THE WEB CONFERENCE ACM

Lyon, France, 25-29 April 2022

Will You Accept the AI Recommendation?

Predicting Human Behavior in AI-Assisted Decision Making

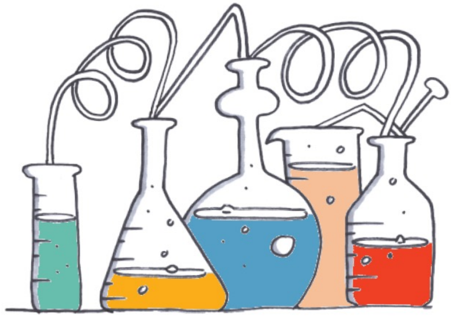
Xinru Wang*, Zhuoran Lu*, Ming Yin

PURDUE UNIVERSITY

AI-driven decision aids are everywhere…

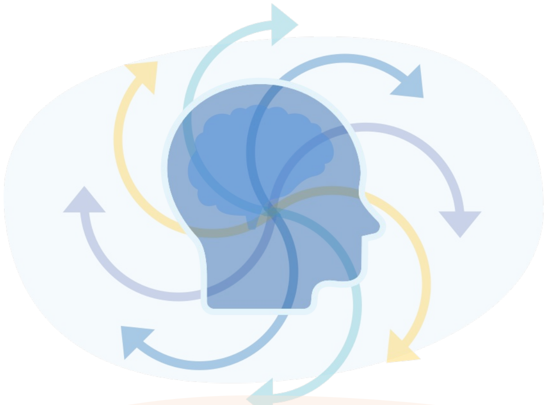
Motivation

Experimental Studies



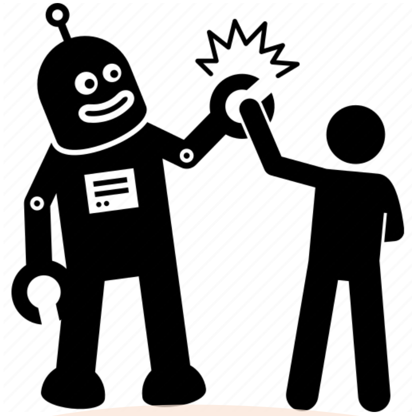
factors that can influence people's trust in AI

Computational Human Models



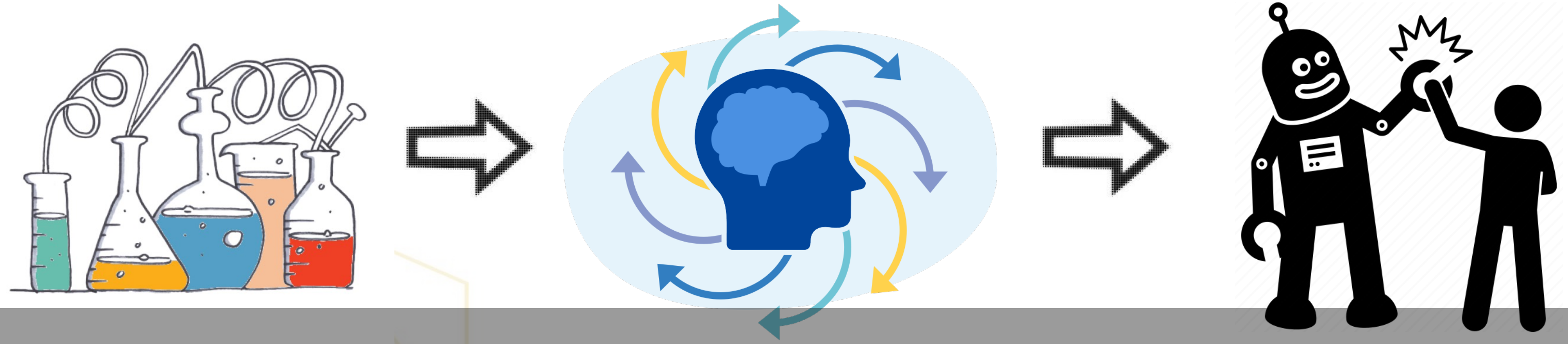
What's the cognitive mechanisms that govern how these factors interact?

Human-AI Collaboration



optimize AI for joint human-AI decision making

Motivation



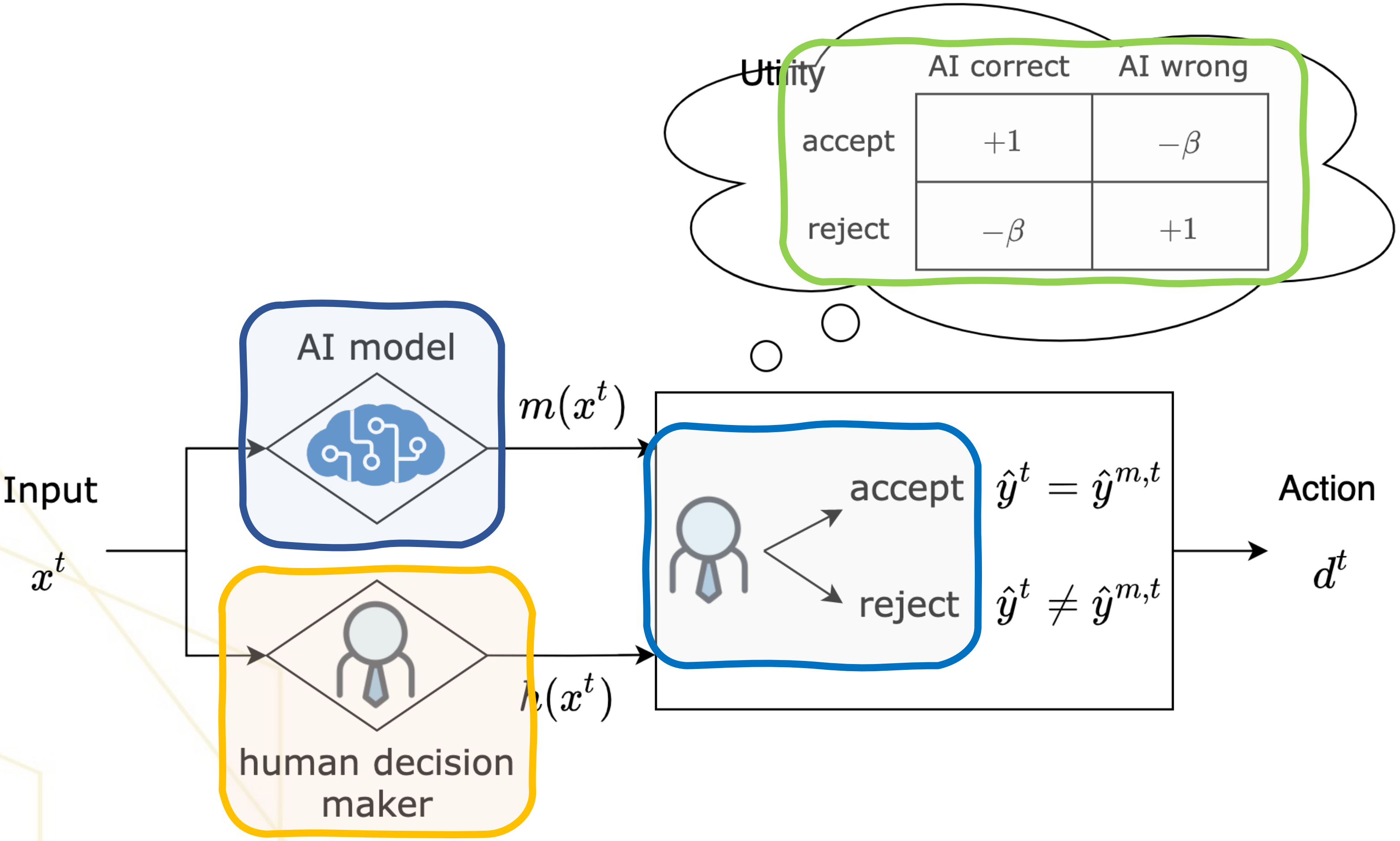
Experimental Studies

Computational Human Models

Human-AI Collaboration

How do humans decide whether to adopt an AI model's recommendation?

