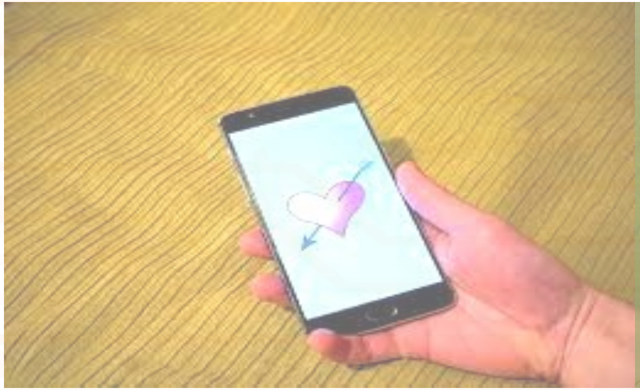
PSYCHONOMIC SOCIETY
62ND ANNUAL MEETING

Metacognitive Bandits: When do humans seek AI assistance?

Aakriti Kumar[1], Trisha Patel[2], Aaron Benjamin[2], Mark Steyvers[1]

University of California Irvine[1]
University of Illinois Urbana Champaign[2]

Abstract # 17



AI is everywhere

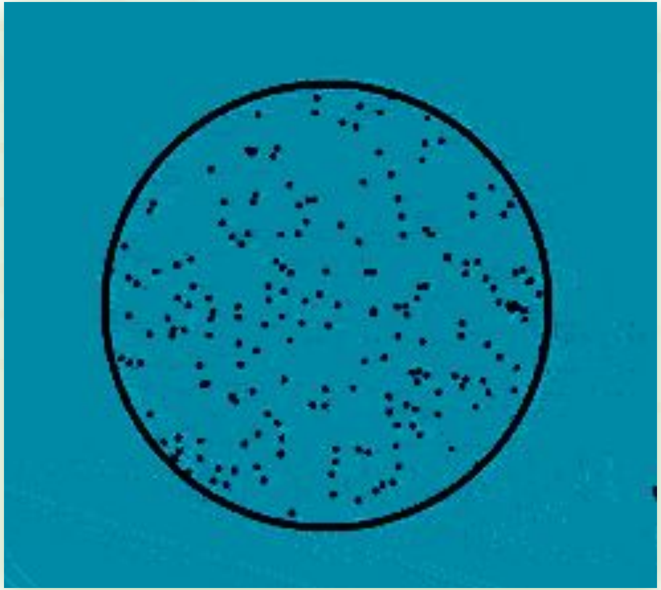
#psynom21

But how do humans use AI assistance?

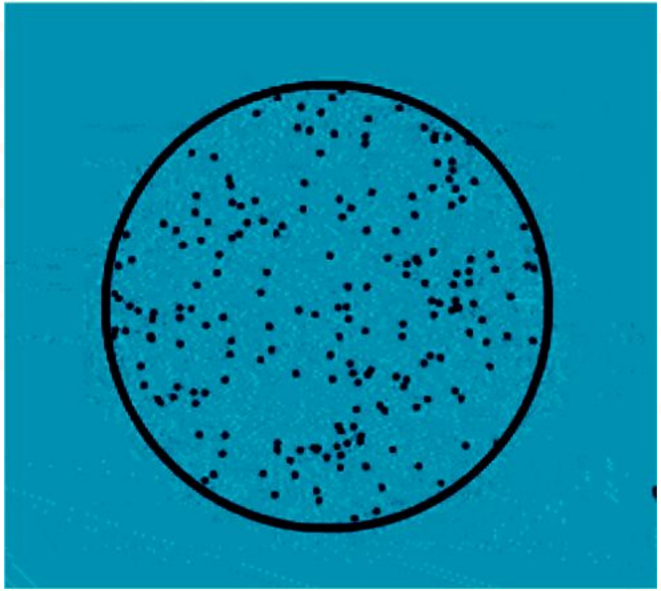
How do humans decide when to ask for AI advice?

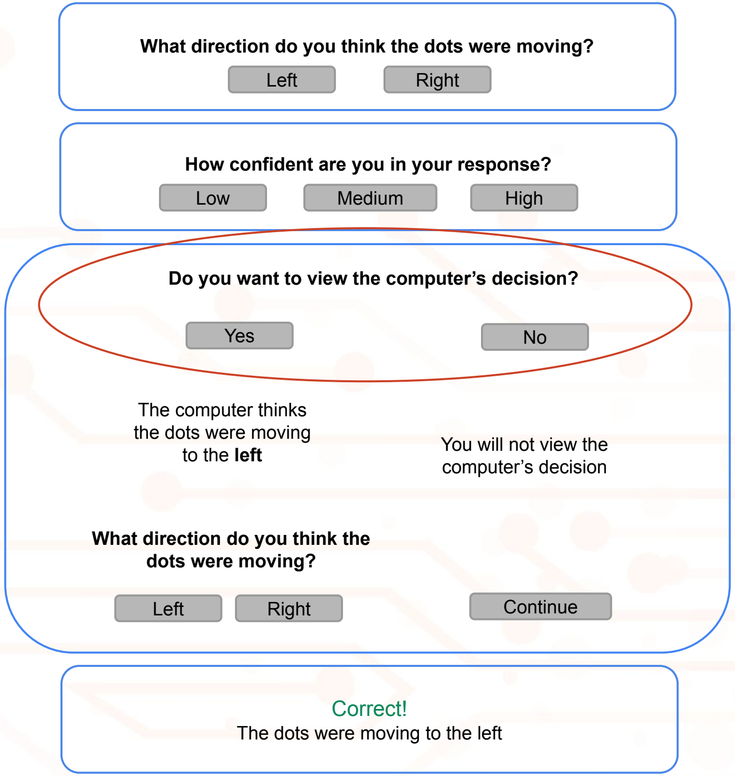
A Motivating Example

A Motivating Example



A Motivating Example





What direction do you think the dots were moving?

How confident are you in your response?

Do you want to view the computer's decision?

The computer thinks the dots were moving to the **left**

You will not view the computer's decision

What direction do you think the dots were moving?

