# Selected Recent Topics: Human-AI Team (1)

Robert Kasumba, Vishesh Patel, Isabelle Hren, Jake Kosowsky

Washington University in St. Louis

# When does an AI system fail?

**Does it even matter?**

# Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, Eric Horvitz

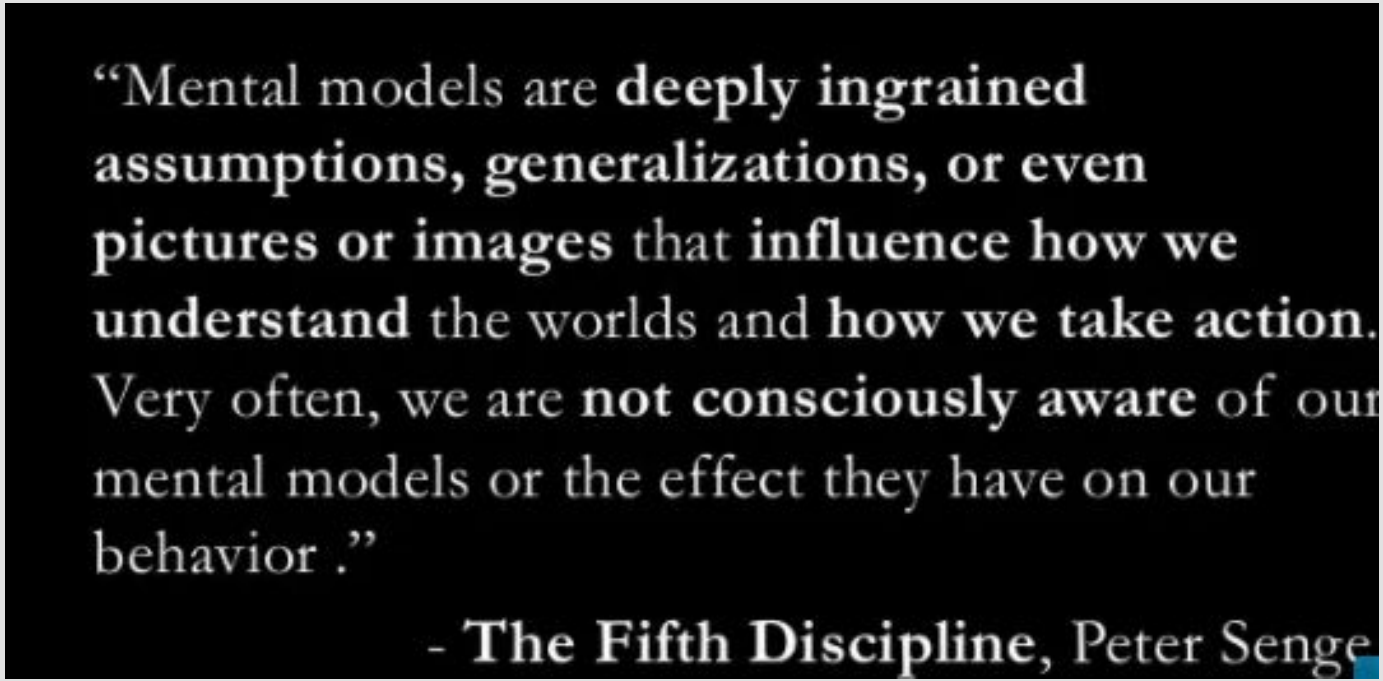
# Background

There are problems that humans or machines cannot solve alone. For example recidivism prediction, medical decisions, credit assessment etc.

For best performance, Humans and AI need to work together- AI accuracy does not always translate to end-to-end team performance.

- In most cases AI developers optimize solely for accurate

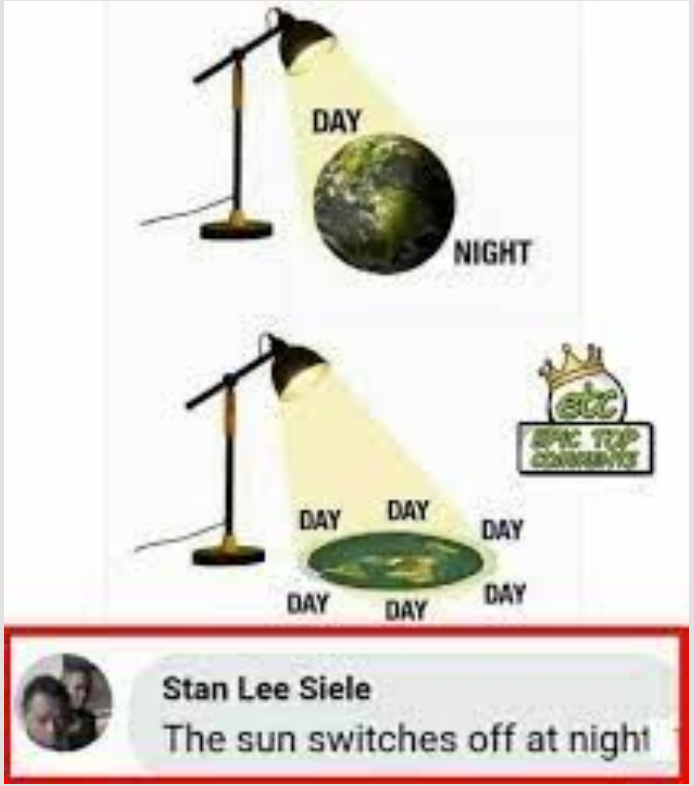
# Mental Models




*Source: https://www.slideshare.net/Managewell/mental-models-28761019*

# Humans will always create mental models...

- Of course some of will be wrong or inconsistent with further observations



# Why Mental Models? Teamwork

- Humans construct insights about the performance of the AI. This helps to know when we should trust it's result and when to override it.
  - When does the AI fail?
  - Whats its **error boundary**?
  - Which kind of inputs or states affect its performance?
- What if a humans misunderstands the error boundary of the AI?
  - Similar to not knowing your teammates weakness.
  - Cost decisions. Human may trust when they shouldn't and mistrust when they should.
  - Lower productivity

# AI Error boundary

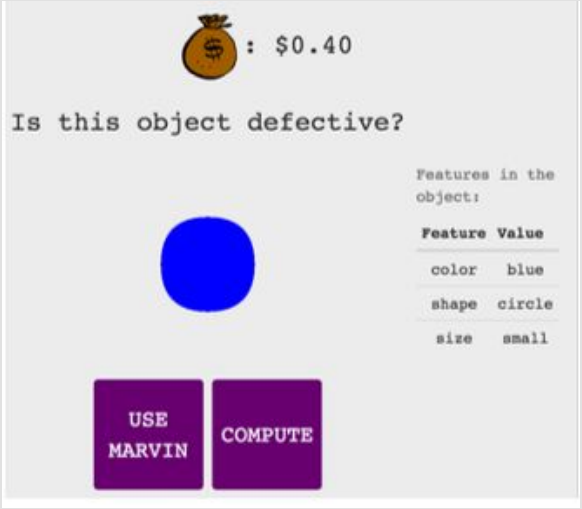
- It a function that defines the input for which the model is correct.

# AI Error boundary: Key Properties

- **Parsimony**: How complex the actual AI model is?
  - Consider an error boundary that only evaluates **conjunction** of two literals e.g. (A **and** B)
- **Stochasticity**: How deterministic or non-deterministics the models outputs are for a given input?
  - Think about a model that on a specific input x1 it's 90% likely to predict correctly and 10% wrongly but on x2 it's 60% likely to be wrong.
  - Another model guaranteed to be correct on x1 input but wrong on x2
- **Task Dimensionality**: How many features are used in the model's prediction?
  - Higher dimensionality tends to higher accuracy.