Problem Description



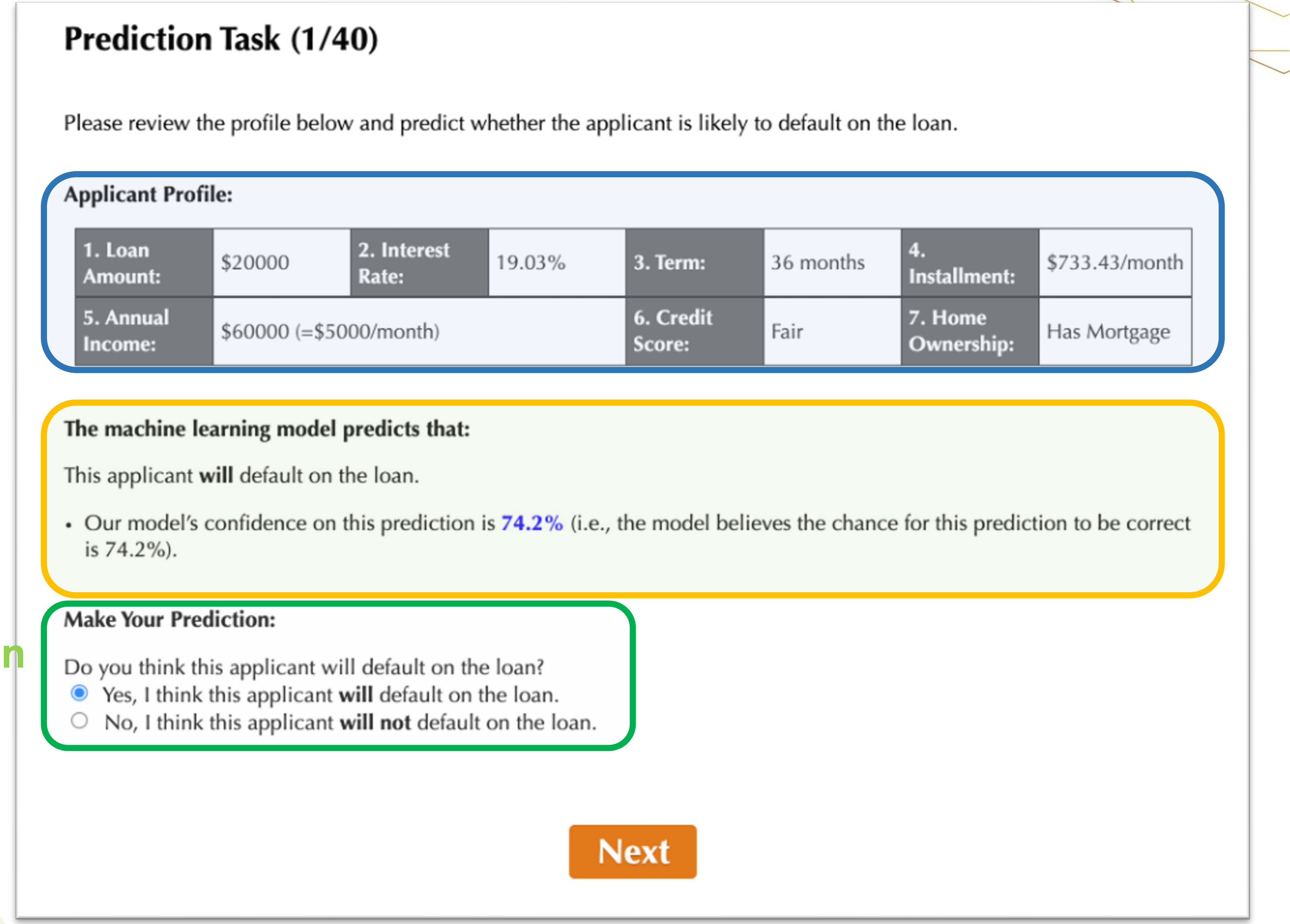
Utility	AI correct	AI wrong
accept	$+1$	$-\beta$
reject	$-\beta$	$+1$

Human-Subject Experiment

- 40 loan risk assessment tasks

AI model's prediction and confidence

make a final prediction



Prediction Task (1/40)

Please review the profile below and predict whether the applicant is likely to default on the loan.

Applicant Profile:

1. Loan Amount:	$20000	**2. Interest Rate:**	19.03%	**3. Term:**	36 months	**4. Installment:**	$733.43/month
5. Annual Income:	$60000 (=$5000/month)			**6. Credit Score:**	Fair	**7. Home Ownership:**	Has Mortgage

The machine learning model predicts that:

This applicant **will** default on the loan.

- Our model's confidence on this prediction is **74.2%** (i.e., the model believes the chance for this prediction to be correct is 74.2%).

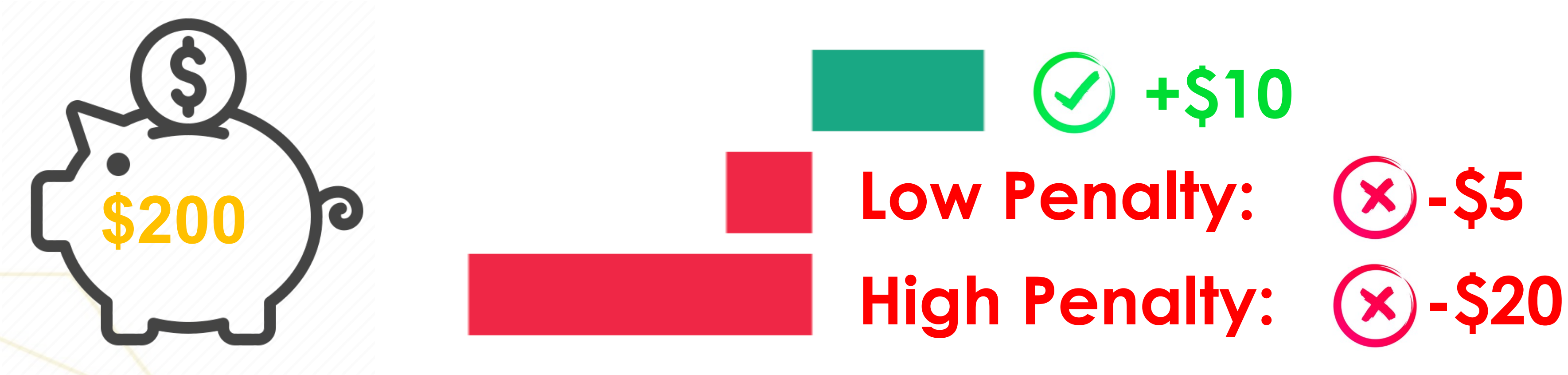
Make Your Prediction:

Do you think this applicant will default on the loan?