Correct!
The dots were moving to the left

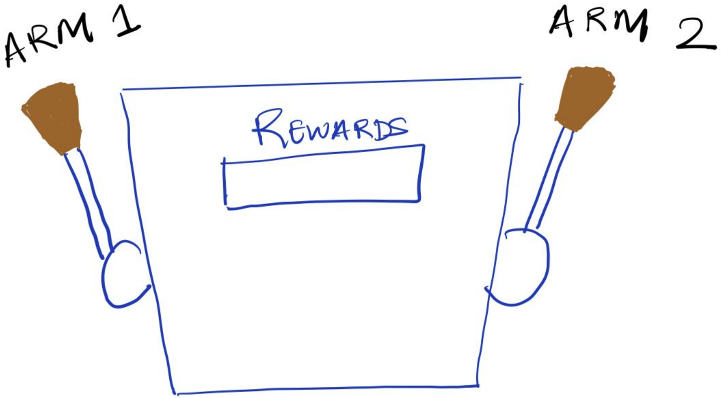
How do humans decide when to take AI advice?

- Combination of **explore/exploit** sequential-decision making and **metacognition**

How do humans decide when to take AI advice?

- Combination of **explore/exploit** sequential-decision making and **metacognition**

- Two-armed bandit



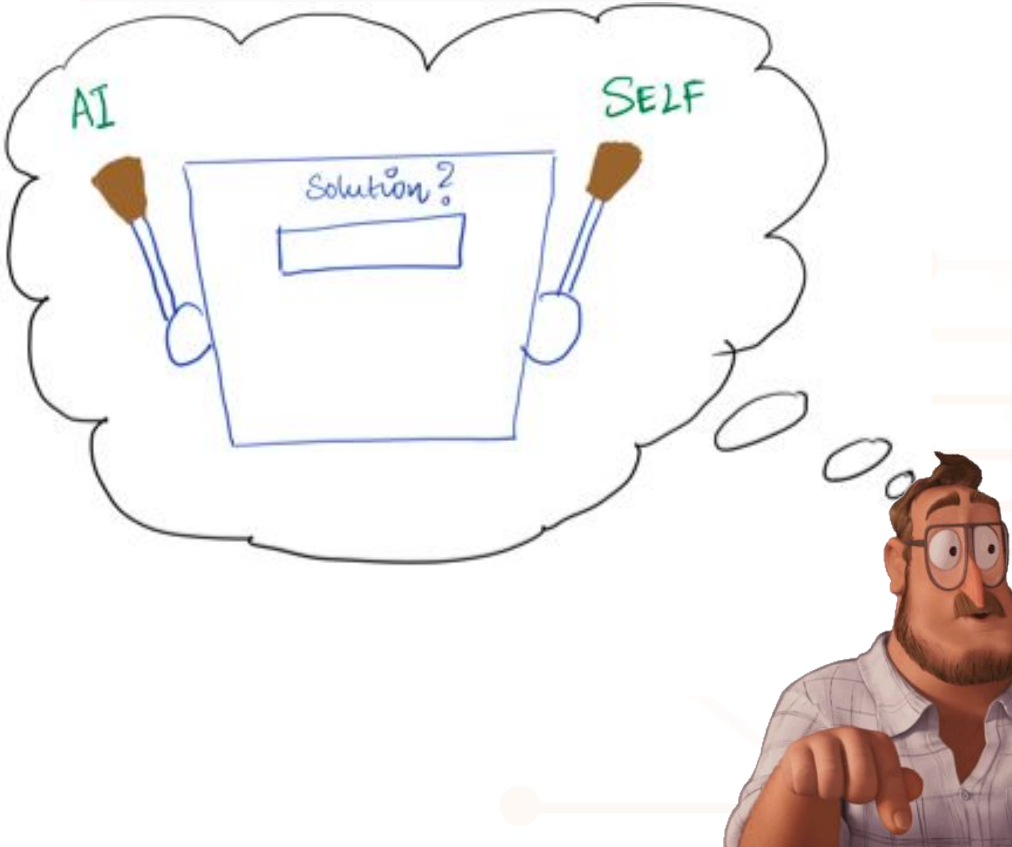
How do humans decide when to take AI advice?

- Combination of **explore/exploit** sequential-decision making and **metacognition**