# Experiment

- AMT workers were recruited to play a game.
- They chose whether to trust or override the result given by AI (Marvin)
- After each round they receive a pay off with a high penalty for accepting wrong recommendation.
- The Turkers where assigned to groups which controlled for parsimony, stochasticity, task dimensionality.
- Experiment set up such that the only way of maximizing pay off is **learning the error boundary.**



# Payoff Matrix

| | Marvin Correct | Marvin Wrong |
|---|---|---|
| Accept | $0.04 | -$0.16 |
| Compute | 0 | 0 |

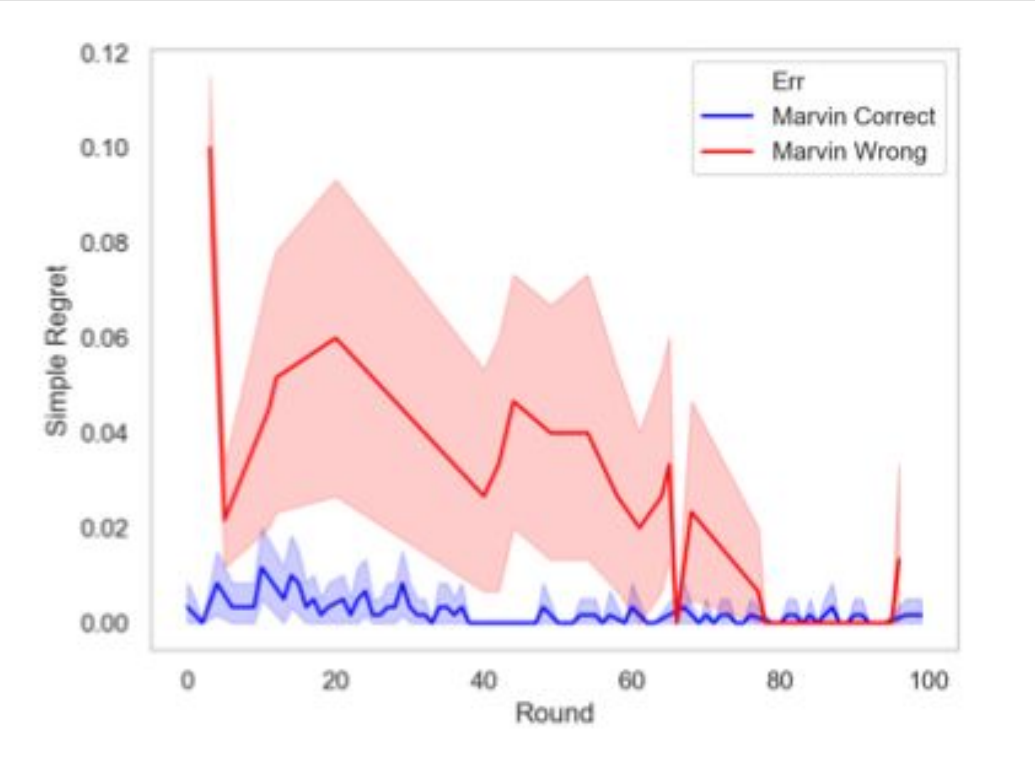
# Research questions:

***Q1: Do people create mental models of the error boundary? How do mental models evolve with interaction?***

*Q2: Do more parsimonious error boundaries facilitate mental model creation?*

*Q3: Do less stochastic error boundaries lead to better mental models?*

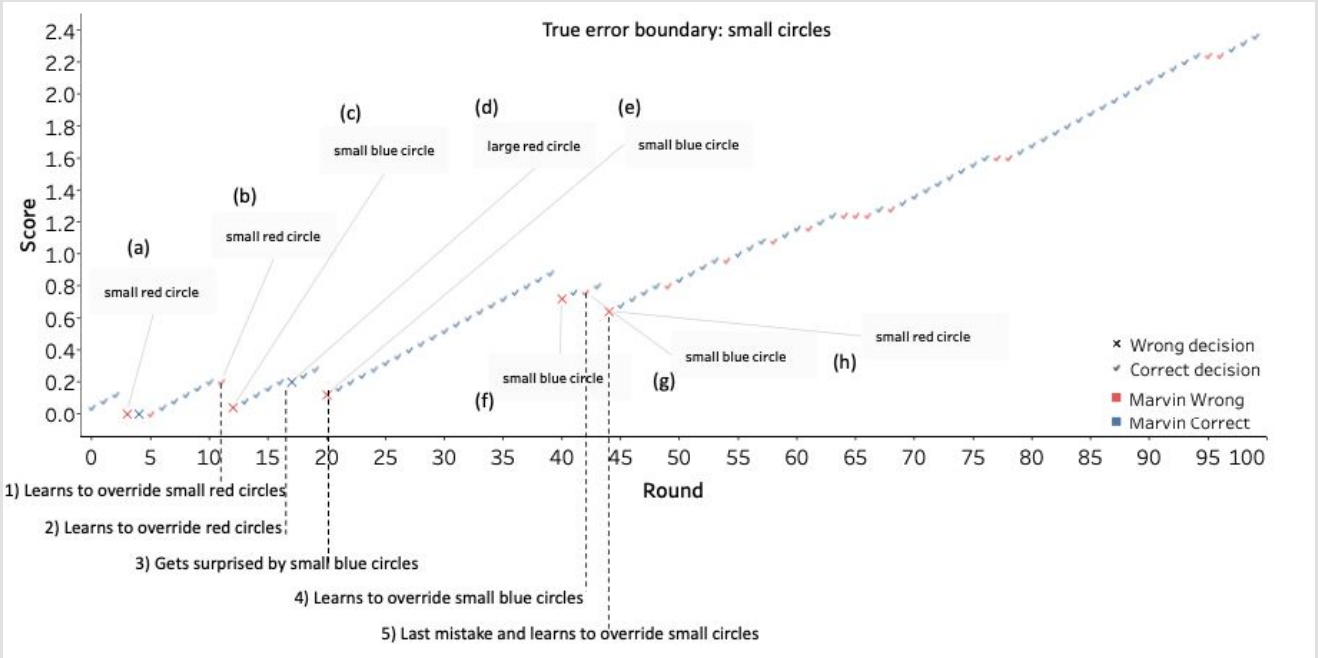
***Q1: Do people create mental models of the error boundary? How do mental models evolve with interaction?***




- The difference between “optimal” payoff and actual payoff reduces as the turkers perform more rounds.
- They are learning Marvin’s error boundary.

# *Q1: Do people create mental models of the error boundary? How do mental models evolve with interaction?*

**Error boundary Condition**: non-stochastic, on conjunction (two literals), task dimensionality as 3.




Evolution of a single given worker

# Check in Quiz:

When does Marvin fail?

A. Red circles
B. Small Red Circles
C. Blue Circles
D. Small Circles
E. All Circles

# Check in Quiz:

When does Marvin fail?

A. Red circles
B. Small Red Circles
C. Blue Circles
**D. Small Circles**
E. All Circles

***Q1: Do people create mental models of the error boundary? How do mental models evolve with interaction?***

- Mental models may suffer from overfitting or overgeneralization.
  - Overgeneralization:
    - E.g: Marvin is always wrong or fails on all colors.
  - Overfitting:
    - E.g: Marvin fails on red small circles.

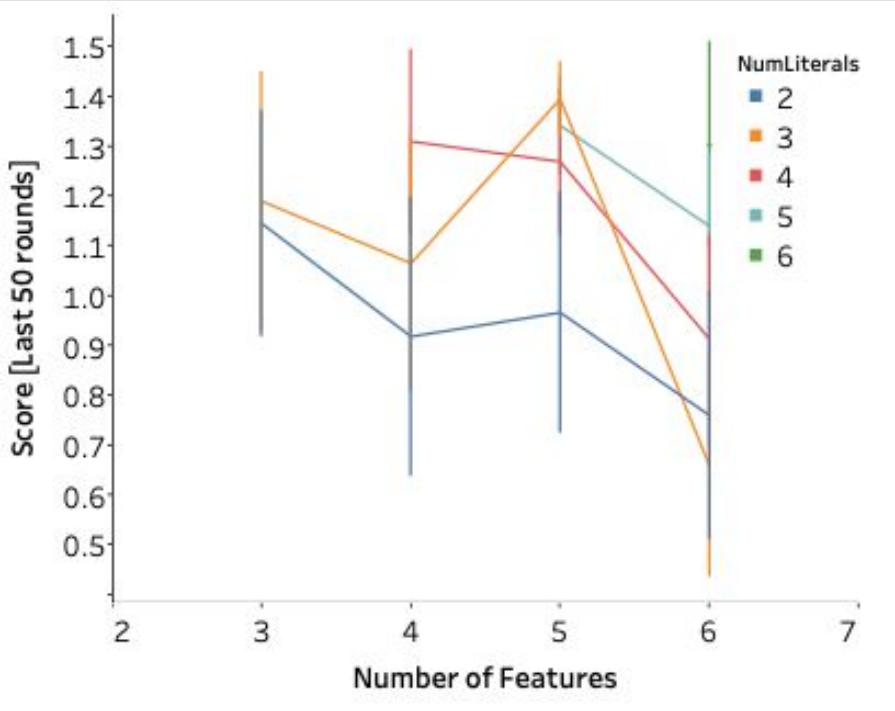
# Research questions:

