Improving Human Sequential Decision-Making with Reinforcement Learning

Park Sinchaisri
Berkeley Haas

with Hamsa Bastani (Wharton)
& Osbert Bastani (Penn)

Learning is Costly

2+ years

to be fully productive

$1,286/worker

training expenses

- Training Magazine 2019

Learning is Costly

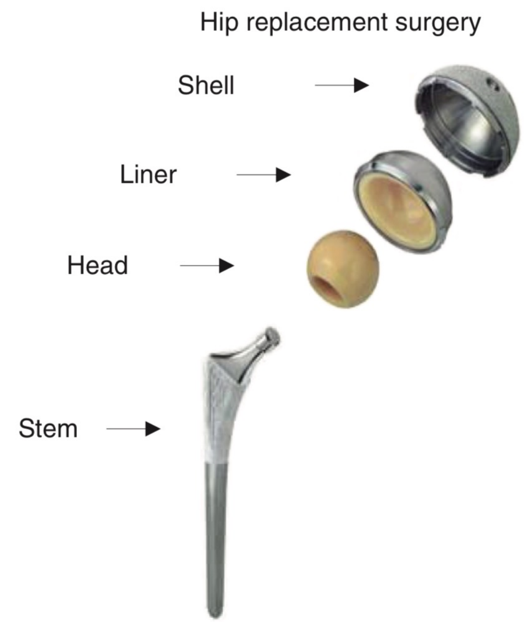
2+ years

to be fully productive

$1,286/worker

training expenses

- Training Magazine 2019



New device = **+32.4%**

surgery duration

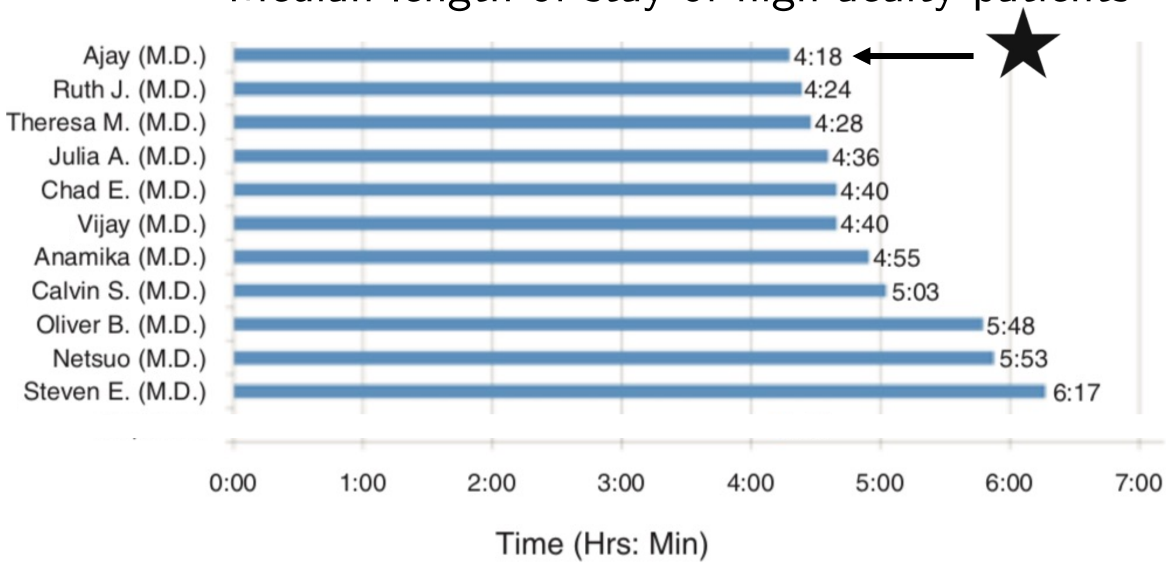
- Ramdas et al. 2018

Also – Tucker et al 2002, Ibanez et al 2017, Gurvich et al 2019, Bavafa & Jonasson 2020, Bloom et al 2020, ...