- Two-armed bandit

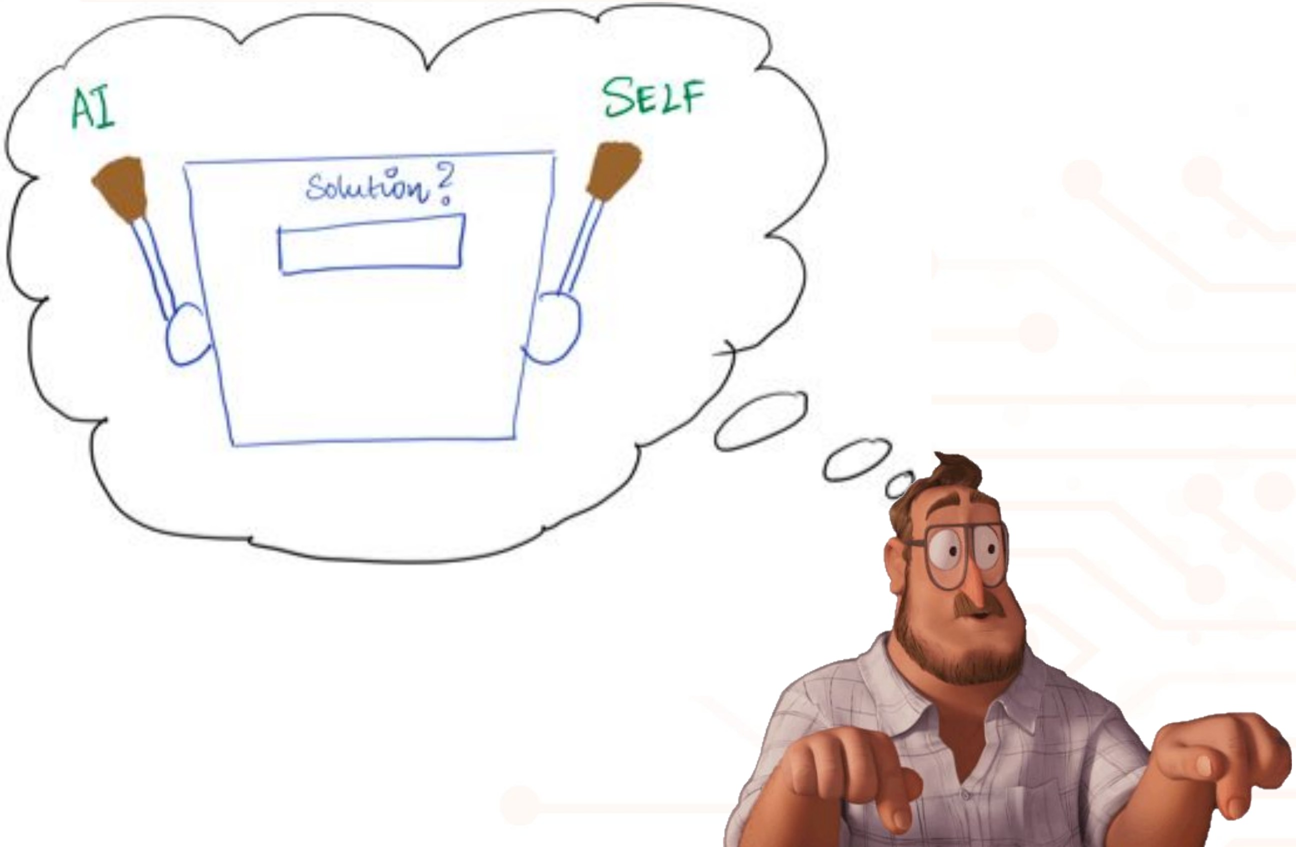
AI
Self
REWARDS

#psynom21



#psynom21

Metacognitive Bandits!



Metacognitive Bandits

- Performance history of both arms (AI and self) to inform the decision of arm selection

Metacognitive Bandits

- Performance history of both arms (AI and self) to inform the decision of arm selection

- Utility inferred by the human

Metacognitive Bandits

- Performance history of both arms (AI and self) to inform the decision of arm selection

- Utility inferred by the human

- Upper confidence bounds (UCB)

#psynom21

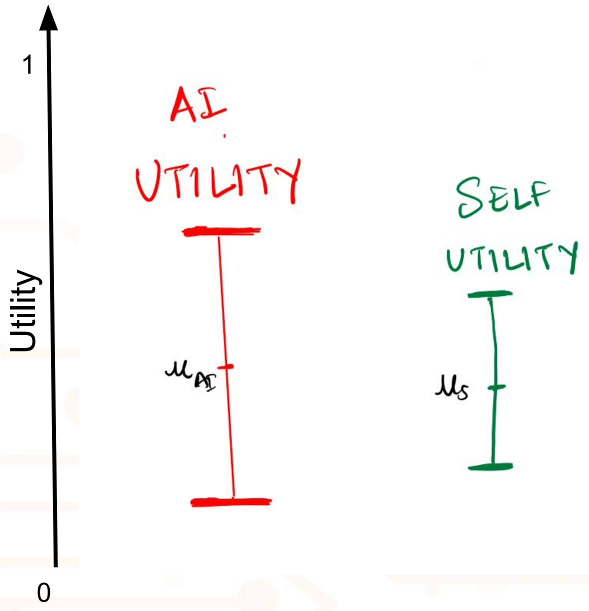
Pavlidis, N. G., Tasoulis, D. K., & Hand, D. J. (2008)

Metacognitive Bandits

- Performance history of both arms (AI and self) to inform the decision of arm selection

- Utility inferred by the human

- Upper confidence bounds (UCB)



#psynom21

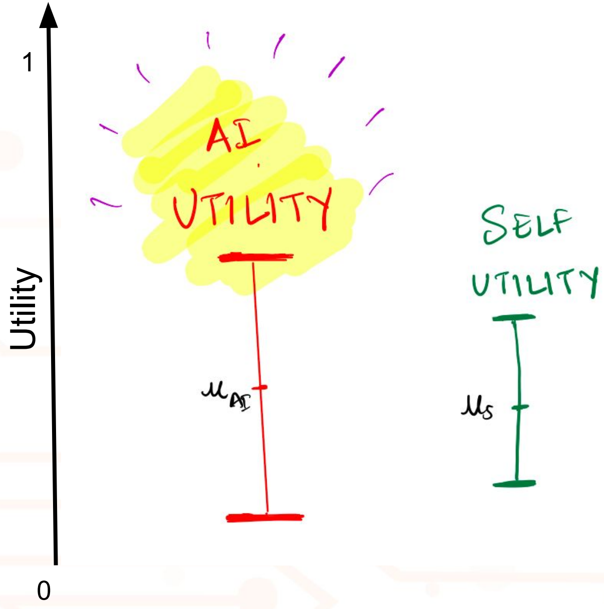
Pavlidis, N. G., Tasoulis, D. K., & Hand, D. J. (2008)

Metacognitive Bandits

- Performance history of both arms (AI and self) to inform the decision of arm selection

- Utility inferred by the human

- Upper confidence bounds (UCB)



#psynom21

Pavlidis, N. G., Tasoulis, D. K., & Hand, D. J. (2008)

Metacognitive Bandits

- Performance history of both arms (AI and self) to inform the decision of arm selection

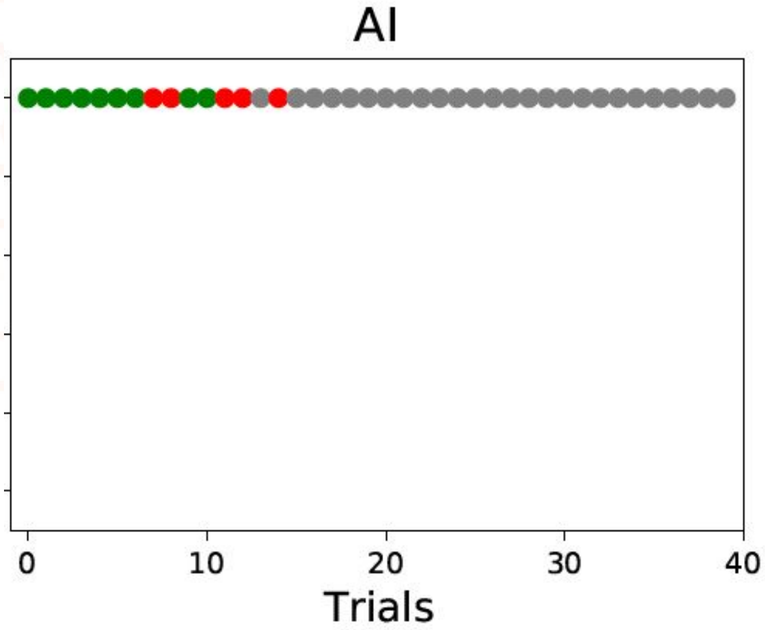
- Utility inferred by the human

- Upper confidence bounds (UCB)