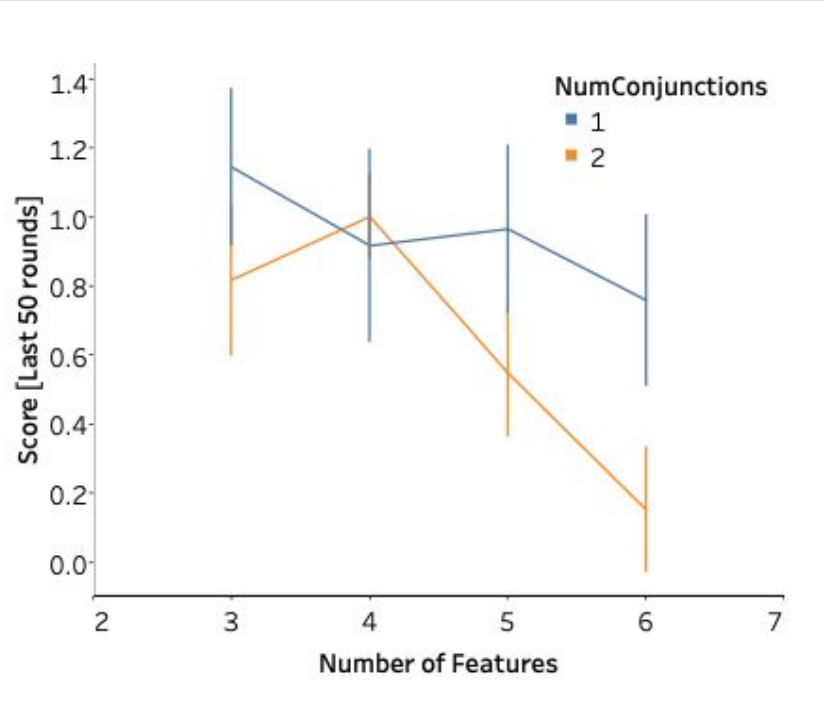
*Q1: Do people create mental models of the error boundary? How do mental models evolve with interaction?*

***Q2: Do more parsimonious error boundaries facilitate mental model creation?***

*Q3: Do less stochastic error boundaries lead to better mental models?*

## *Q2: Do more parsimonious error boundaries facilitate mental model creation?*



# Research questions:

*Q1: Do people create mental models of the error boundary? How do mental models evolve with interaction?*

*Q2: Do more parsimonious error boundaries facilitate mental model creation?*

***Q3: Do less stochastic error boundaries lead to better mental models?***

## *Q3: Do less stochastic error boundaries lead to better mental models?*

In Q1 and Q2, the experiments were non-stochastic i.e Marvin makes a mistake if only and only if the input satisfies a formula.

For Q3, the experiment was modified to introduce stochasticity. Two variables were varied i.e  P(err/¬f) and P(err/f) where f is the error boundary.

P(err/¬f): The probability of marvin being wrong given that the input does not satisfy the error formula f.

P(err/f): The probability of marvin being wrong given that the input satisfies the error formula f.

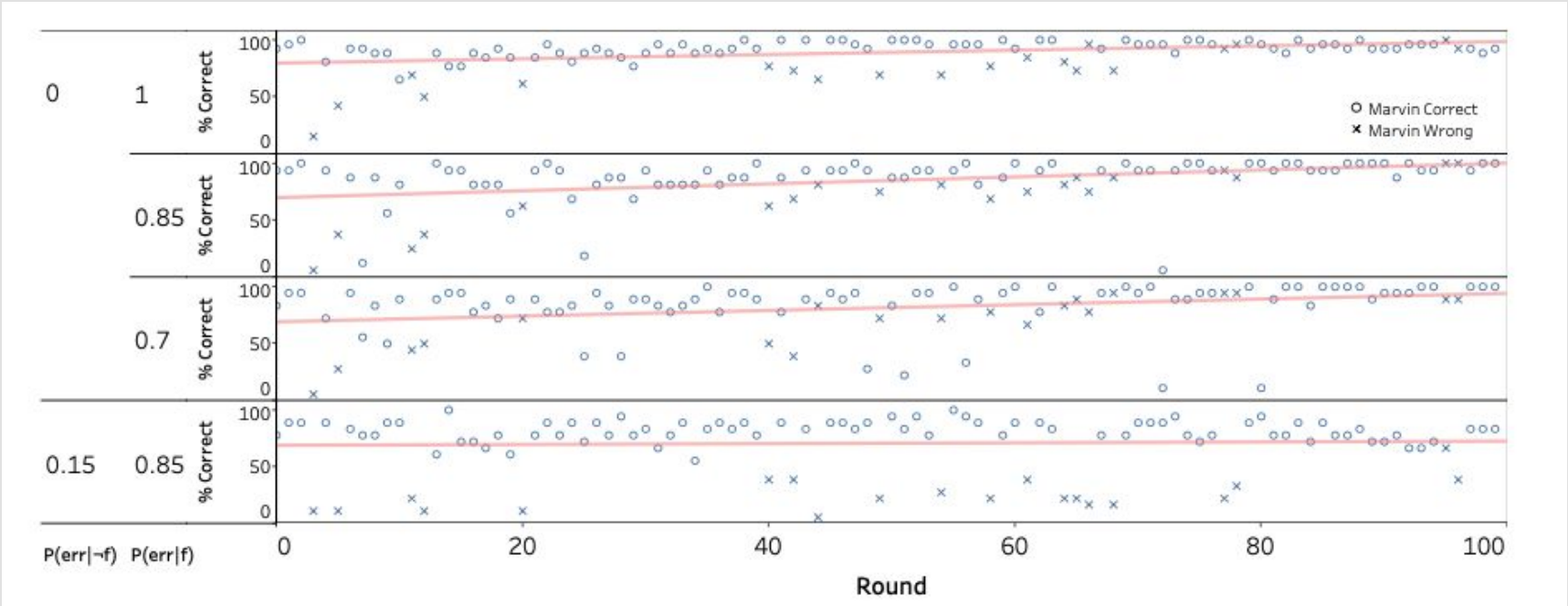
## *Do less stochastic error boundaries lead to better mental models?*

Experiment Settings with different levels of stochasticity.

| | P(err/¬f) | P(err/f) | Stochasticity |
|---|---|---|---|
| A | 0 | 1 | Non-stochastic |
| B | 0 | 0.7 | stochastic |
| C | 0 | 0.85 | stochastic |
| D | 0.15 | 0.85 | Highly stochastic |

## *Do less stochastic error boundaries lead to better mental models?*

The error boundary is harder to learn with increased stochasticity and the performance barely changes for the highly stochastic settings.



# Recommendations

1. Build AI systems with parsimonious error boundaries.
2. Minimize the stochasticity of system errors.
3. Reduce the task dimensionality when possible (e.g  by removing features that are not very important)
4. Deploy models whose error boundaries are backward compatible to previous models to leverage the experience of users that has been gained in the past.

# Discussion

- One example of AI-Human team applications is the Doctor and recommendation system. Can you think of another AI-Human team that relates to you? Think of an example where you use a recommendation system and when you override or take its result.

- One key property presented by the paper is parsimony which relates to how complex the model is. Do you think more parsimonious models (less complex) but less accurate should be preferable to more accurate and less parsimonious ones for successful AI-human collaboration?

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

Yunfeng Zhang, Q. Vera Liao, Rachel K. E. Bellamy

# Background and Research Question

- AI/ML is probabilistic -> no guarantee of correctness

- **AI-assisted decision making -** have humans use suggestions from AI

- Authors explore whether human trust in the model leads to better decision outcomes
  - Knowing in which cases when to trust or distrust the AI's predictions

- **Does understanding features about the model lead to improved outcomes?** 2 studies.

# Background (cont.)

- Many models are black boxes (we've seen examples of this in the justice system discussion) which makes fully trusting in them difficult/risky
- Users/humans need a good understanding of the AI error boundaries to form a strong mental model of when to trust the AI
- Premise: Little is known about human trust in confidence scores
- This paper is about **human trust calibration** and how **trust (or lack thereof) is key to the success of AI assisted decision-making**

# Some Terminology

- **AI assisted decision-making:** The process of using AI predictions in conjunction with human expertise to determine the best outcomes. Deciding whether to use an AIs prediction.
- **Calibrate (trust):** Helping people correctly distinguish when to or when not to trust an AI
- **Confidence score**: A score that measures whether the AI will be correct or not in an outcome
- **Local explanation:** provide the rationale for a single prediction made by the algorithm
- **Enhance (trust):** repeatedly demonstrating high performing AI will make people generally trust the power of AI more
- **Automation bias:** The humans overly rely on the algorithm's rationale
- **Algorithm aversion**: The humans stop trusting the algorithm after seeing it make a few mistakes

# Experiment One (E1): Confidence Scores

**Hypothesis 1:**
**Showing AI confidence score improves trust calibration in AI** such that people trust the AI more in cases where the AI has higher confidence.