● Yes, I think this applicant **will** default on the loan.

○ No, I think this applicant **will not** default on the loan.

Experimental Treatments



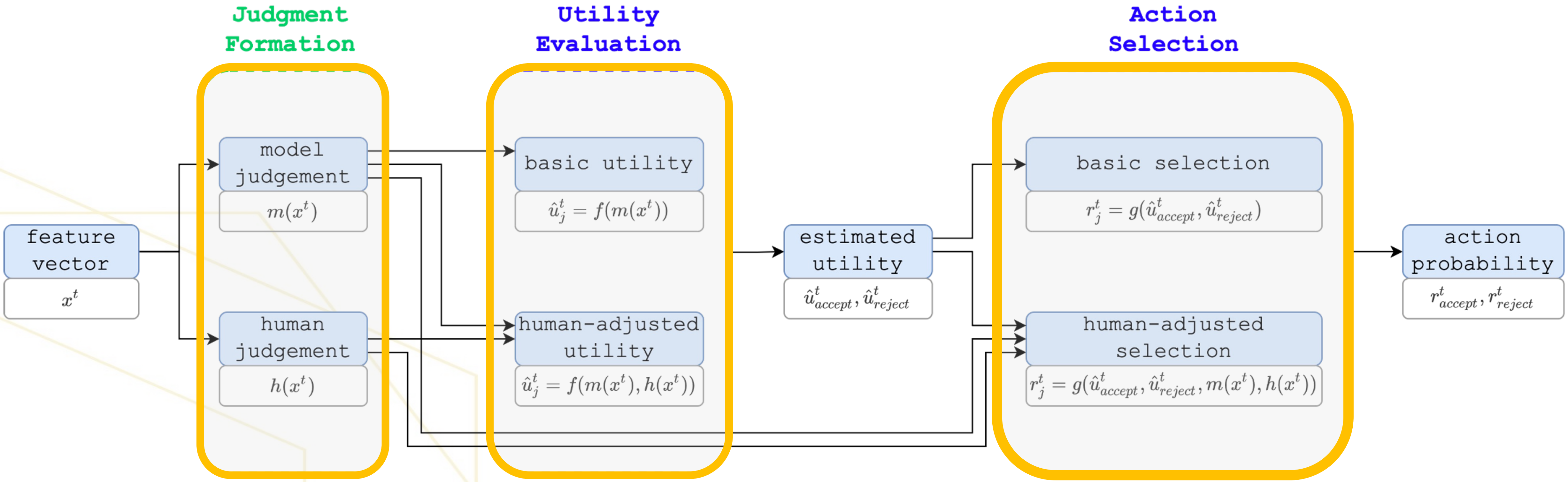
✓ +$10

Low Penalty: ☒ -$5

High Penalty: ☒ -$20

High Penalty Treatment: **214** participants
Low Penalty Treatment : **190** participants