- Incorporate Rasch model to account for difficulty of the trial

Example Runs from Metacognitive Bandit

Example run from metacognitive bandit



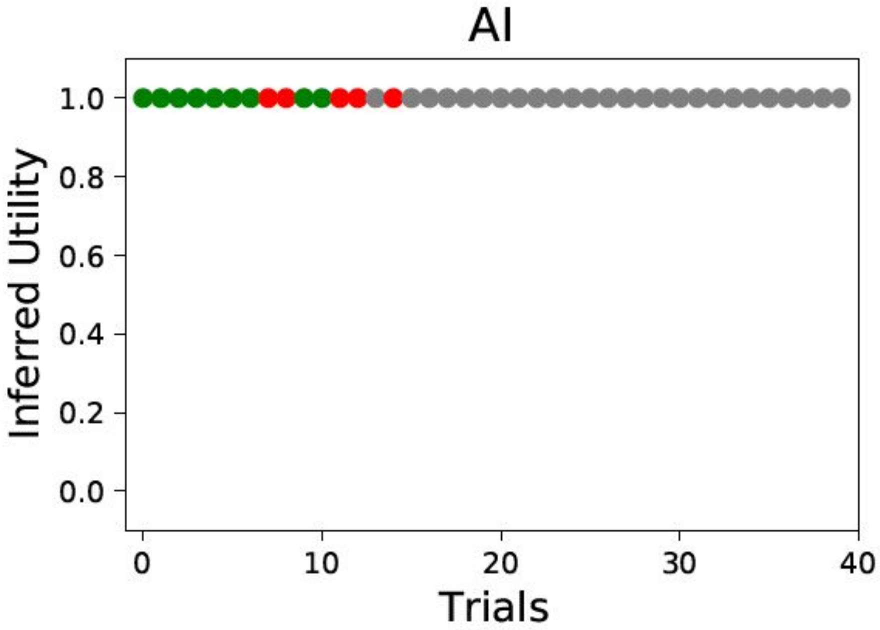
Green - Correct Response

Red - Incorrect Response

Gray - Advice not Solicited

#psynom21

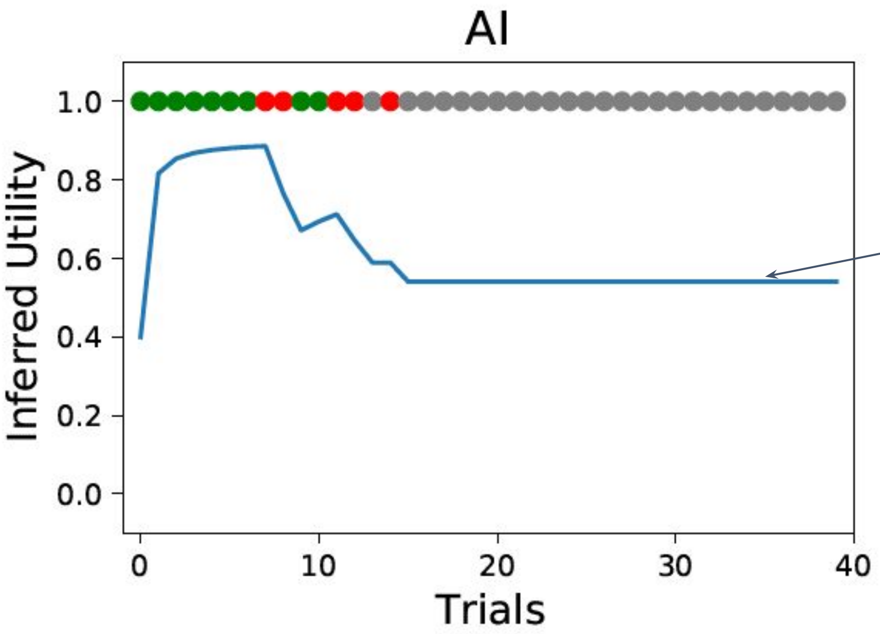
Example run from metacognitive bandit



Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

Example run from basic metacognitive bandit

