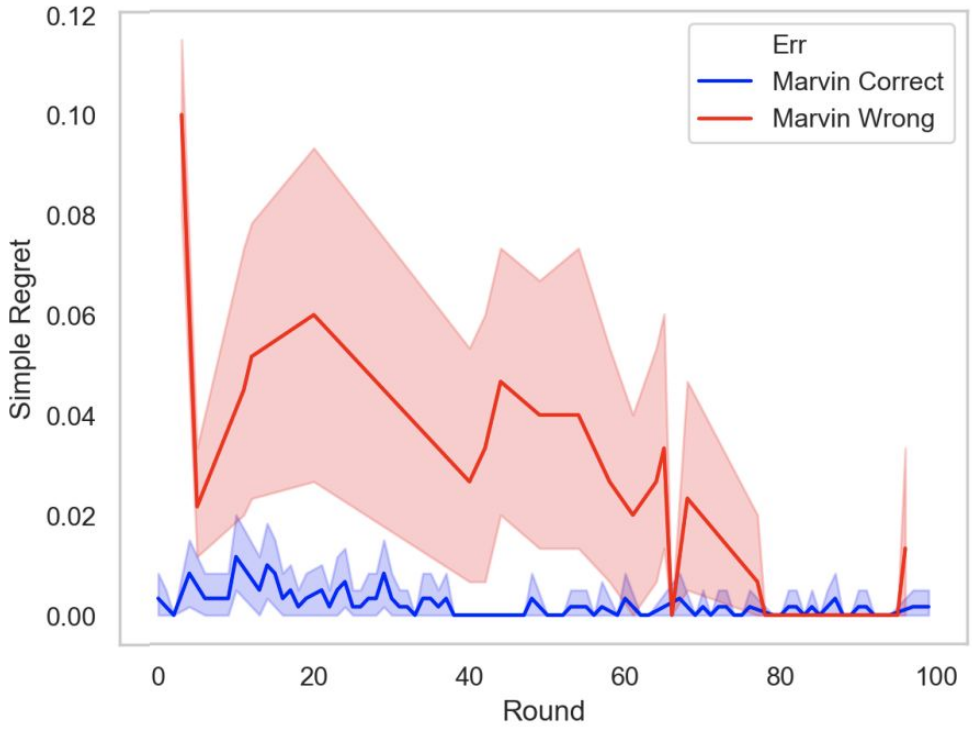

Figure 3: With more rounds of interaction, users perform closer to the optimal policy.