Learning from Experts

Learning from Experts

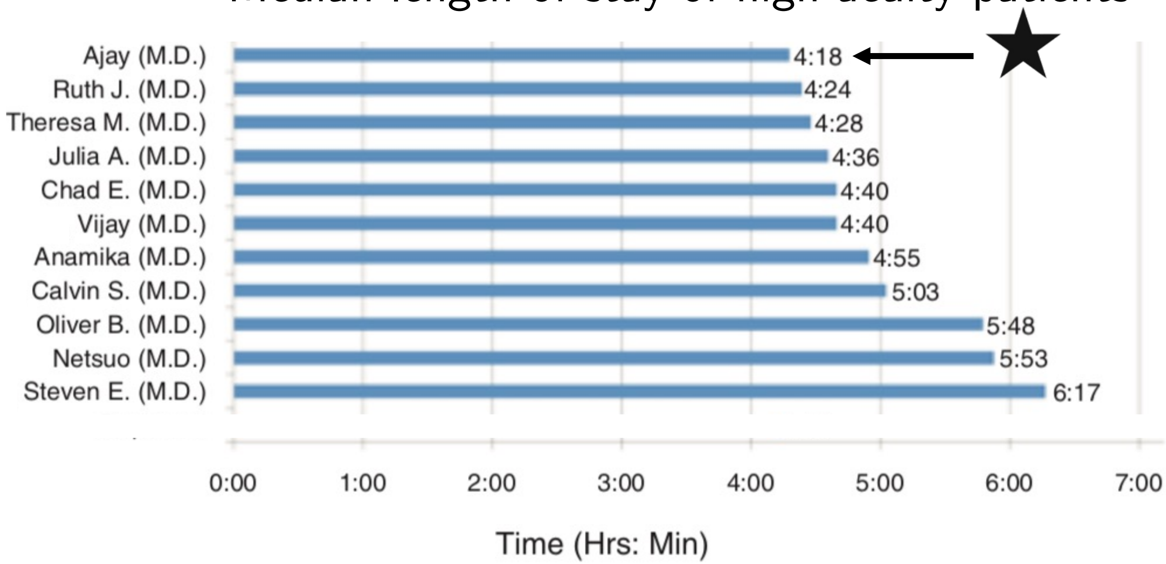
Median length of stay of high acuity patients



- Song et al. 2018

Learning from Experts

Median length of stay of high acuity patients



+10.9%
productivity
- Song et al. 2018

Also – Chan et al 2014, Herkenhoff et al 2018, Tan & Netessine 2019, Jarosch et al 2019, …