Two-Component Models



Judgment Formation

Utility Evaluation

Action Selection

feature vector

x^t

model judgement

$m(x^t)$

human judgement

$h(x^t)$

basic utility

$\hat{u}^t_j = f(m(x^t))$

human-adjusted utility

$\hat{u}^t_j = f(m(x^t), h(x^t))$

estimated utility

$\hat{u}^t_{accept}, \hat{u}^t_{reject}$

basic selection

$r^t_j = g(\hat{u}^t_{accept}, \hat{u}^t_{reject})$

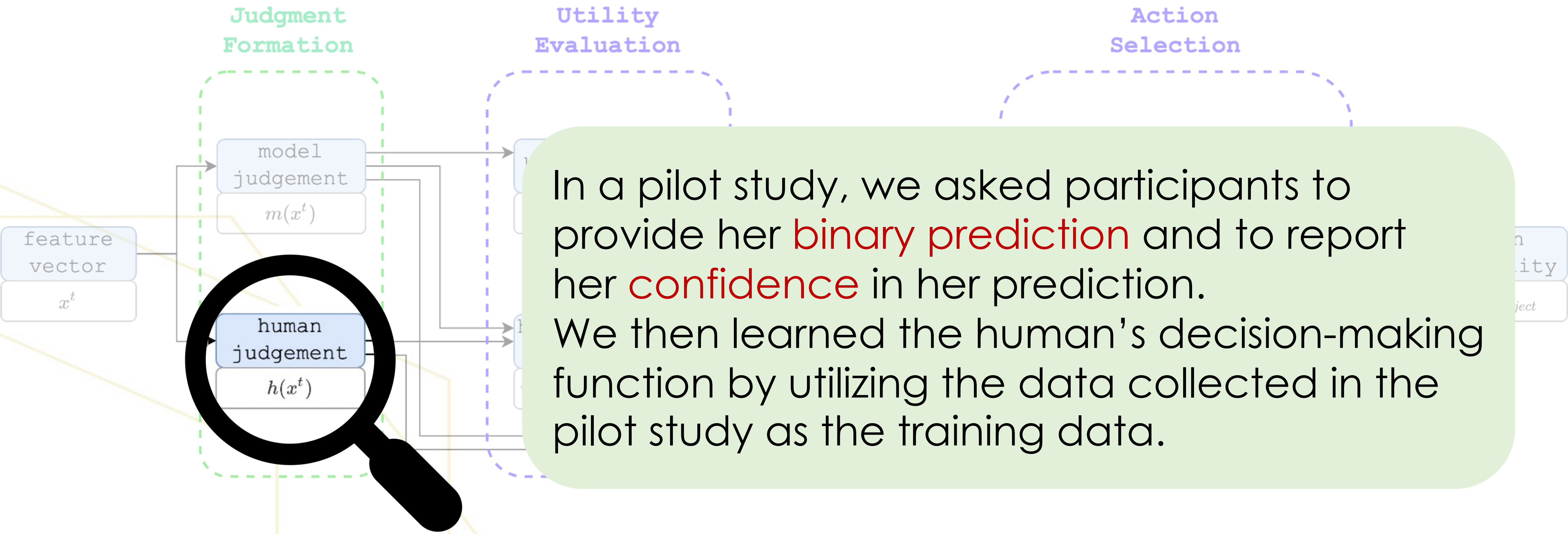
human-adjusted selection

$r^t_j = g(\hat{u}^t_{accept}, \hat{u}^t_{reject}, m(x^t), h(x^t))$

action probability

$r^t_{accept}, r^t_{reject}$

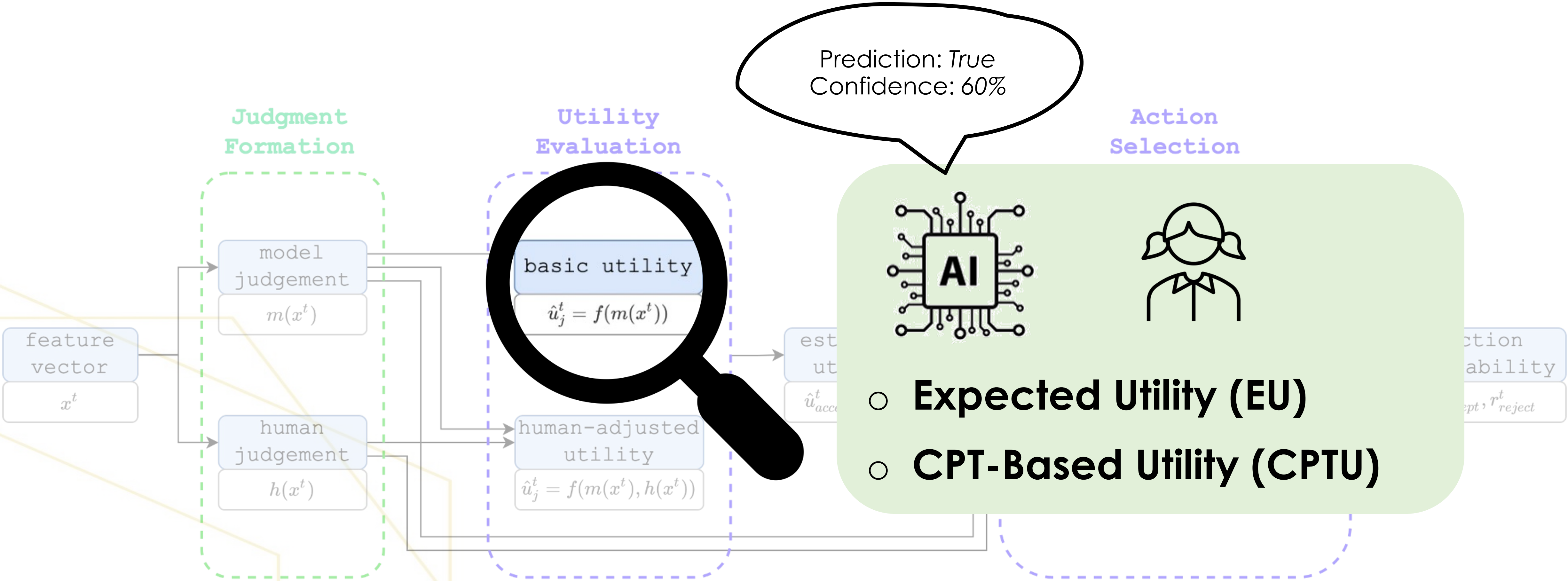
Two-Component Models



In a pilot study, we asked participants to provide her binary prediction and to report her confidence in her prediction.
We then learned the human's decision-making function by utilizing the data collected in the pilot study as the training data.

Two-Component Models

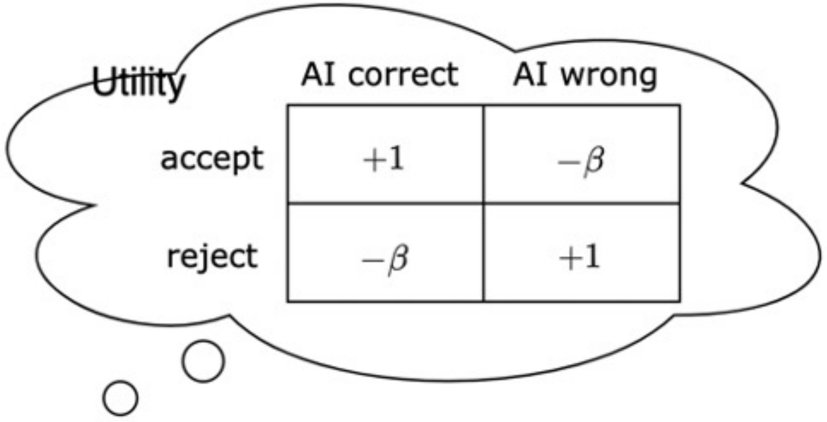
Utility Component



- ○ **Expected Utility (EU)**
- ○ **CPT-Based Utility (CPTU)**

Two-Component Models

Utility Component



Utility	AI correct	AI wrong
accept	$+1$	$-\beta$
reject	$-\beta$	$+1$

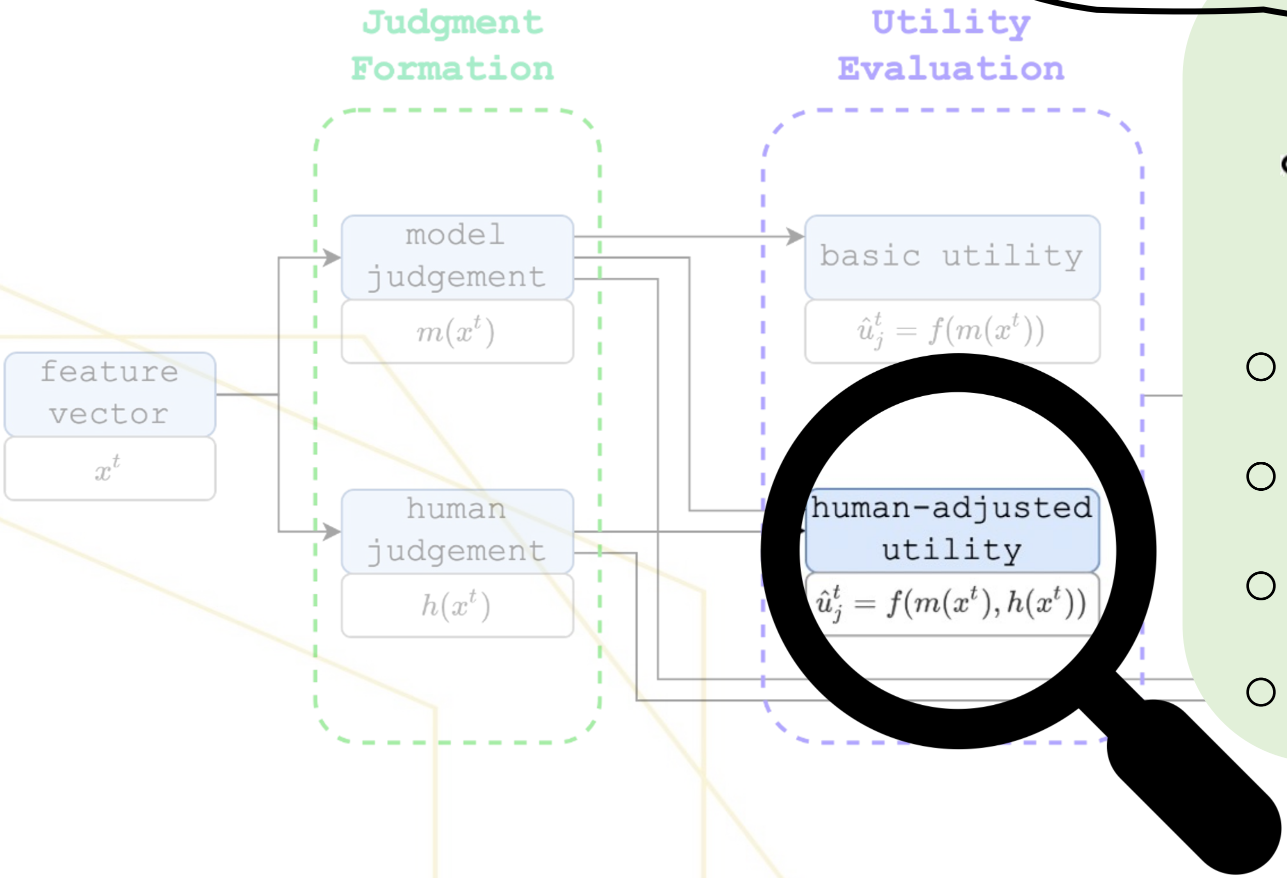
- the utility of accepting AI is $0.6 - 0.4\beta$
- the utility of rejecting AI is $-0.6\beta + 0.4$