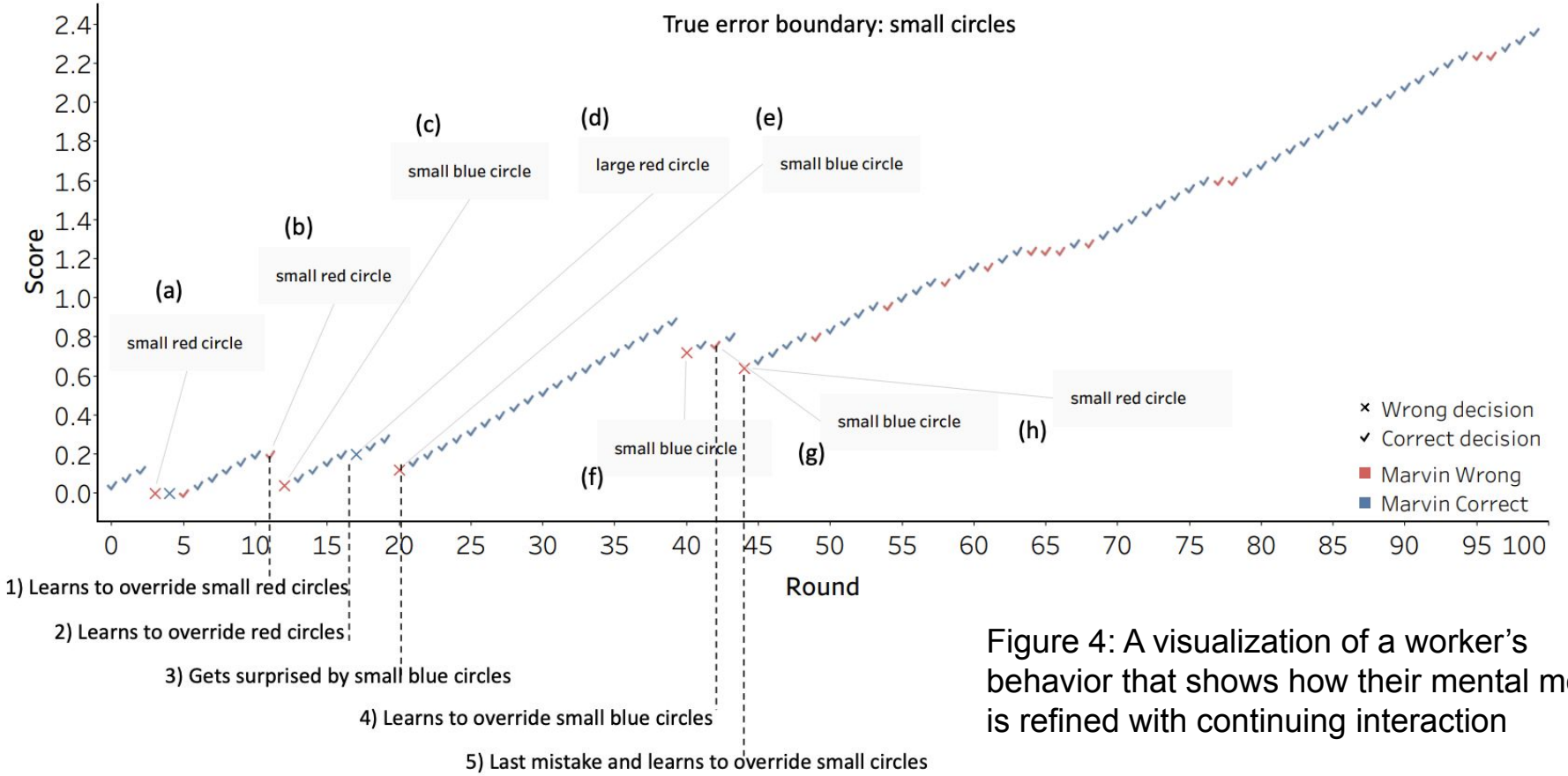

Figure 4: A visualization of a worker's behavior that shows how their mental model is refined with continuing interaction

Results - Parsimony

- Do more parsimonious error boundaries facilitate mental model creation?

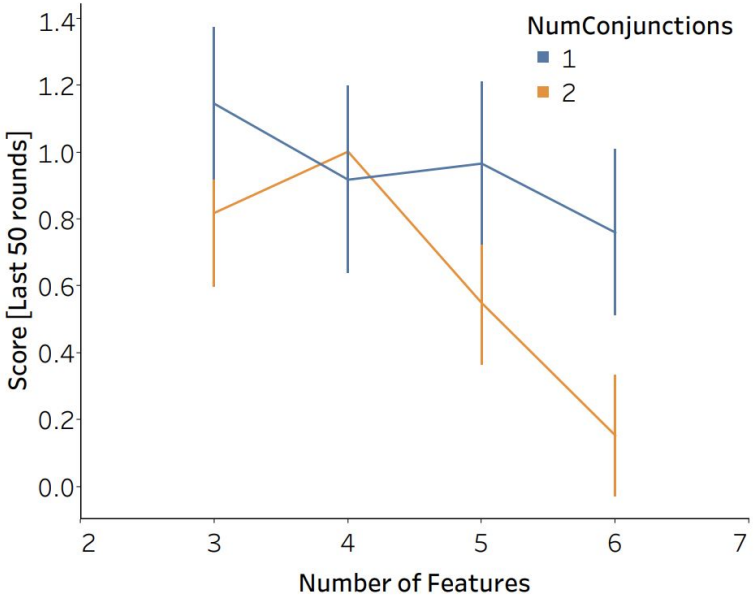

Figure 5: Team performance decreases as the number of conjuncts in the error boundary is increased. Number of literals were fixed to 2.