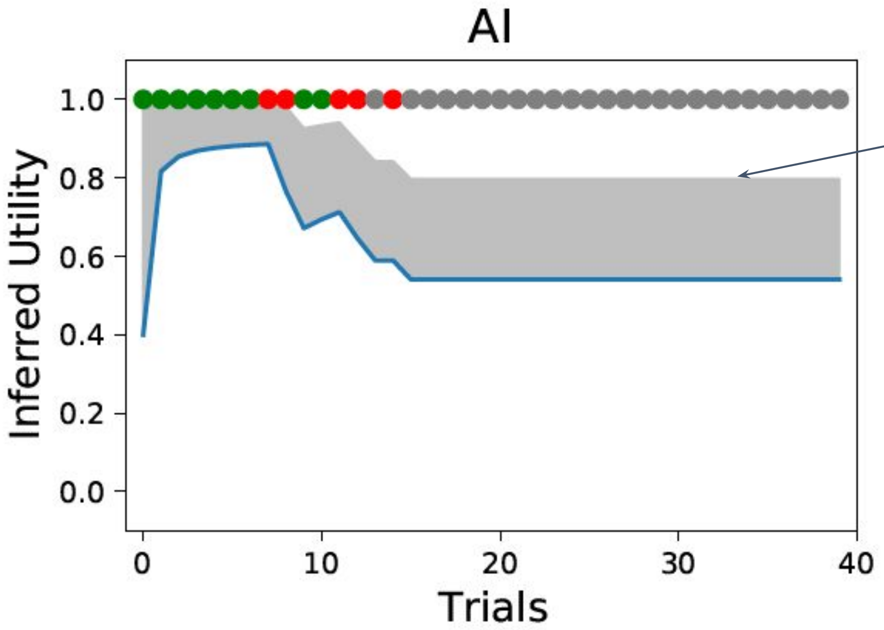


Mean utility inferred by the human

Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

Example run from metacognitive bandit

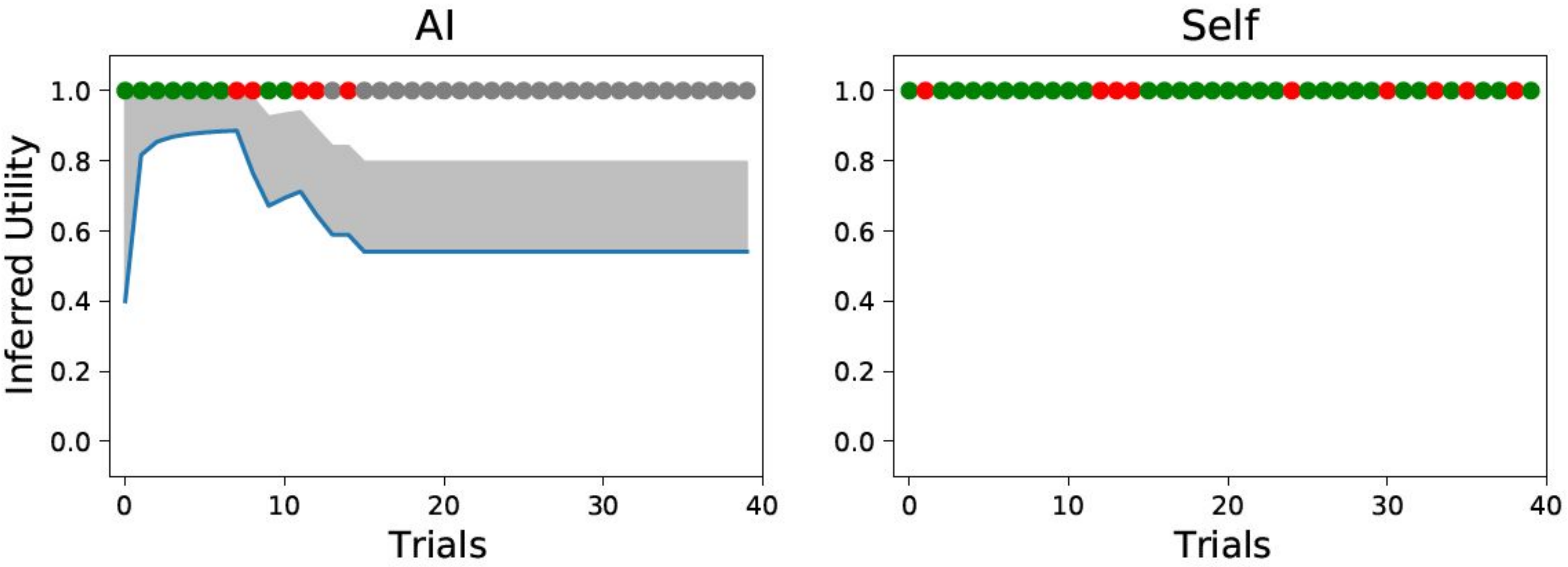


Upper bound of the posterior uncertainty

Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

Example run from metacognitive bandit



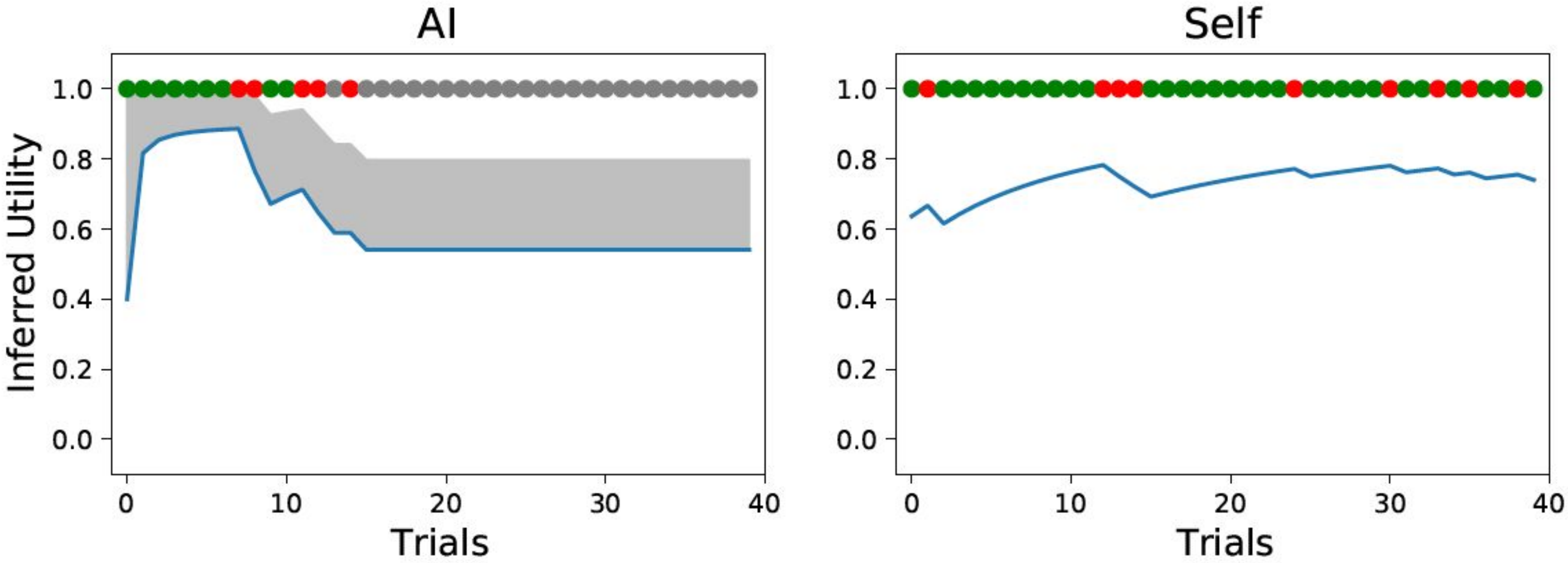
Green - Correct Response

Red - Incorrect Response