**Hypothesis 2**:
**Showing** AI confidence **score improves accuracy of AI-assisted predictions.**

*H2 is based on assumption that if H1 holds, humans will pick the AI recommendation at the right times

# E1: Research Questions

**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?

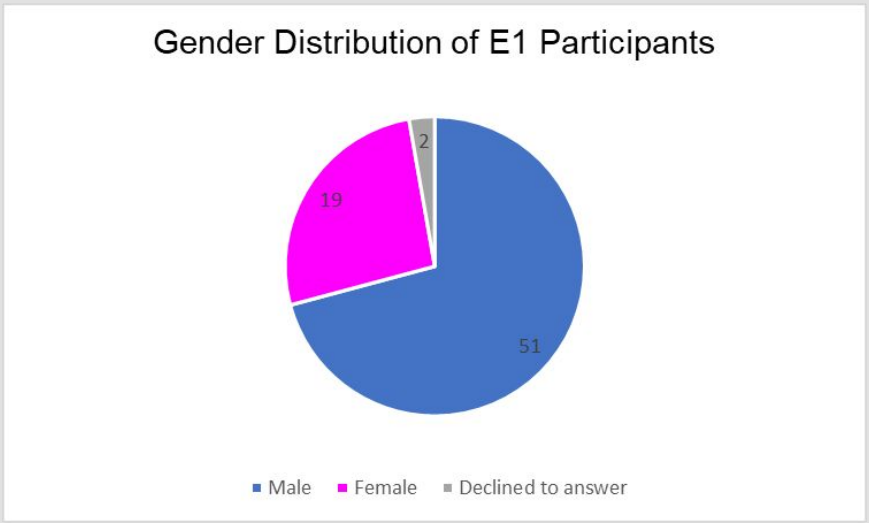
**Research Question 2**: How does knowing to have (knowing that you as a human have) more domain knowledge than the AI affect humans' trust, accuracy of AI-assisted predictions, and the effect of confidence score on trust calibration?

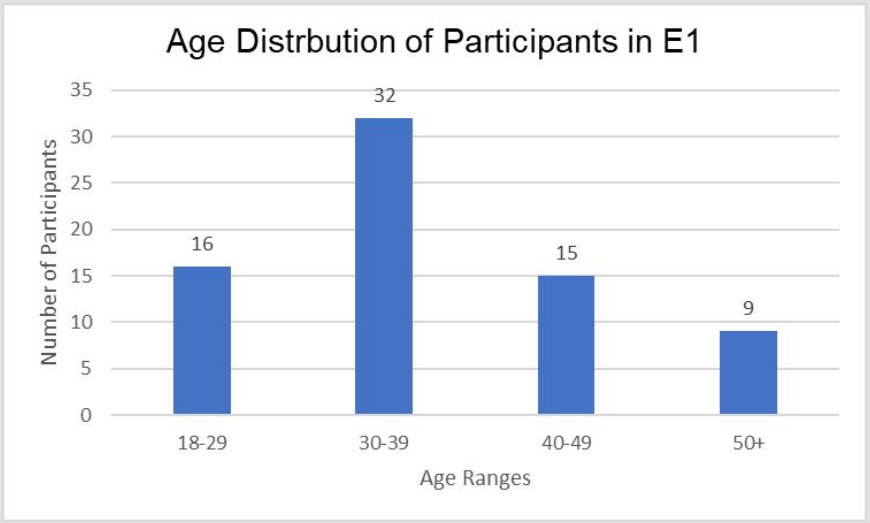
# E1: Experimental Design

**Income Prediction Task**

- Predict whether a person's annual income will exceed $50k given 8 demographic characteristics (binary prediction)
- 48,842 observed observations from 1994 census data
- Base pay $3, bonus of 5 cents if right, loss of 2 cents if wrong
- Training/testing split 70/30

# E1: Experimental Design (cont.)




72 total participants

# E1: Model Information

| Attributes | Value | Chance |
|---|---|---|
| Age | 50 | 4 |
| Sex | Male | 3 |
| Race | White | 3 |
| Marital Status | Married, spouse civilian | 4 |
| Years of Education | 10 | 2 |
| Workclass | Private | 2 |
| Occupation | Executive & managerial | 5 |
| Hours per week | 40 | 2 |

**Figure 1: A screenshot of a profile table shown in the experiment. The table lists eight attribute values and their corresponding chances (out of 10) that a person with the same attribute value would have income above $50K.**

**Data provided to increase participant domain knowledge on annual income**

# E1 Scenarios and Conditions

1. Show or not show AI confidence score
   - Stated as “The model’s prediction is correct x times out of 10”
2. Show or not show AI prediction
   - Show the user what the AI prediction is. Sometimes the human had to choose whether to delegate the task to the AI without seeing their prediction
3. Full vs partial model
   - This was meant to see if humans behaved differently when they knew if they had more information than the AI.
   - The partial model didn’t have marital status

# E1 Procedure

Each user gets 20 training trials

- Told what AI prediction was
- Told what correct answer was on last 10 trials

Next, 40 task trials

- User makes prediction
- User is shown AI prediction (or confidence score instead depending on scenario)
- User decides whether to go with their prediction or AI prediction

# E1 Results (Trust)

**Switch percentage**: the % of times the participant used the AI's prediction and not their own – considered to be a stricter measure of trust.

**Agreement percentage**: the % of all trials where the participant's prediction agreed with the AI's prediction

Partial or full models (completeness) did not affect percentages significantly - participants did not distrust the partial model

"high confidence scores encouraged participants to delegate the decision task to the AI even without seeing its predictions."

# E1 Results (Accuracy)

Humans were less accurate than AI in all trials (65% accuracy for humans vs 75% for AI)

AI assisted decision-making did not increase accuracy in this trial, disproving Hypothesis 2

Humans did not disagree much with AI in high-confidence cases. In the low-confidence zone, human decisions were not better than AI decisions.

Model decision uncertainty and human decision uncertainty were similar in low-confidence cases. These situations were also challenging for humans

# E1 Results

“participants switched to the AI’s predictions (or decided to use AI in the without-prediction conditions) more often when the AI’s confidence scores were displayed.”

“However, trust calibration did not translate into improvement in AI assisted decision-making outcomes

# Experiment Two: Local Explanations

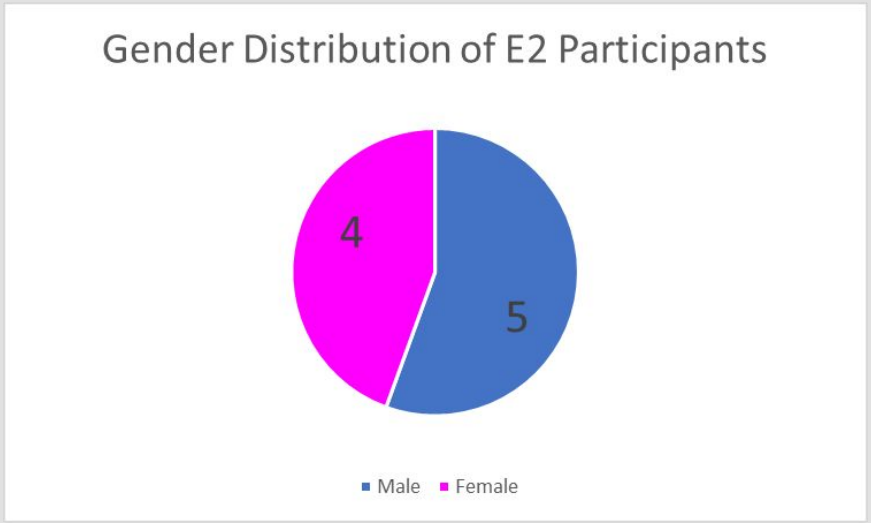
Same setup as Experiment One

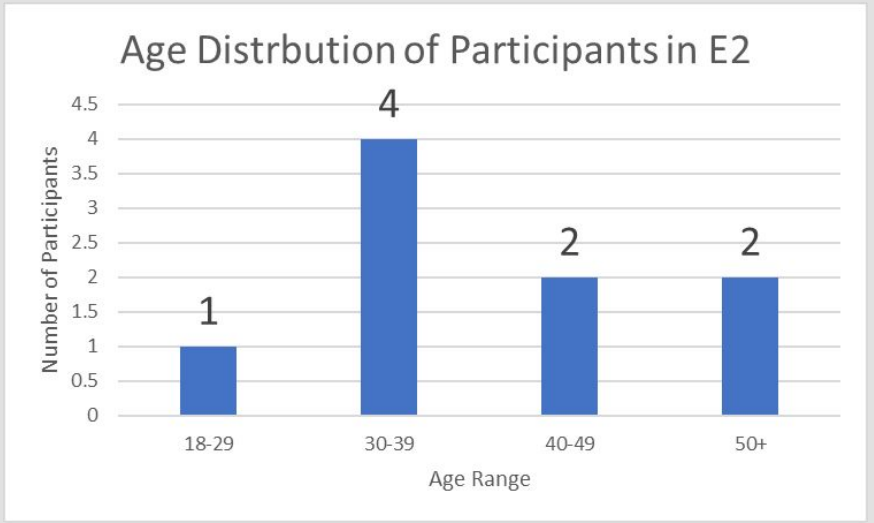
- Partial/complete model scenarios were removed based on E1 results that they were not statistically significant