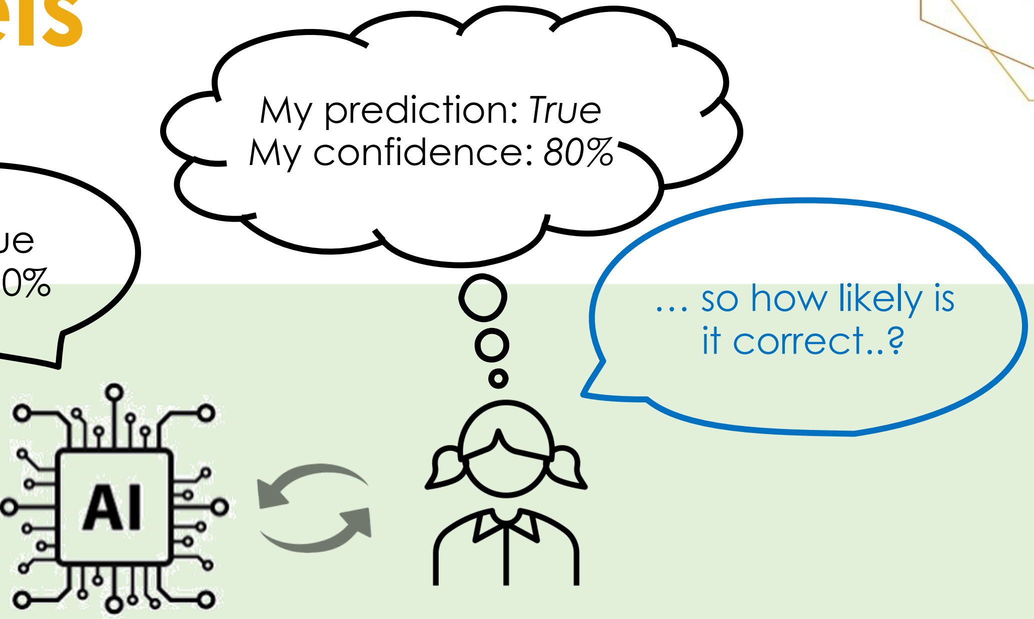
Prediction: *True*
Confidence: *60%*

- ○ **Expected Utility (EU)**
- ○ **CPT-Based Utility (CPTU)**

Two-Component Models

Utility Component

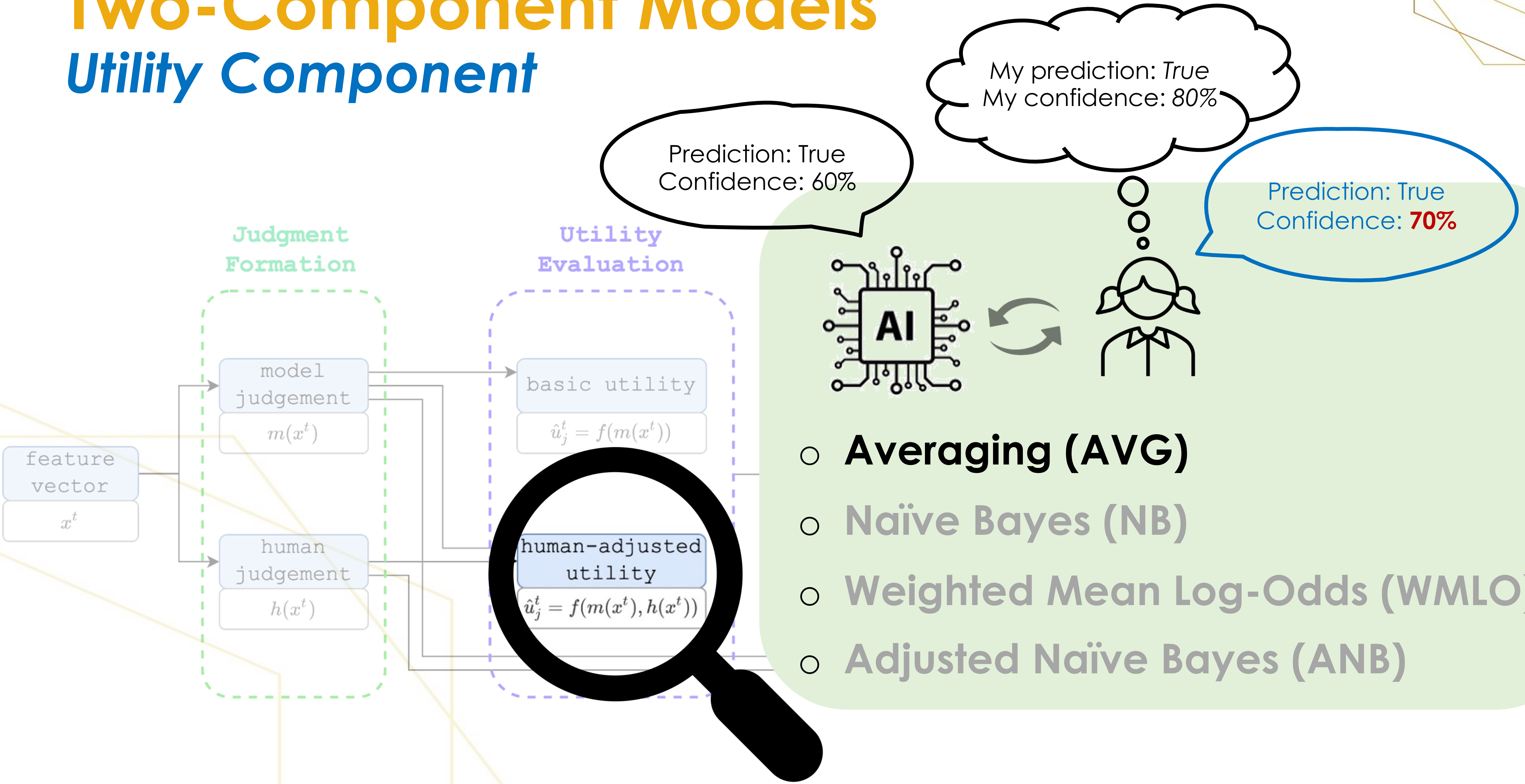




- ○ **Averaging (AVG)**
- ○ **Naïve Bayes (NB)**
- ○ **Weighted Mean Log-Odds (WMLO)**
- ○ **Adjusted Naïve Bayes (ANB)**