Gray - Advice not Solicited

#psynom21

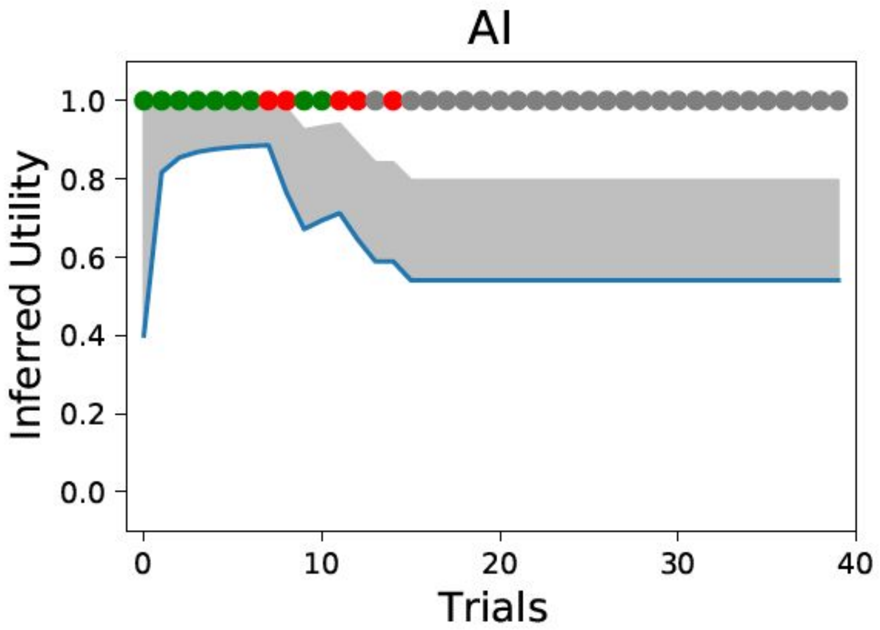
Example run from metacognitive bandit

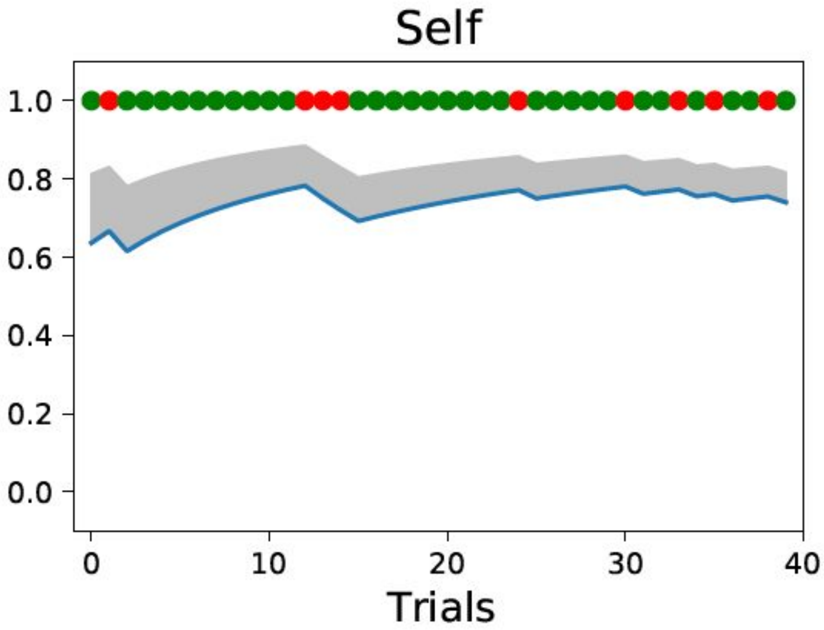


Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

Example run from metacognitive bandit

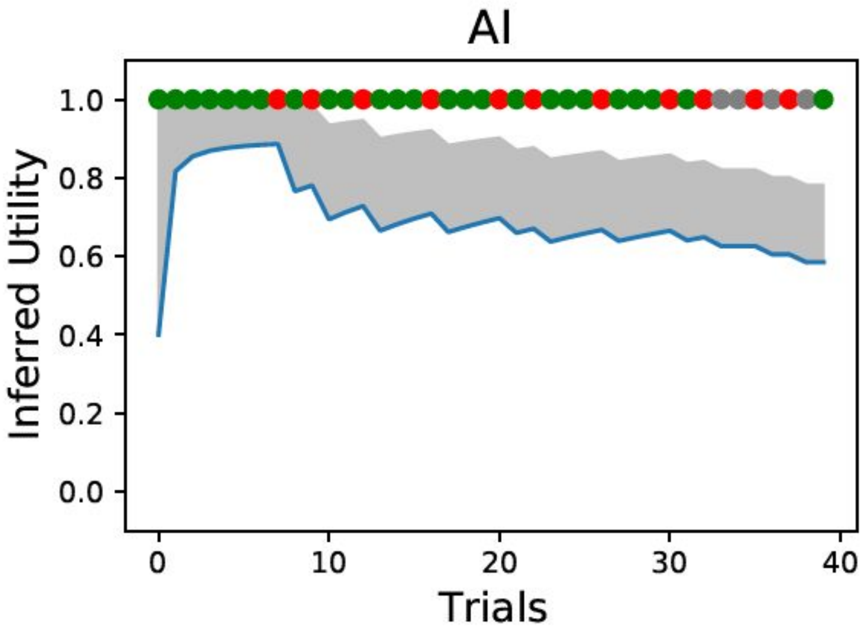


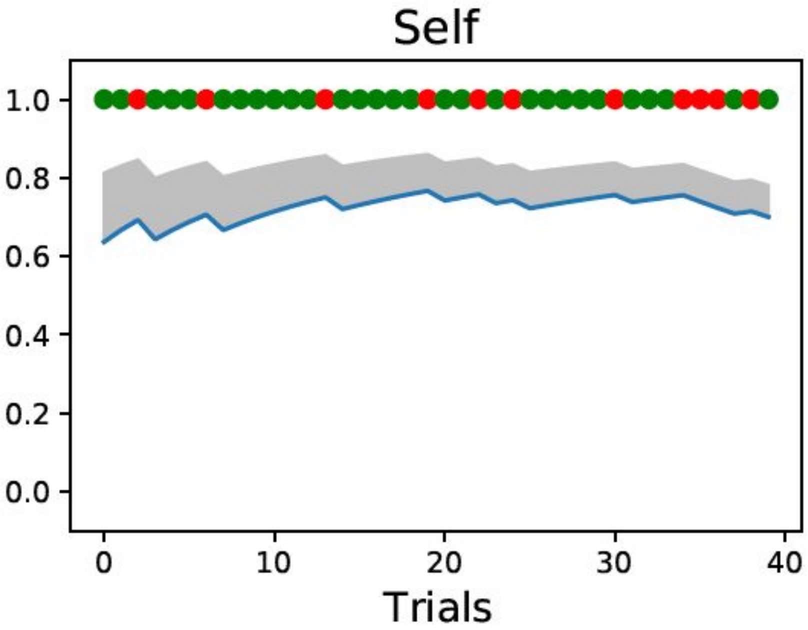


Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

Another example run from metacognitive bandit