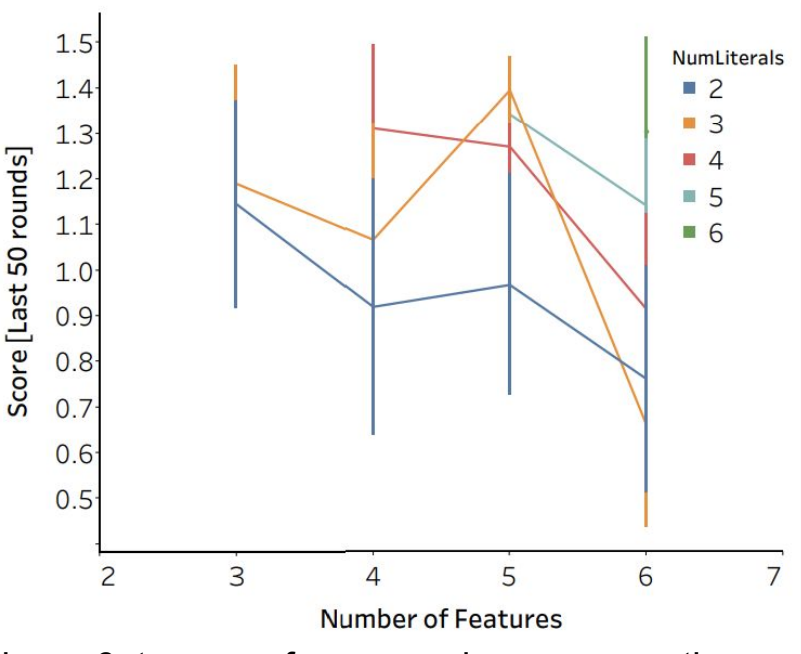

Figure 6: team performance decreases as the number of human-visible features increases, which is consistent with previous findings on human reasoning about features

Results - Non-Stochasticity

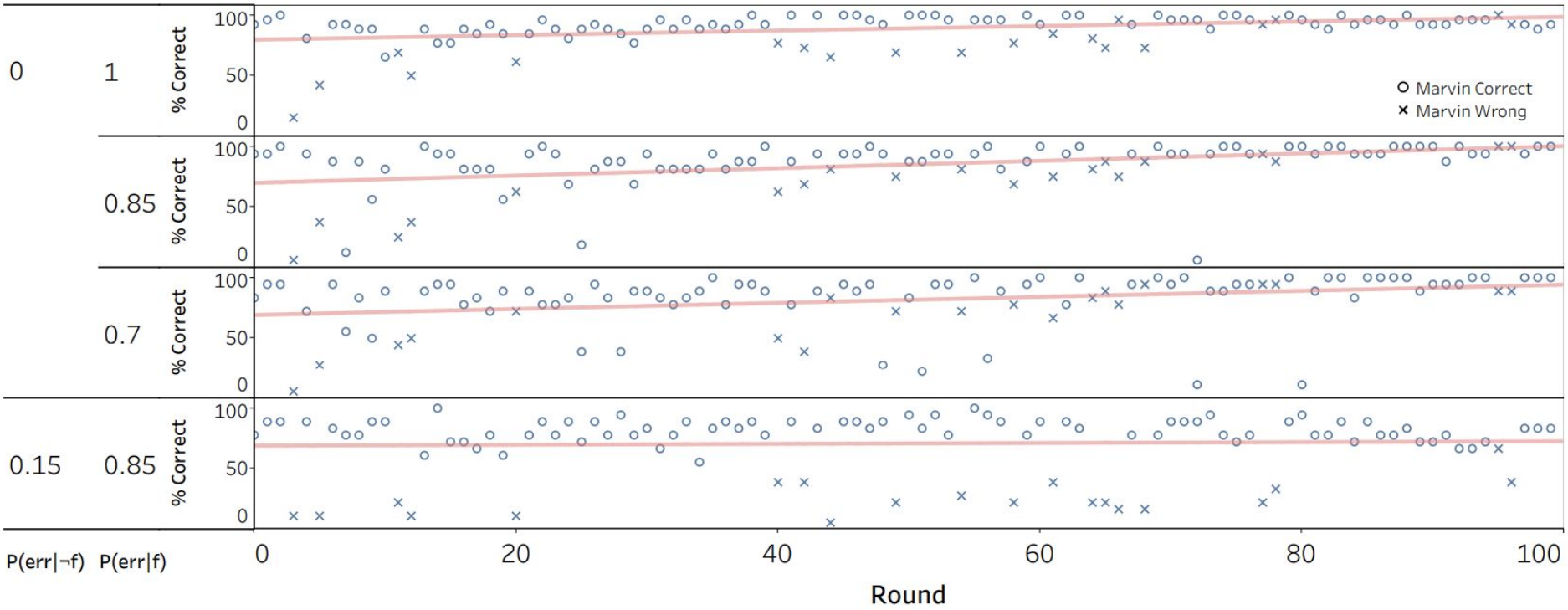

Figure 7

Recommendation

1. Build AI systems with parsimonious error boundaries.
2. Minimize the stochasticity of system errors.
3. Reduce task dimensionality
4. Deploy models whose error boundaries are backward compatible, i.e. by regularizing in order to minimize the introduction of new errors on instances where the user has learned to trust the system.