**Hypothesis 3**: Showing a local explanation could support trust calibration

**Hypothesis 4:** ... and improve (accuracy of) AI-assisted prediction

# E2 Participant Pool




9 total participants

# E2 Local Explanation Visual

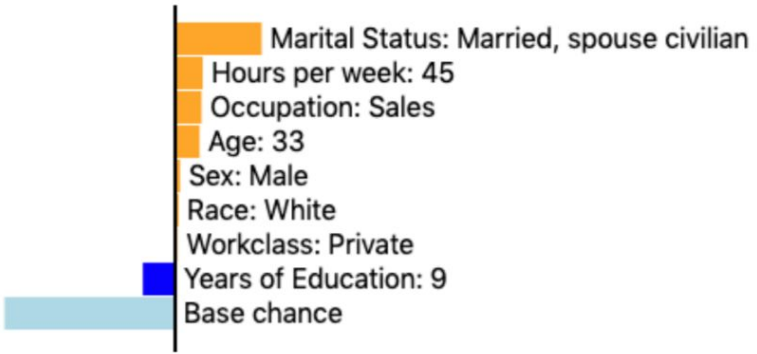



Figure 7: A screenshot of the explanation shown for a particular trial. Participants were told that orange bars indicate that the corresponding attributes suggest higher likelihood of income above 50K, whereas blue bars indicate higher likelihood of income below 50K. The light blue bar at the bottom indicates the base chance—a person with average values in all attributes is unlikely to have income above 50K.

**An example of a local explanation given to a user in Experiment Two**

# E2 Results (Trust)

Explanations did not seem to affect participant trust in model predictions

Showing the model confidence made participants trust the model in high-confidence cases

Rejected H3 - Found no evidence that explanations was more effective in trust calibration than baseline

# E2 Results (Accuracy)

Again, humans were less accurate than AI in all trials (63% accuracy for humans vs 75% for AI)

Similar to E1, AI assisted decision-making did not measurably increase accuracy in this trial

Alternatively, showing the explanation resulted in a **decrease** in AI-assisted accuracy

Thus, explanations are ineffective for both trust calibration and improvement of accuracy in this experiment

# Implications and Limitations

- Specific confidence information can improve trust calibration
- Showing confidence scores may be more helpful than explaining the model rationale locally or global
- Related literature has had different outcomes when it comes to accuracy on AI-assisted decision making
  - In this study, humans were likely to err when AI was also likely to err
- More research is needed to better understand how explainability can be used to improve accuracy and mitigate AI aversion, especially to guide the growth of highly accurate black-box models

# The Principles and Limits of Algorithm-in-the-Loop Decision Making

Ben Green, Yiling Chen

# Background

Decisions are increasingly being made through “algorithm-in-the-loop” processes, where machine learning models inform people

Often the focus is on the statistical properties of these tools (like accuracy and fairness)

There is little research on how the interactions between people and models actually influence human decisions

# Proposed Criteria and Research Goal

Accuracy

People using the algorithm should make more accurate predictions than they could without the algorithm.

Reliability

People should accurately evaluate on their own and the algorithm's performance, and people should calibrate their use of the algorithm to account for its accuracy and errors.

Fairness

People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

Research goal is to evaluate how people interact with risk assessments and study whether people satisfy these principles when making predictions

# Study Design

Broken down into two stages:

1. **Creating risk assessments for pretrial detention and financial loan applications**
2. Running experiments on Amazon Mechanical Turk

# Pretrial Detention Overview

When an individual is arrested, courts can either:

(a) Hold that individual in jail until their trial is held

(b) Release them with a mandate to return for their trial

If the perceived risk of a defendant is high such that, if released, that individual would most likely fail to return to court for their trial and/or commit any further crimes, then is is more likely that the court will detain that person until their trial.

# Pretrial Detention Data and Algorithm

| | All N=47,141 | Black N=26,246 | White N=20,895 | Sample N=300 | Black N=178 | White N=122 |
|---|---|---|---|---|---|---|
| **Background** | | | | | | |
| Male | 76.7% | 77.7% | 75.4% | 85.7% | 87.6% | 82.8% |
| Black | 55.7% | 100.0% | 0.0% | 59.3% | 100.0% | 0.0% |
| Mean age | 30.8 | 30.1 | 31.8 | 27.7 | 27.4 | 28.2 |
| Drug crime | 36.9% | 39.2% | 34.0% | 44.3% | 49.4% | 36.9% |
| Property crime | 32.7% | 30.7% | 35.3% | 36.0% | 32.0% | 41.8% |
| Violent crime | 20.4% | 20.9% | 19.8% | 14.7% | 14.0% | 15.6% |
| Public order crime | 10.0% | 9.3% | 10.8% | 5.0% | 4.5% | 5.7% |
| Prior arrest(s) | 63.4% | 68.4% | 57.0% | 55.0% | 66.9% | 37.7% |
| # of prior arrests | 3.8 | 4.3 | 3.1 | 3.6 | 4.6 | 2.2 |
| Prior conviction(s) | 46.5% | 51.2% | 40.7% | 39.7% | 50.0% | 24.6% |
| # of prior convictions | 1.9 | 2.2 | 1.6 | 2.2 | 2.8 | 1.3 |
| Prior failure to appear | 25.1% | 28.8% | 20.4% | 23.7% | 30.3% | 13.9% |
| **Outcomes** | | | | | | |
| Rearrest | 15.0% | 16.9% | 12.6% | 19.0% | 24.2% | 11.5% |
| Failure to appear | 20.3% | 22.6% | 17.5% | 23.3% | 28.1% | 16.4% |
| Violation | 29.8% | 33.1% | 25.6% | 32.3% | 39.9% | 21.3% |

47,141 defendants

- 76.7% male
- 55.7% black
- Average age of 30.8 years

Trained a model using gradient boosted trees

- Age, defense type, number of prior arrests, number of prior convictions, and previous failure to appear

# Financial Lending Overview

When an individual applies for a financial loan, the lender often will assess the risk that the borrower will fail to pay back the money (also called “defaulting” on the loan).

If an individual appears to be more likely to pay off the loan, then the lender is more likely to provide money to that individual.

# Financial Lending Data and Algorithm

| | All<br>N = 206,913 | Sample<br>N = 300 |
|---|---|---|
| **Applicant** | | |
| Annual income | $78,093.47 ($73,474.56) | $83,190.08 ($83,681.52) |
| Credit score | 695.3 (30.5) | 693.9 (30.3) |
| "Good" credit score | 71.2% | 70.7% |
| Home owner | 10.2% | 10.0% |
| Renter | 40.1% | 40.3% |
| Has mortgage | 49.7% | 49.7% |
| **Loan** | | |
| Loan amount | $15,133.51 ($8,575.05) | $15,377.75 ($8,520.84) |
| 36 months to pay off loan | 70.5% | 73.3% |
| 60 months to pay off loan | 29.5% | 26.7% |
| Monthly payment | $448.49 ($251.44) | $462.19 ($253.86) |
| Interest rate | 12.9% (4.5%) | 13.05% (4.5%) |
| **Outcomes** | | |
| Fully paid | 74.1% | 74.0% |
| Charged off | 25.9% | 26.0% |

206,913 issued loans

- Average loan was for $15,133.51
- Average income of income of $78,093.47
- Average of "Good" credit score

Trained a model using gradient boosted trees

- Applicant's annual income, credit score, and home ownership status; the value and interest rate of the loan; and the number of months to pay off the loan and the value of each monthly installment