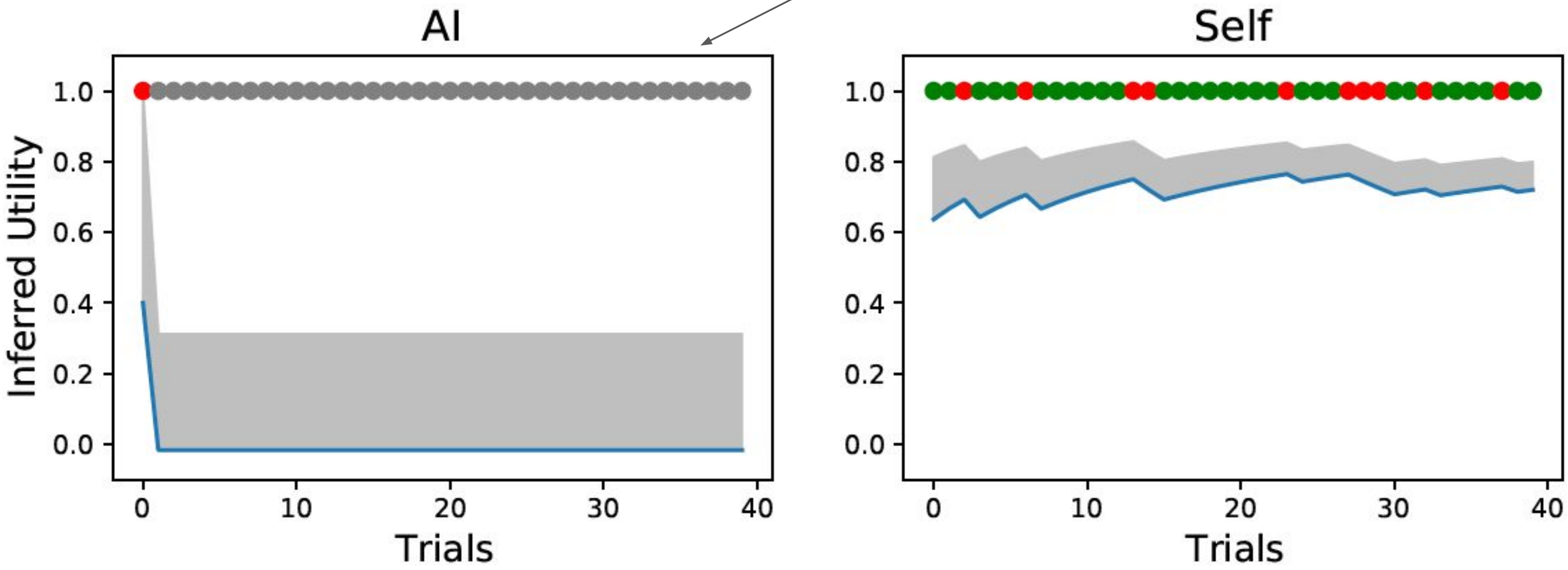
Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

Algorithm Aversion

Another example run from metacognitive bandit

Example of Algorithm Aversion

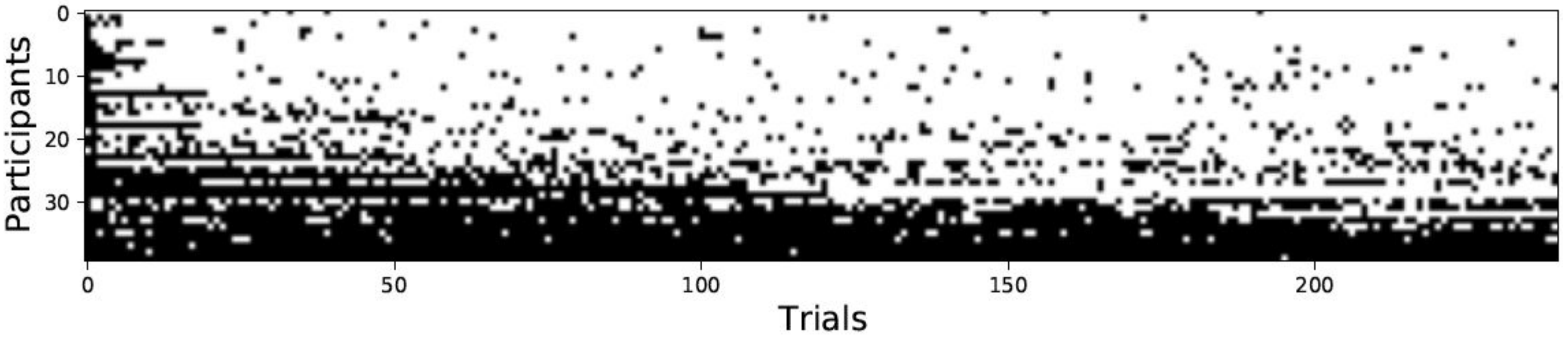


Green - Correct Response
Red - Incorrect Response
Gray - Advice not Solicited

#psynom21

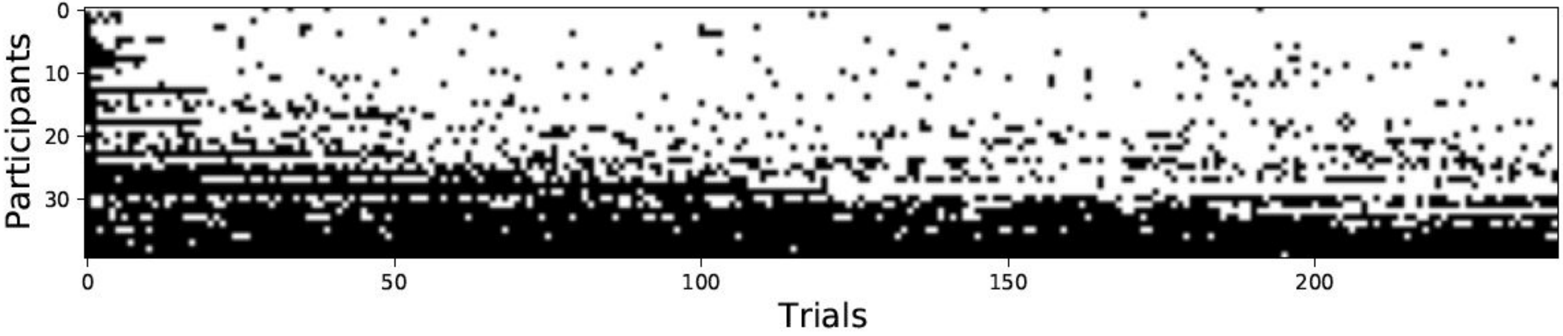
Advice soliciting behavior for actual and simulated participants on 240 trials