Two-Component Models

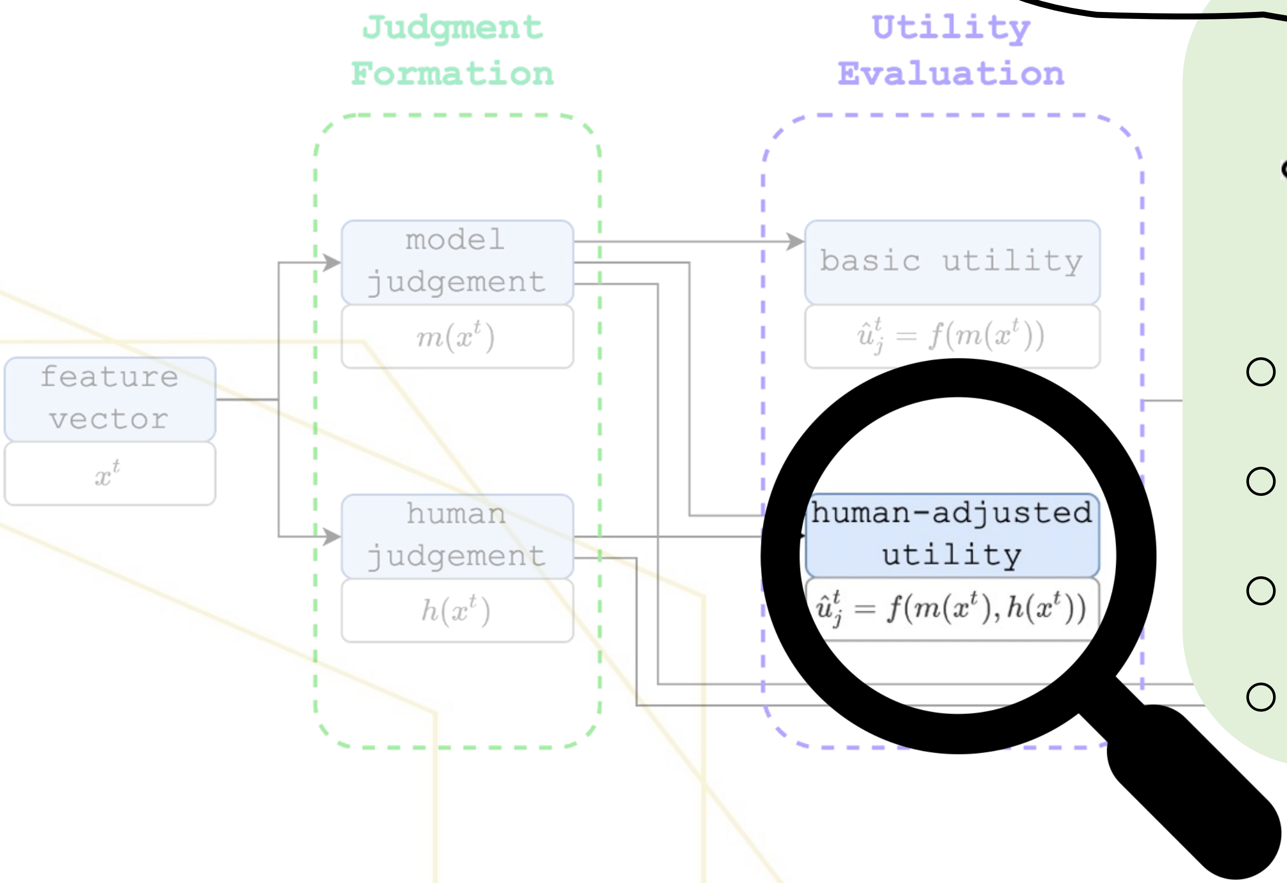
Utility Component

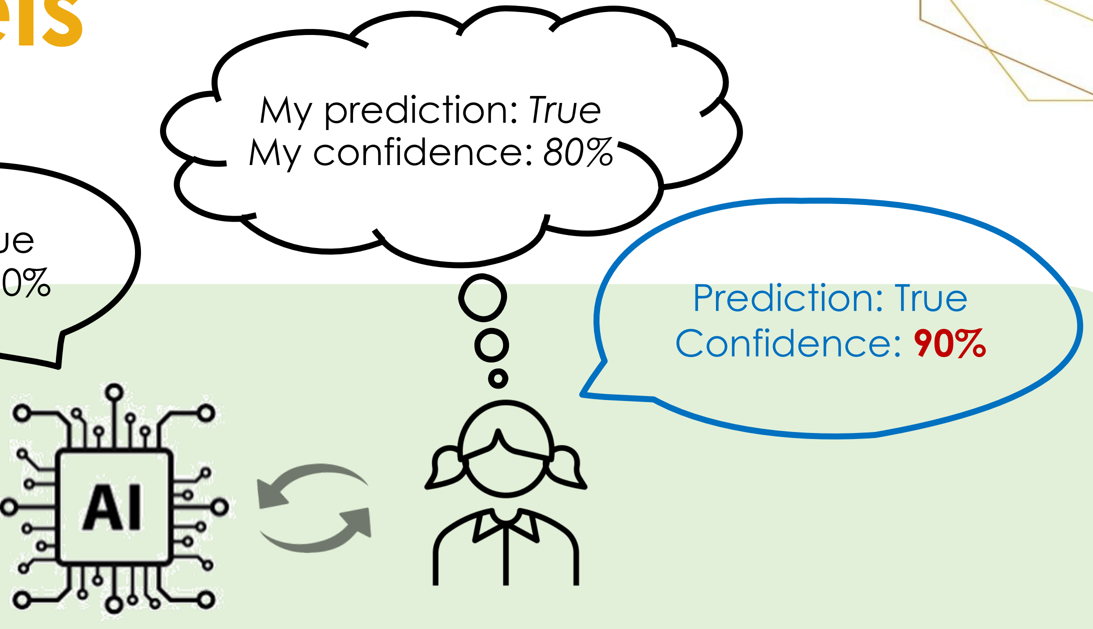


- **Averaging (AVG)**
- Naïve Bayes (NB)
- Weighted Mean Log-Odds (WMLO)
- Adjusted Naïve Bayes (ANB)