Learning from Experts

+10.9%

productivity

- Song et al. 2018

Learning from Experts

Median length of stay of high acuity patients

+10.9%

productivity

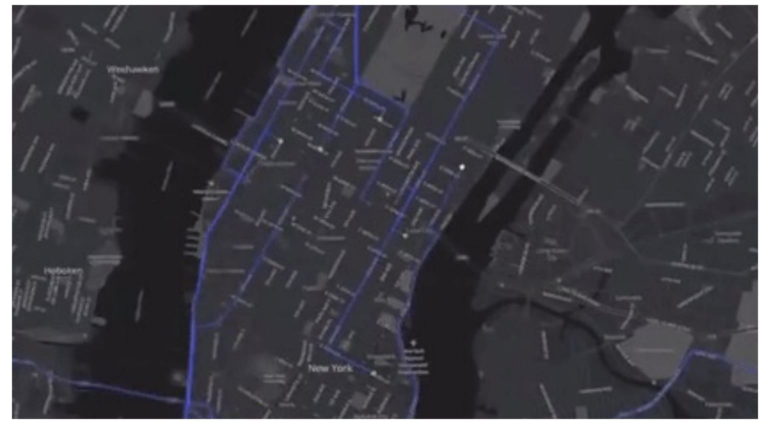
- Song et al. 2018

Trace Data is Everywhere

Physicians

• ROACH,TRISTIN	Fibrinogen, INR, PT, PTT AMD_996304_76		MILLER,ALEX,MD *status: Unreviewed*	05•19•17
• ROACH,TRISTIN	Lipitor 80 mg		MILLER,ALEX,MD *status: Unreviewed*	05•18•17
• LEON,ERIN	Geriatric Wellness Visit		JONES,CAMERON,MD *status: Unreviewed*	05•16•17
• BECK,ALIVIA	Zocor 20 mg		JACK,JACK,MD *status: Unreviewed, held*	05•18•17
NORTON,BETHANY	Norvasc 10 mg		MILLER,ALEX,MD *status: Unreviewed*	05•18•17
MONTGOMERY,BLAINE	Glucophage 850 mg		OSHEA,JAMIE,MD *reviewed by: PPMD_AKN...* *status: Reviewed*	05•18•17
KLECK,MICHAEL	Office Visit - Abbreviated		JONES,CAMERON,MD *reviewed by: SUSAN* *status: Reviewed*	05•12•17
MCARDLE,HELEN	Office Visit - Mobile		JONES,CAMERON,MD *status: Unreviewed*	05•12•17

Uber Drivers



Trace Data is Everywhere

Physicians

• ROACH,TRISTIN	Fibrinogen, INR, PT, PTT AMD_996304_76	MILLER,ALEX,MD *status: Unreviewed*	05•19•17
• ROACH,TRISTIN	Lipitor 80 mg	MILLER,ALEX,MD *status: Unreviewed*	05•18•17
• LEON,ERIN	Geriatric Wellness Visit	JONES,CAMERON,MD *status: Unreviewed*	05•16•17
• BECK,ALIVIA	Zocor 20 mg	JACK,JACK,MD *status: Unreviewed, held*	05•18•17
NORTON,BETHANY	Norvasc 10 mg	MILLER,ALEX,MD *status: Unreviewed*	05•18•17
MONTGOMERY,BLAINE	Glucophage 850 mg	OSHEA,JAMIE,MD *reviewed by: PPMD_AKN...* *status: Reviewed*	05•18•17
KLECK,MICHAEL	Office Visit - Abbreviated	JONES,CAMERON,MD *reviewed by: SUSAN* *status: Reviewed*	05•12•17
MCARDLE,HELEN	Office Visit - Mobile	JONES,CAMERON,MD *status: Unreviewed*	05•12•17