# Study Design

Broken down into two stages:

(1) Creating risk assessments for pretrial detention and financial loan applications

**(2) Running experiments on Amazon Mechanical Turk**

# MTurk Task Overview

Participants randomly presented one of the two settings:

(1) Pretrial
(2) Loans

Participants randomly presented one of the six conditions:

(1) Baseline
(2) RA Prediction
(3) Default
(4) Update
(5) Explanation
(6) Feedback

# Condition Overview

## Baseline

Presents the narrative profile, without any information regarding the risk assessment

Represents when people made decisions without the aid of algorithms

## RA Prediction

Presents the narrative profile as well as the risk assessment's prediction in simple numeric form

Represents the simplest presentation of a risk assessment in numerical or categorical form as a factor for the human decision maker to consider

## Default

Presents the narrative profile as well as the risk assessment's prediction, except that the prediction form was automatically set to the risk assessment's prediction, such that participants could select any desired value

Looks at the algorithm's prediction first and then consider whether to deviate from and override that value

Treats the model's prediction as the presumptive default and requires further thought to override

## Update

First, presents just the narrative profile to make a prediction by themselves

Second, after making a prediction, presents the narrative profile as well as the risk assessment's prediction, asked again to make the prediction

Prompts people to focus on the narrative profile before considering the risk assessment's prediction

## Explanation

Presents the narrative profile as well as the risk assessment's prediction, along with an explanation that indicated which features made the risk assessment predict notably higher or lower levels of risk

Represents use of explanations of machine learning, indicating which attributes strongly influenced the risk assessment's prediction

May prevent people from double counting features that the model had already considered

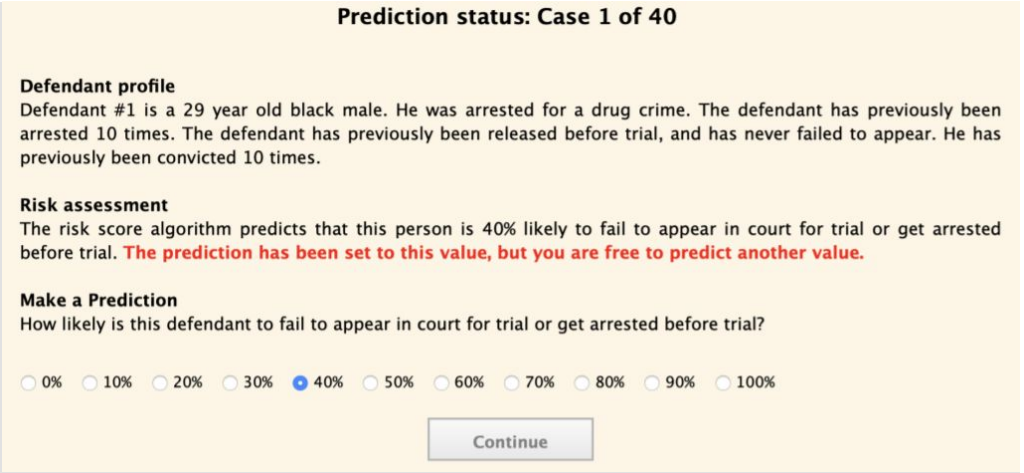
## Feedback

First, presents the RA Prediction condition

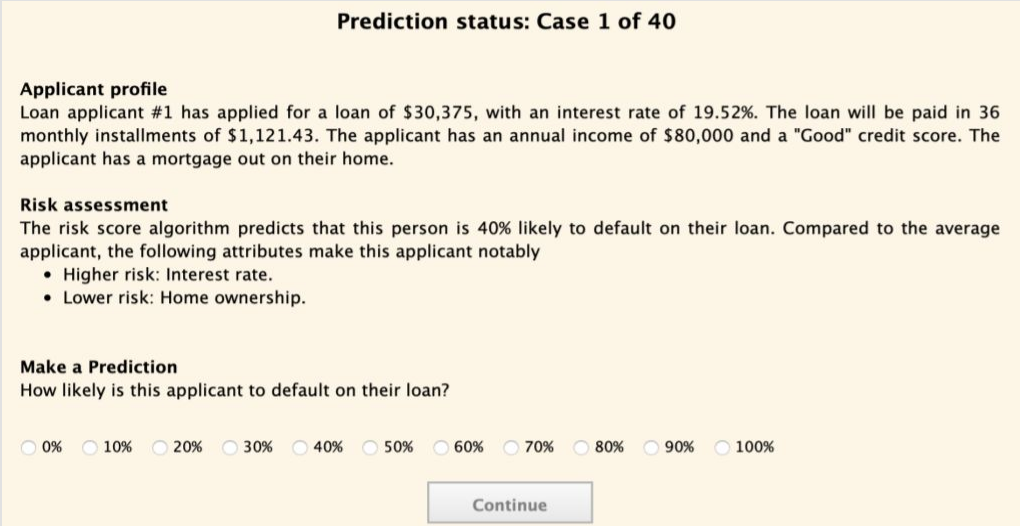
Second, after submitting each prediction, presents an alert indicating the outcome of that case

Provides a way of training for users of machine learning systems, which is an essential ingredient for effective implementation of risk assessments

Example of Pretrial setting with Default condition:




Example of Loans setting with Explanation condition:



# Proposed Criteria

**Accuracy**

**People using the algorithm should make more accurate predictions than they could without the algorithm.**

Reliability

People should accurately evaluate on their own and the algorithm's performance, and people should calibrate their use of the algorithm to account for its accuracy and errors.

Fairness

People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

# Results: Accuracy

| | Pretrial N=1156 | Loans N=732 |
|---|---|---|
| **Demographics** | | |
| Male | 55.3% | 53.0% |
| Black | 7.1% | 7.2% |
| White | 77.2% | 77.6% |
| 18-24 years old | 8.4% | 7.9% |
| 25-34 years old | 42.4% | 44.5% |
| 35-59 years old | 45.0% | 43.2% |
| 60+ years old | 4.2% | 4.4% |
| College degree or higher | 70.9% | 71.7% |
| Criminal justice familiarity | 2.8 | 2.9 |
| Financial lending familiarity | 2.7 | 2.9 |
| Machine learning familiarity | 2.4 | 2.5 |
| **Treatment** | | |
| Baseline | 16.5% (N=191) | 15.3% (N=112) |
| Risk Assessment | 17.3% (N=200) | 16.9% (N=124) |
| Default | 16.9% (N=195) | 17.6% (N=129) |
| Update | 16.1% (N=186) | 17.9% (N=131) |
| Explanation | 15.1% (N=174) | 16.8% (N=123) |
| Feedback | 18.2% (N=210) | 15.4% (N=113) |
| **Outcomes** | | |
| Participant hourly wage | $15.20 | $17.18 |
| Experiment clarity | 4.4 | 4.4 |
| Experiment enjoyment | 3.5 | 3.7 |

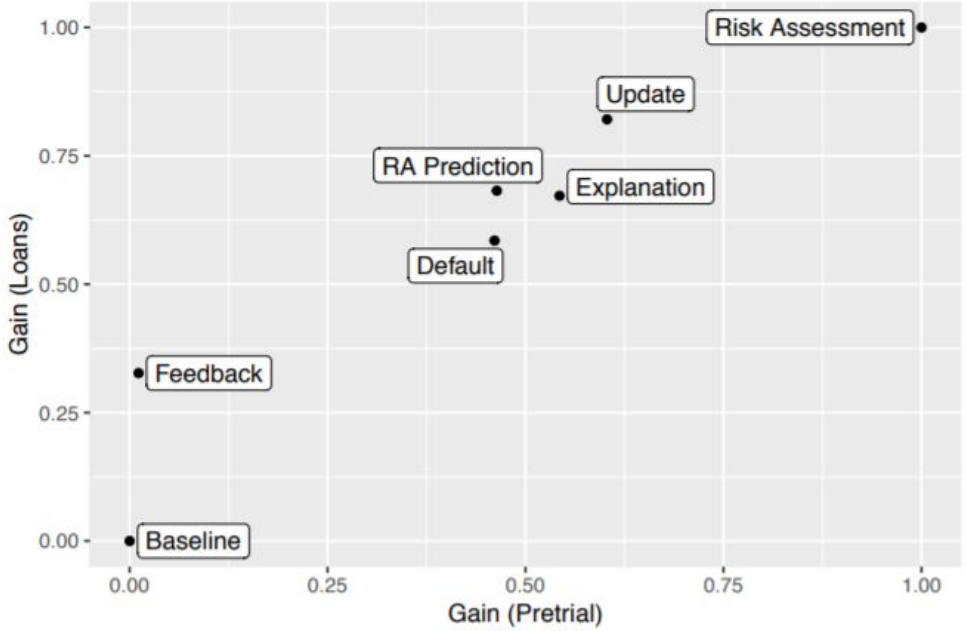



Fig. 2. The relative performance gain (Equation 1) achieved by each experimental condition across the pretrial and loans settings. In both settings, the Update treatment performed statistically significantly better than RA Prediction and the Feedback treatment performed statistically significantly worse. Across the two settings, the gain of the conditions was highly correlated.