Two-Component Models

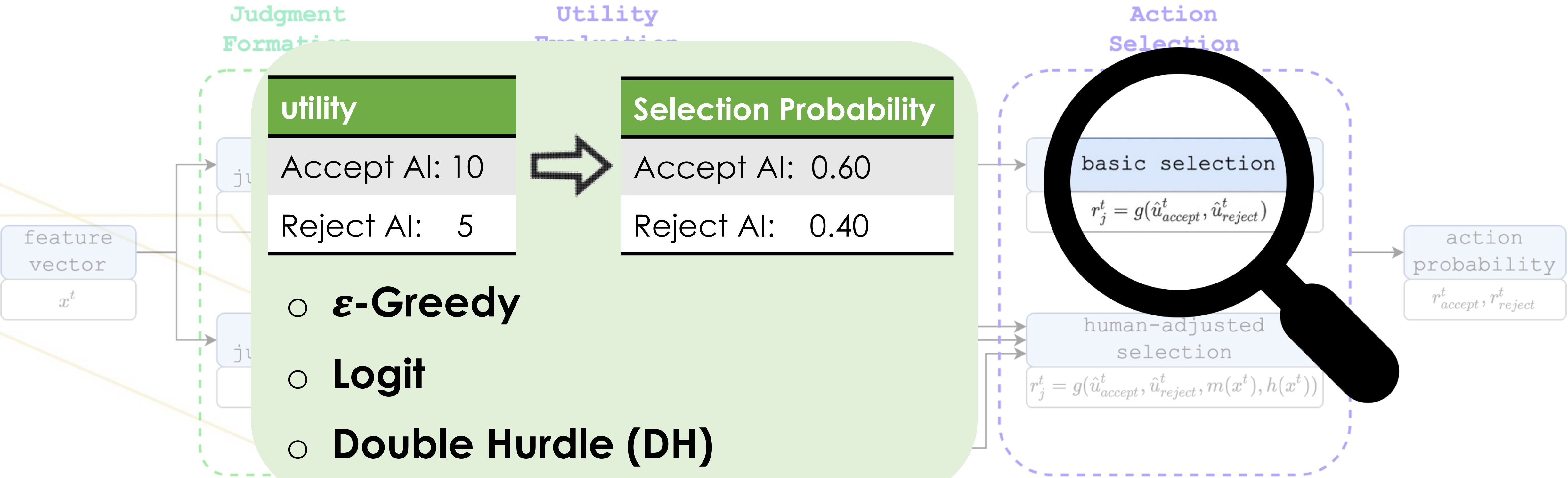
Utility Component





- Averaging (AVG)
- **Naïve Bayes (NB)**
- Weighted Mean Log-Odds (WMLO)
- Adjusted Naïve Bayes (ANB)

Two-Component Models
Selection Component


Judgment
Utility
Action Selection
feature vector
x^t
basic selection
$r_j^t = g(\hat{u}^t_{accept}, \hat{u}^t_{reject})$
human-adjusted selection
$r_j^t = g(\hat{u}^t_{accept}, \hat{u}^t_{reject}, m(x^t), h(x^t))$
action probability
$r^t_{accept}, r^t_{reject}$

utility	
Accept AI:	10
Reject AI:	5

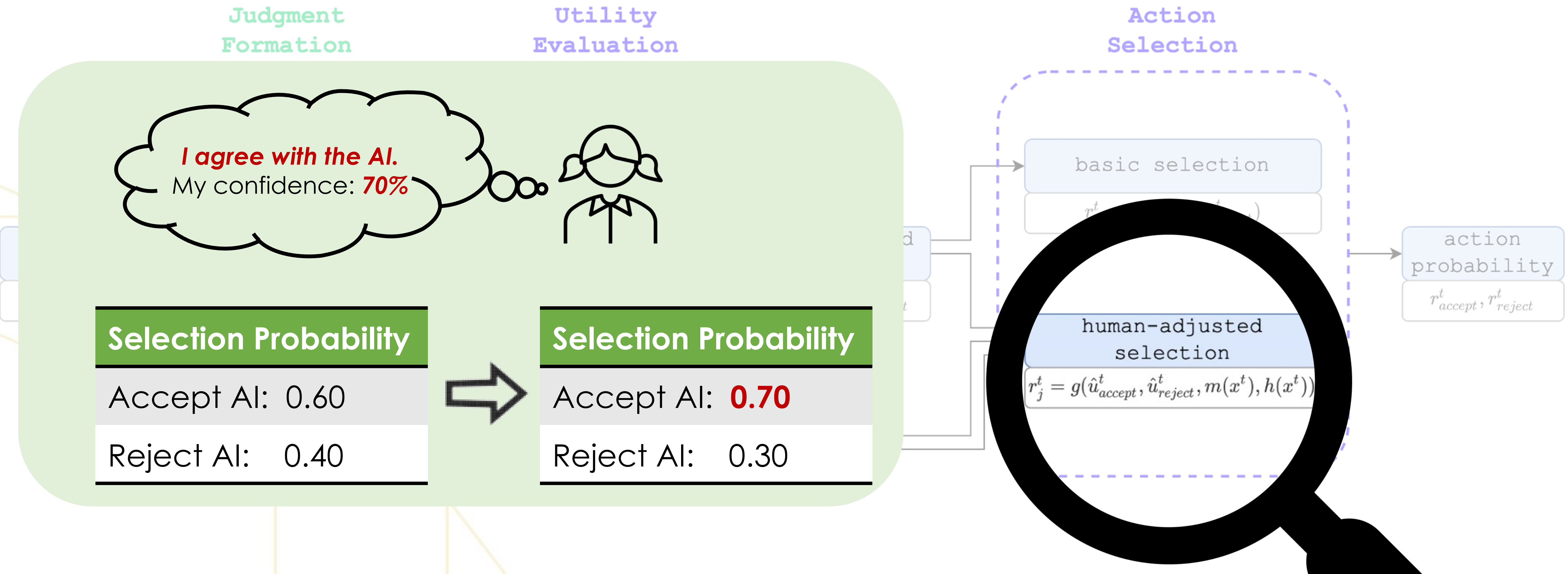
⇨

Selection Probability	
Accept AI:	0.60
Reject AI:	0.40

- **ε-Greedy**
- **Logit**
- **Double Hurdle (DH)**

Two-Component Models

Selection Component



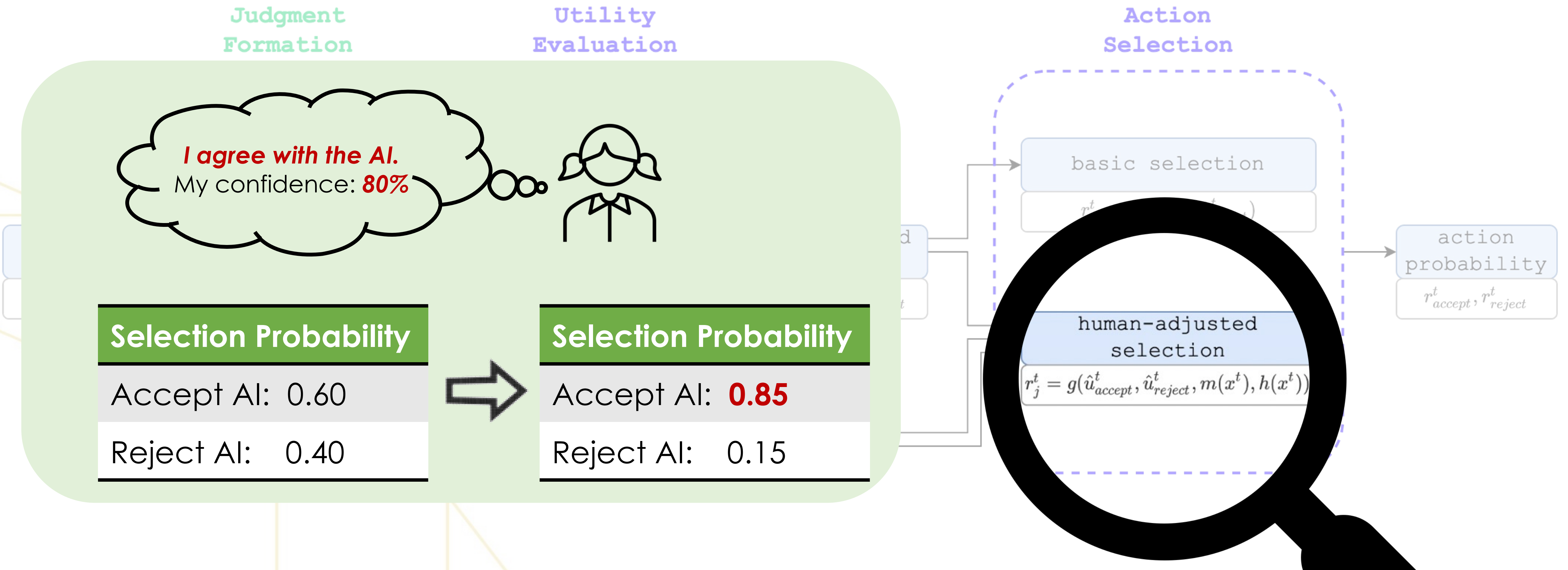
Selection Probability	
Accept AI:	0.60
Reject AI:	0.40