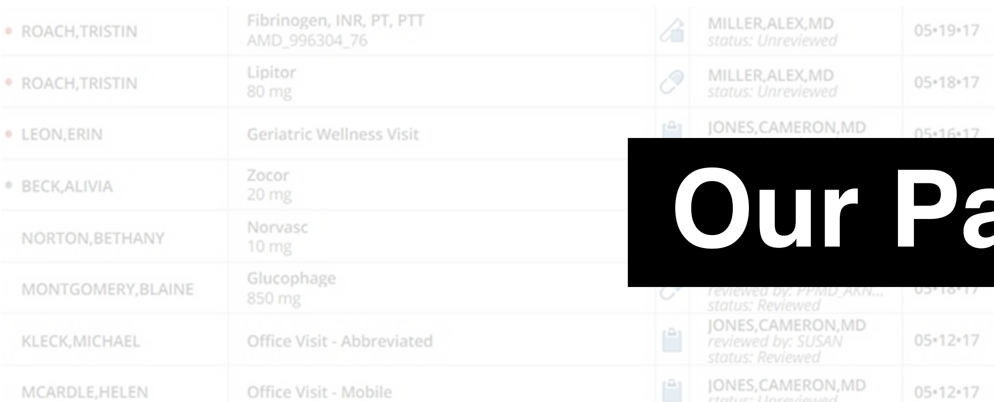
Uber Drivers

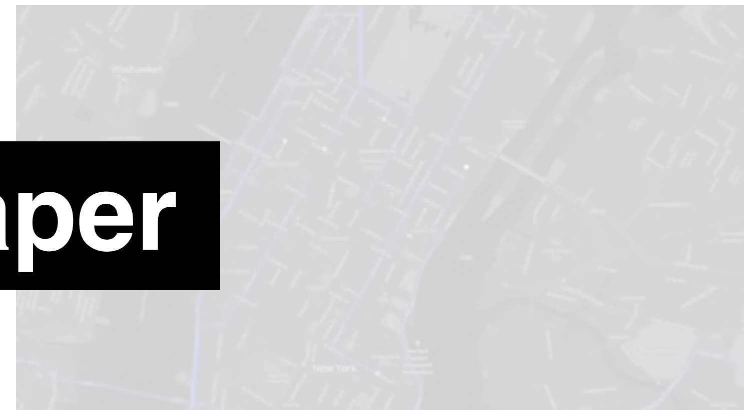
Trace data → Tips

Trace Data is Everywhere

Physicians

Uber Drivers





Our Paper

Trace data → Machine Learning → Tips

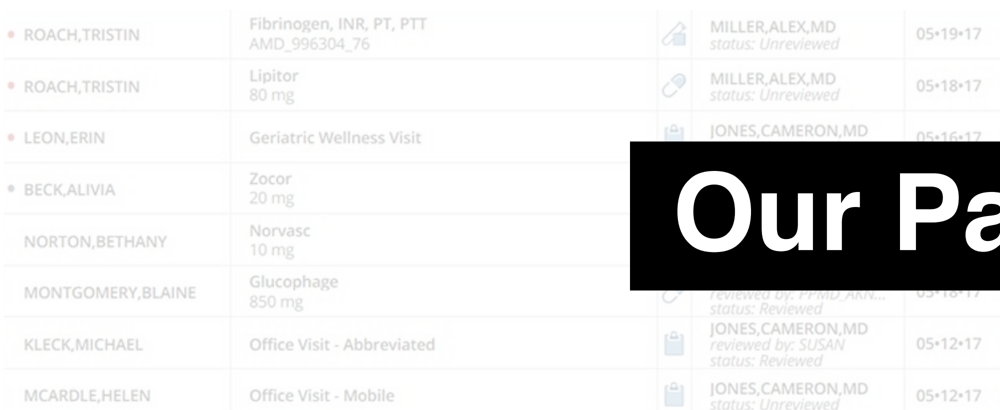