# Proposed Criteria

Accuracy

People using the algorithm should make more accurate predictions than they could without the algorithm.

**Reliability**

**People should accurately evaluate on their own and the algorithm's performance, and people should calibrate their use of the algorithm to account for its accuracy and errors.**

Fairness

People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

# Results: Reliability

Table 4. Summary of participant abilities to evaluate performance (first two columns) and to calibrate their predictions (third column). The columns measure the relationships between between participant confidence and actual performance (Confidence), participant estimates of the algorithm's performance and its actual performance (RA Accuracy), and participant reliance on the risk assessment and the risk assessment's performance (Calibration). + signifies a positive and statistically significant relationship, - signifies a negative and statistically significant relationship, and 0 signifies no statistically significant relationship. In all cases, + means that the desired behavior was observed.

| | **Confidence** | | **RA Accuracy** | | **Calibration** | |
|---|---|---|---|---|---|---|
| | Pretrial | Loans | Pretrial | Loans | Pretrial | Loans |
| RA Prediction | 0 | 0 | 0 | 0 | - | 0 |
| Default | 0 | - | 0 | - | 0 | 0 |
| Update | 0 | - | - | - | 0 | 0 |
| Explanation | 0 | 0 | 0 | 0 | - | + |
| Feedback | 0 | 0 | 0 | 0 | - | 0 |

# Proposed Criteria

Accuracy

People using the algorithm should make more accurate predictions than they could without the algorithm.

Reliability

People should accurately evaluate on their own and the algorithm's performance, and people should calibrate their use of the algorithm to account for its accuracy and errors.

**Fairness**

**People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.**

# Results: Fairness

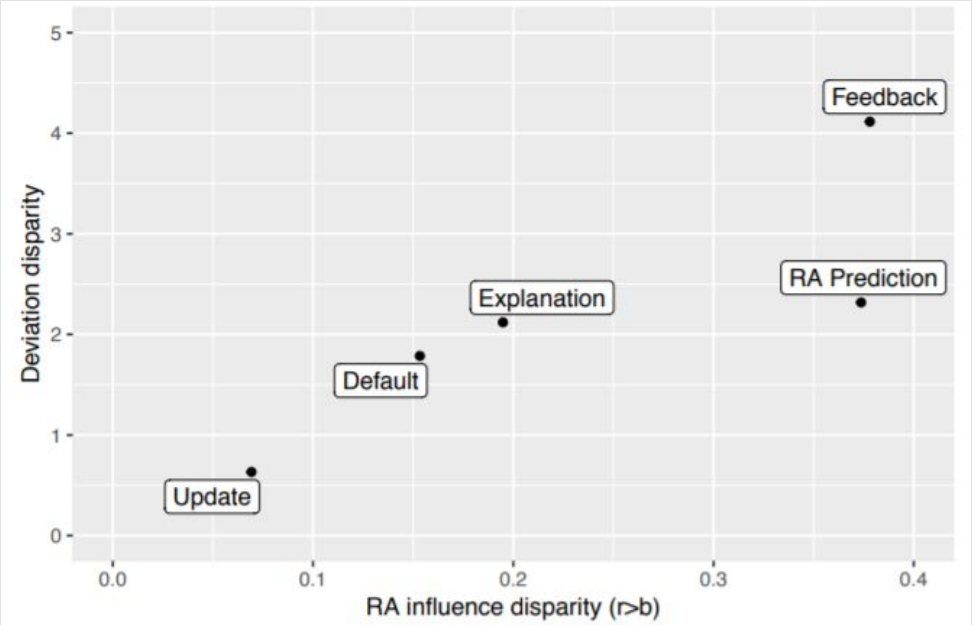



Fig. 3. The disparate interactions present in each treatment in the pretrial setting, measured by the disparities in risk assessment influence (Equation 3) and in participant deviations (Equation 4) for black versus white defendants. In both cases, values closer to 0 indicate lower levels of bias. The Update treatment yielded the smallest disparate interactions along both metrics, reducing the disparities (compared to RA Prediction) by 81.5% and 73.9%, respectively.

# Discussion Session

(1) Besides accuracy, fairness, and reliability, what other criteria could characterize an ethical and responsible decision when a person is informed by an algorithm?

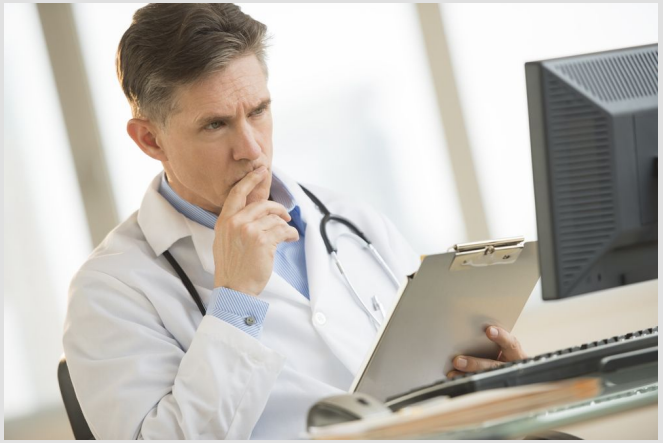
(2) Do you think the ways that people make decisions when informed by an algorithm satisfy these criteria?

# Understanding the Effect of Accuracy on Trust in Machine Learning Models

Ming Yin, Jennifer Wortman Vaughan, Hanna Wallach

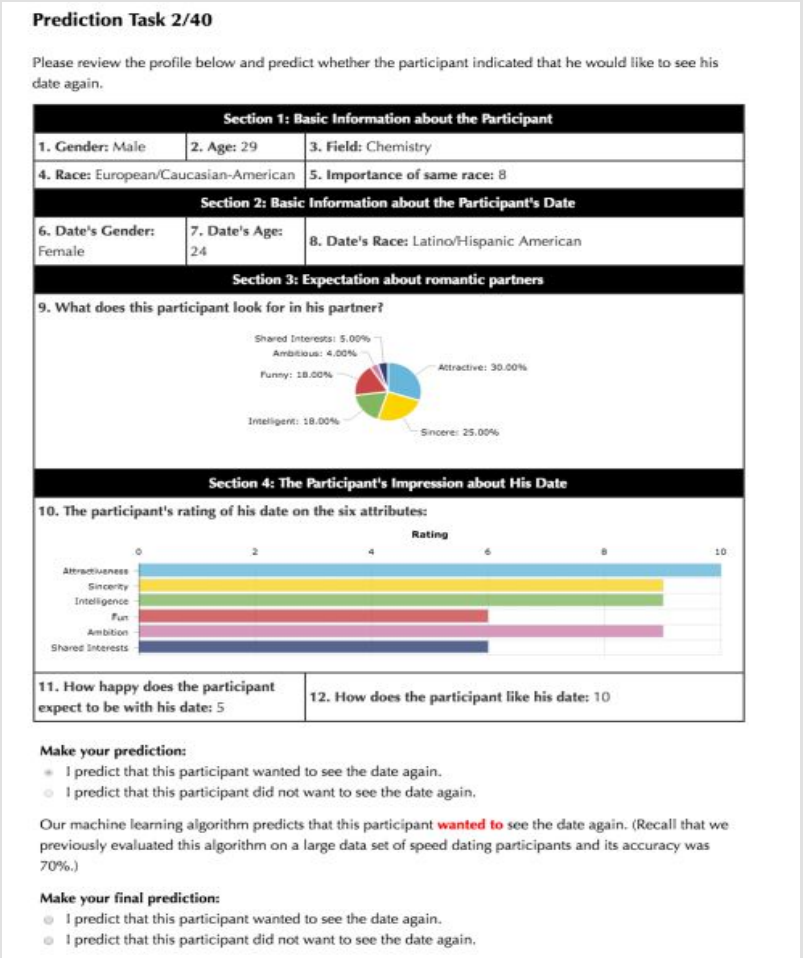
# Background

- In some situations, humans are making choices while being advised by some AI
- This paper aims to understand how a model's stated and observed accuracy affect a person's reliance on an algorithm