Empirical data

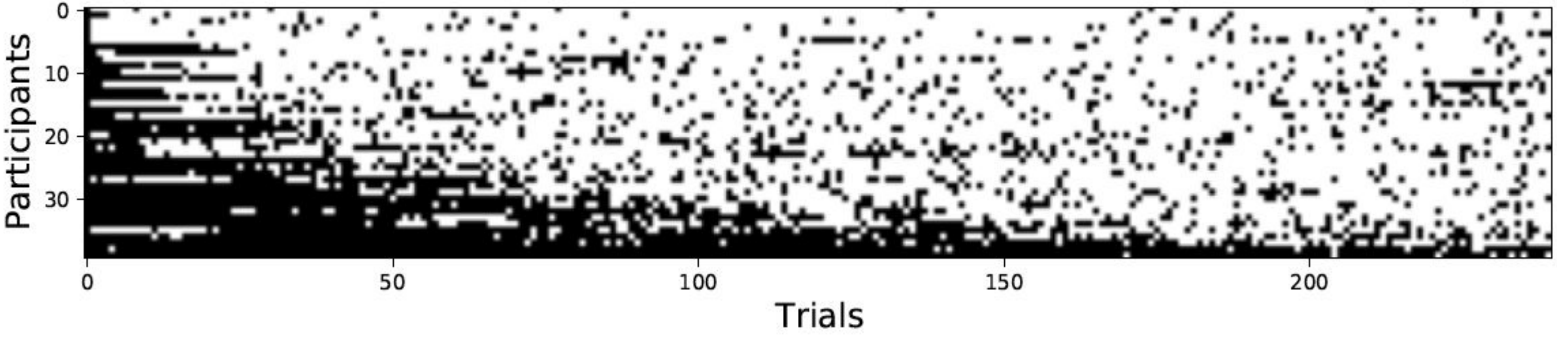


Advice soliciting behavior for actual and simulated participants on 240 trials

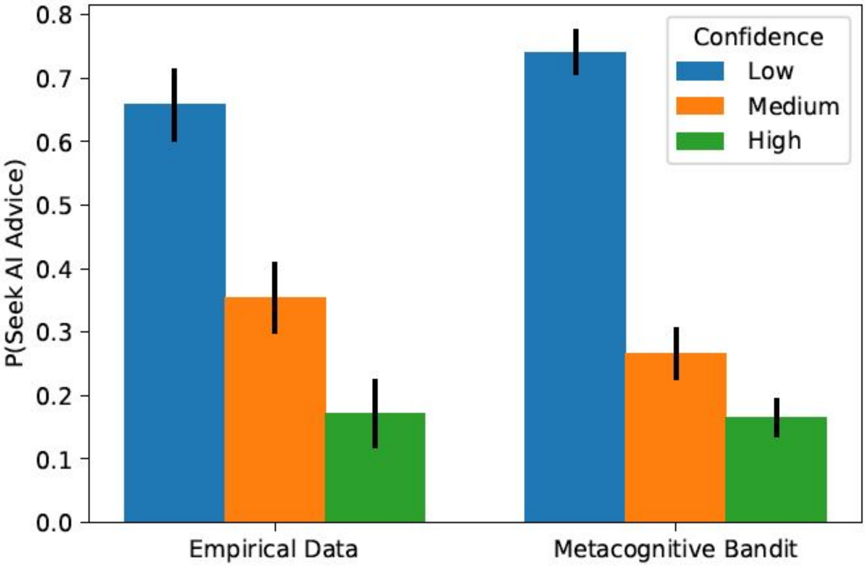
Empirical data



Metacognitive Bandit

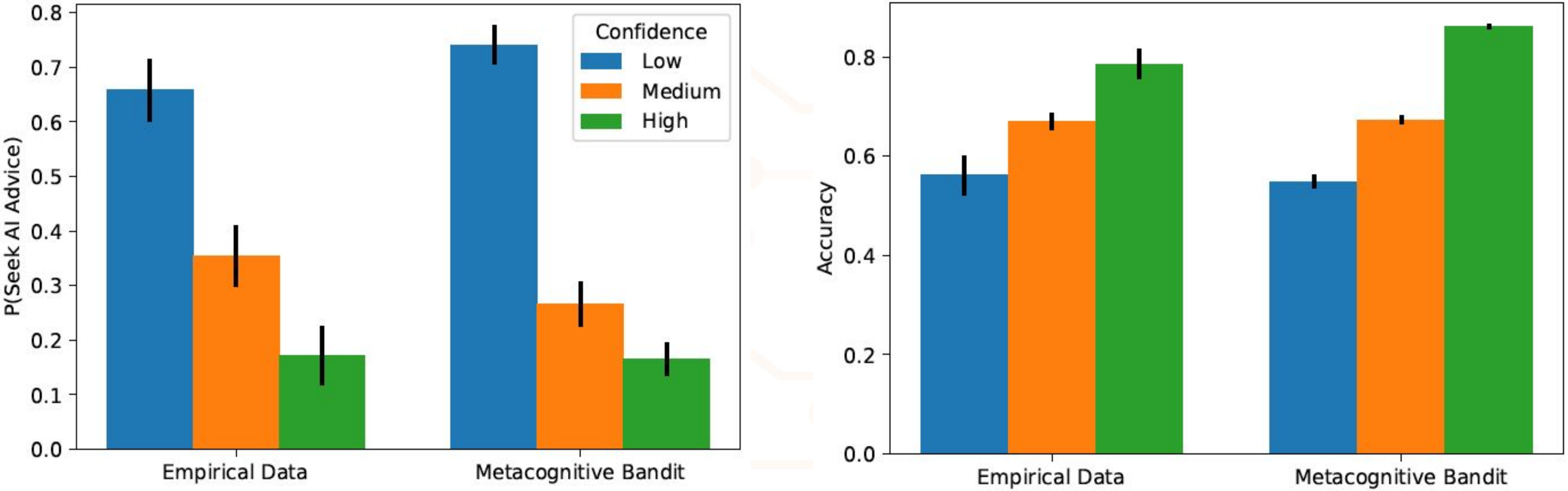


Model captures qualitative trends



#psynom21

Model captures qualitative trends



Limitations and Future Work

- Use real ML algorithms
- How AI advice is integrated in decision-making?

Thank you!

Questions?