⇨

Selection Probability	
Accept AI:	**0.70**
Reject AI:	0.30

Two-Component Models

Selection Component

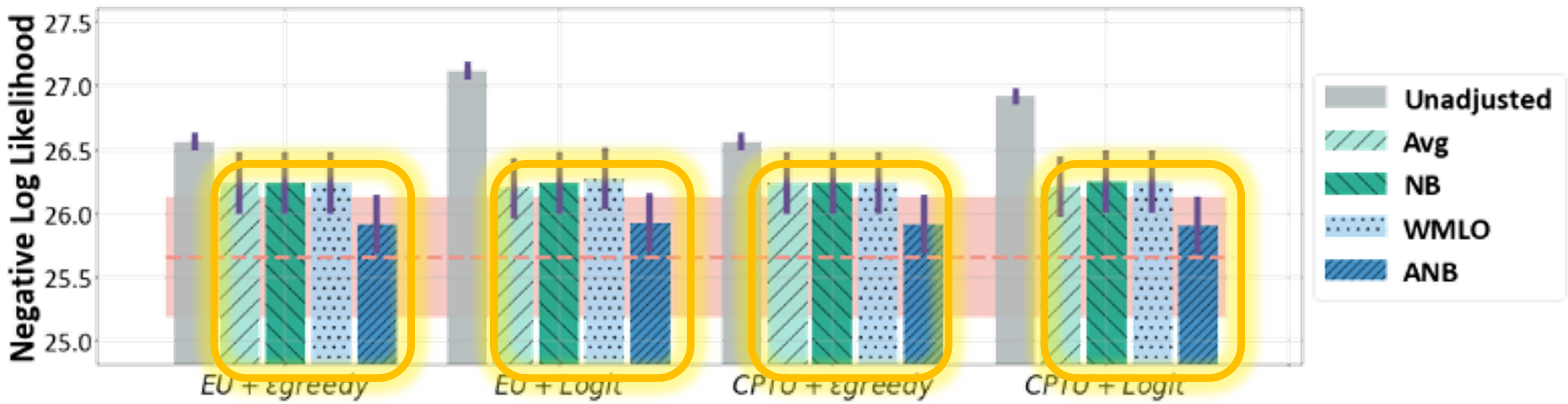


Selection Probability	
Accept AI:	0.60
Reject AI:	0.40

⇨

Selection Probability	
Accept AI:	**0.85**
Reject AI:	0.15

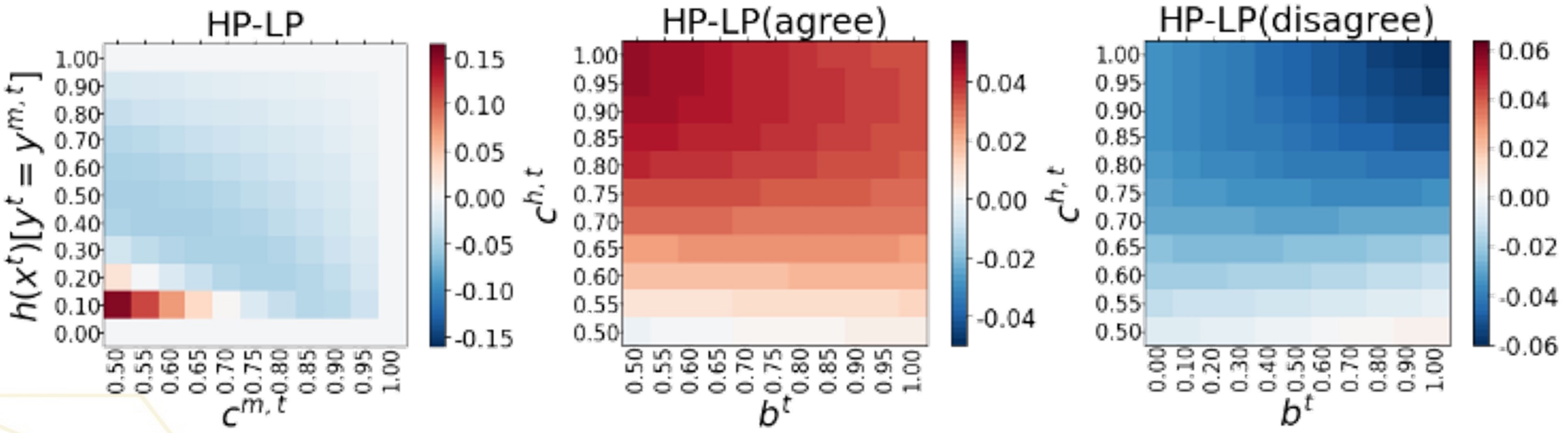
Comparing Model Performance



(b) High Penalty Treatment

- Average human decision makers incorporate their own judgement to decide whether to accept an AI model's recommendation.
- Such judgement may influence their behavior through multiple steps in their cognitive reasoning processes.

Comparing Behavior Across Treatments



HP-LP

HP-LP(agree)

HP-LP(disagree)

(a) Difference in b^t (b) Difference in r^t_{accept} (c) Difference in r^t_{accept}

When the decision stakes are higher, people

- lower their belief in the AI model being correct, and
- rely more on their own judgements

Summary

- We try to **quantitatively** model humans' adoption behavior of the AI recommendation in AI-assisted decision making.

- Our results show that the human-adjusted models outperform models that are only based on the AI model's outputs.

- When the decisions stakes are larger, people tend to lower their belief in AI recommendation's correctness and rely more on their own judgement.

Thank you!

Will You Accept the AI Recommendation?

Predicting Human Behavior in AI-Assisted Decision Making

Xinru Wang*, Zhuoran Lu*, Ming Yin