# Experimental Setup

- Task: Predict whether or not participant would want to see date
- Next, they were given a model's prediction and the model's accuracy and were given the option to revise their decision
- Halfway through the experiment, feedback was given on the Model's accuracy and the human's accuracy



**Prediction Task 2/40**

Please review the profile below and predict whether the participant indicated that he would like to see his date again.

| Section 1: Basic Information about the Participant | | |
|---|---|---|
| **1. Gender:** Male | **2. Age:** 29 | **3. Field:** Chemistry |
| **4. Race:** European/Caucasian-American | | **5. Importance of same race:** 8 |
| **Section 2: Basic Information about the Participant's Date** | | |
| **6. Date's Gender:** Female | **7. Date's Age:** 24 | **8. Date's Race:** Latino/Hispanic American |

**Section 3: Expectation about romantic partners**

**9. What does this participant look for in his partner?**



**Section 4: The Participant's Impression about His Date**

**10. The participant's rating of his date on the six attributes:**



| | |
|---|---|
| **11. How happy does the participant expect to be with his date:** 5 | **12. How does the participant like his date:** 10 |

**Make your prediction:**

- I predict that this participant wanted to see the date again.
- I predict that this participant did not want to see the date again.

Our machine learning algorithm predicts that this participant **wanted to** see the date again. (Recall that we previously evaluated this algorithm on a large data set of speed dating participants and its accuracy was 70%.)

**Make your final prediction:**

- I predict that this participant wanted to see the date again.
- I predict that this participant did not want to see the date again.

# Experimental Setup

Two Variables to measure “trust”:

- **Agreement fraction**: the number of tasks for which the subject’s final prediction agreed with the model’s prediction, divided by the total number of tasks.

- **Switch fraction**: the number of tasks for which the subject’s initial prediction disagreed with the model’s prediction and the subject’s final prediction agreed with the model’s prediction, divided by the total number of tasks for which the subject’s initial prediction disagreed with the model’s prediction

Experiment was run twice with:

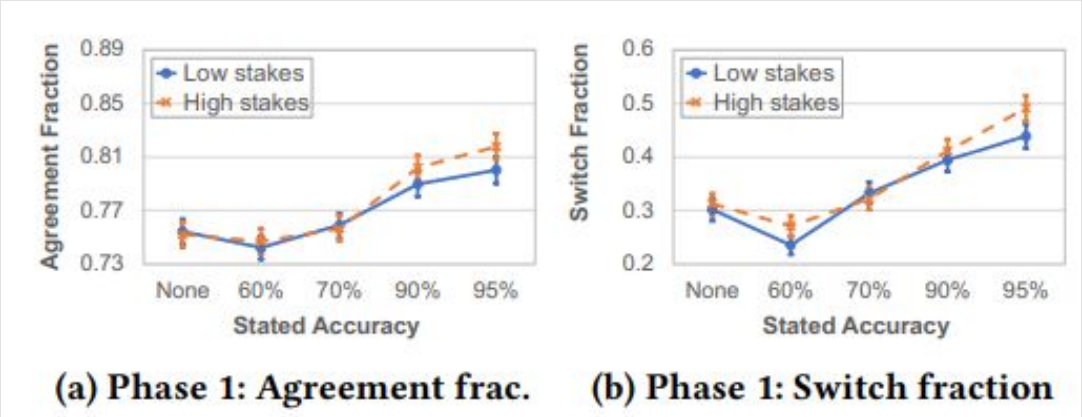
- **High Stakes:** Participants received bonus for correct final prediction

- **Low Stakes:** Participants paid flat rate

# Experiment 1:

**Question**: Does a model's stated accuracy affect laypeople's trust?

**Results:**



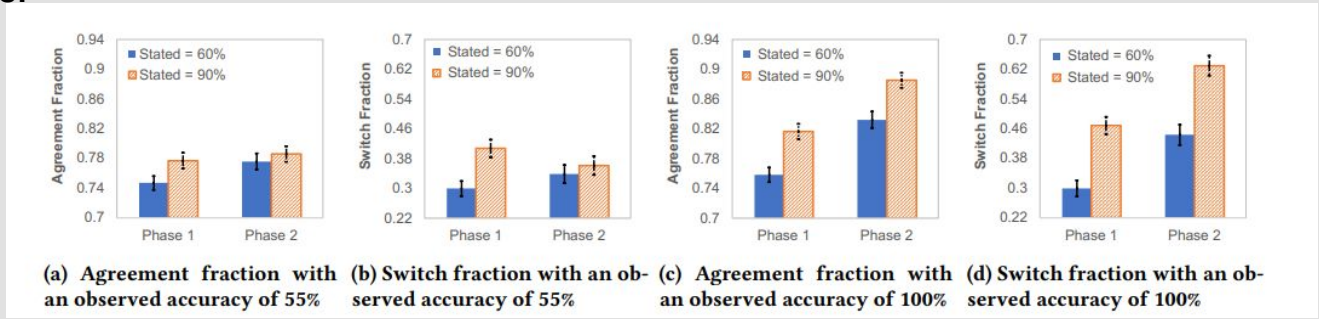

- The model's stated accuracy had significant effect on both agreement rate and switch fraction
- Not shown here, but once the observed results come into play, people tend to care less about a model's stated accuracy
- Stakes overall have little effect in this experiment

# Experiment 2:

**Question**: Does this effect differ based on the observed accuracy?

- Recall that halfway through the experiment the observed accuracy of the algorithm is revealed (i.e. correct 14/20 times)
- In this experiment, two groups of people are shown an algorithm with an observed accuracy of 55% and 100%

**Results:**

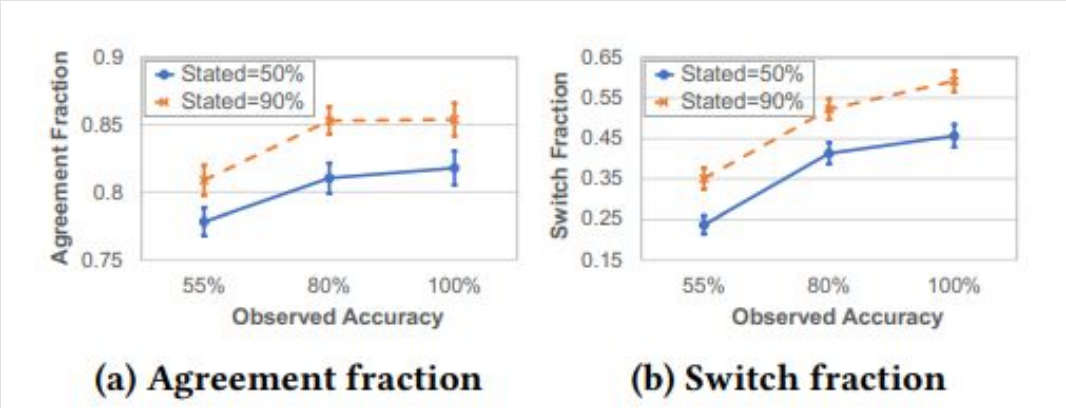



- When the observed accuracy is high, the stated accuracy still has a significant effect on people's trust in the algorithm
- When the observed accuracy is low, people tend to not care about the stated accuracy any more

# Experiment 3:

**Question**: Does a model's observed accuracy affect laypeople's trust?

**Results:**




- For both stated accuracies, the observed accuracy has significant impact on people's trust of the algorithm

# Exploratory Analysis

3 theories

- **Stated vs. observed**: Subjects increase their trust in the model if its observed accuracy is higher than its stated accuracy, and decrease their trust otherwise (Not Supported by the data)

- **Self vs. stated**: Subjects increase their trust in the model if the model's stated accuracy is higher than their own accuracy in Phase 1, and decrease their trust otherwise (Not Supported by the data)

- **Self vs. observed**: Subjects increase their trust in the model if the model's observed accuracy is higher than their own accuracy in Phase 1, and decrease their trust otherwise (Mostly Supported by the data)

# Key Takeaways:

- A model's stated accuracy on held-out data affects people's trust in the model, but that the effect size is smaller after people observe the model's accuracy in practice

- Generally, people choose to trust a model based on a model's observed accuracy even if its sample size is small

- People are more likely to trust a model if the observed accuracy is higher than their own personal accuracy

# Discussion Session

If your personal accuracy was as good as a model's accuracy, would you still consider the model's advice in your future predictions? Is this model even worth building?

In what scenarios would you depend more on a model’s observed accuracy as opposed to its stated accuracy?