Paper 2 - Understanding the Effect of Accuracy on Trust in Machine Learning Models

- Research question: Does laypeople's trust in a model vary depending on the model's stated accuracy on held-out data and on its observed accuracy in practice?
- New challenge
- Growing concerns
  - Societal biases
- Resolution
  - Interpretable machine learning
  - Existing work
- What is missing?

Paper 2 - Experiments

- Experiment 1: Does a model's stated accuracy affect laypeople's trust?
  - Does a model's stated accuracy on held out data affect laypeople's trust in the model
  - If so, does it continue to do so after they have observed the model's accuracy in practice?
- Experiment 2: Does this effect change if the observed accuracy is low/high?
- Experiment 3: Does a model's observed accuracy affect laypeople's trust?

Paper 2 - Results

- A model's stated accuracy on test data affects people's trust in the model
- Effect size is smaller after people observe the model's accuracy in practice
- If a model's observed accuracy is low, then its stated accuracy has at most a very small effect on people's trust in the model
- People's trust in the model can be significantly affected by its observed accuracy regardless of its stated accuracy
- After observing a model's accuracy in practice, people are more likely to increase their trust in the model if the model's observed accuracy is higher than their own accuracy
    - One exception

Paper 2 - Limitation & DIscussion

- Measured people’s trust in terms of both behavioral measures (i.e., agreement fraction and switch fraction) and self-reports
- However, due to the experimental design (in particular, the difference between a model’s stated and observed accuracies), it is possible that trust is affected by other factors such as surprise, confusion, and cognitive dissonance, all of which may mediate the effect of accuracy
- How do you think trust should be measured?

Paper 3 - Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making

- Research Question 1: How does showing AI's prediction versus not showing, affect trust, accuracy of AI-assisted predictions, and the effect of confidence score on trust calibration?
- Research Question: How does knowing to have more domain knowledge than the AI affect humans' trust, accuracy of AI-assisted predictions, and the effect of confidence score on trust calibration?

Paper 3 - Experiment 1 - Effect of Showing AI Confidence Scope

- Hypothesis 1 (H1): Showing A confidence score improves trust calibration in AI such that people trust the AI more in cases where the AI has higher confidence
- Hypothesis 2 (H2): Showing A confidence score improves accuracy of AI-assisted predictions

Paper 3 - Experiment 1 - Experimental Design

- Participants could achieve comparable performance to an AI model
- This task served as the foundation for both the first and the second experiment

Paper 3 - Experiment 1 - Results

- H1 supported
  - Confidence score improved trust calibration and increased people's willingness to rely on AI's prediction in high confidence cases
- RQ1:
  - This trust calibration effect held in AI-assisted decision scenarios where the AI's recommendation was shown, and in scenarios where people had to make blind delegation without seeing the AI's recommendation
- H2 rejected
  - In this case study, trust calibration did not translate into improvement in AI-assisted decision outcome, potentially because there was not enough complementary knowledge for people to draw on
- RQ2:
  - While we explored a scenario where participants knew they had additional knowledge that the AI did not have access to, it did not make significant difference in the AI-assisted prediction task

Paper 3 - Experiment 2 - Effect of Local Explanation

- Same setup
- Hypotheses:
- H3: explanation could support trust calibration
- H4: Explanation could improve AI-assisted predictions

Paper 3 - Experiment 2 - Results

- H3
- H4

Paper 3 - Experiment 2 - Limitations

1. Participants are not experts
2. Contrived prediction
3. Dependance on the model’s predicted probabilities being well calibrated to the true outcome probabilities

Paper 3 - Discussion

- Anything you can think of that might help cope with the limitation in futures studies?
- What happens when humans do not have enough unique knowledge when AI tries to use human knowledge to make decisions? What remedies can you think of to improve AI-assisted decision making?

Papers Mentioned

Beyond Accuracy: On the Role of Mental Models in Human-AI Teams. Bansal et al. HCOMP 2019.

Understanding the Effect of Accuracy on Trust in Machine Learning Models. Yin et al. CHI 2019.

Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Zhang et al. FAT* 2020.