Extract best practices

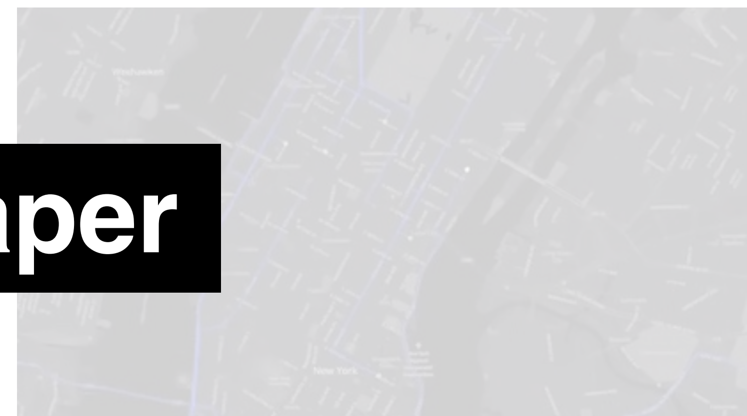
Mine simple tips

Trace Data is Everywhere

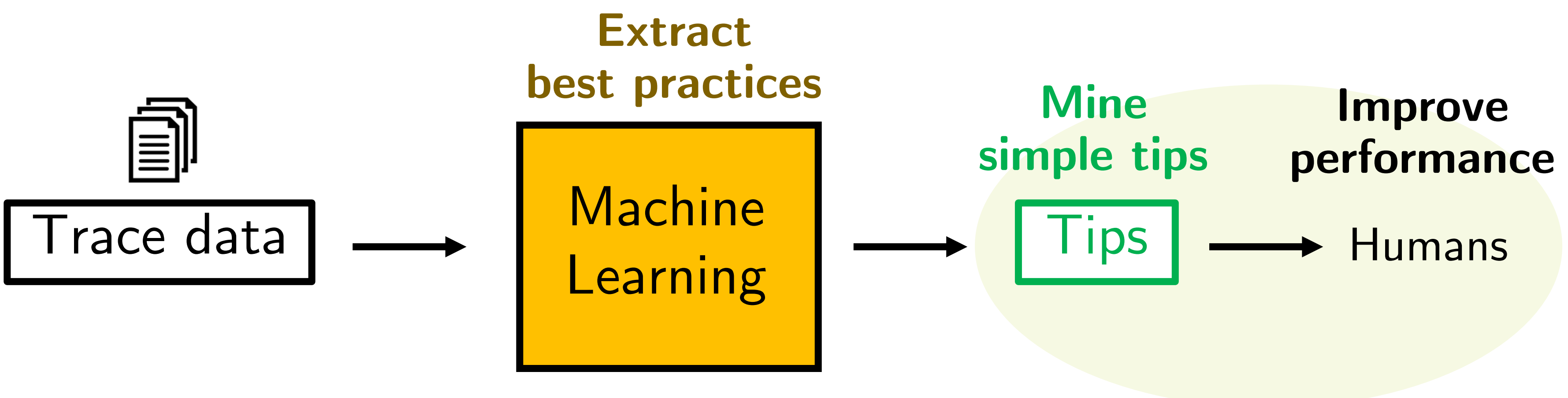
Physicians

Uber Drivers

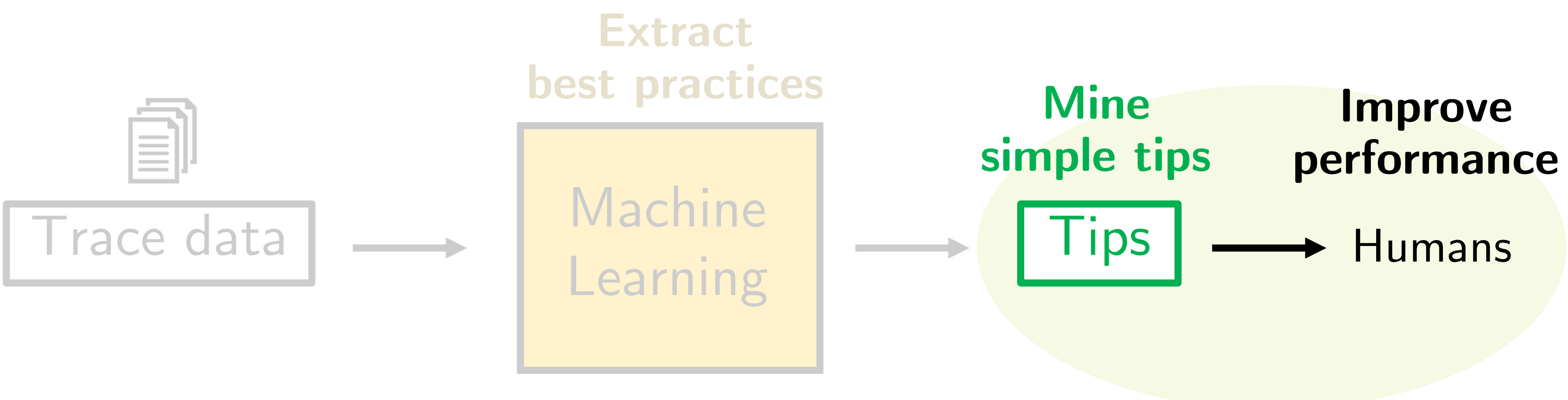




Our Paper

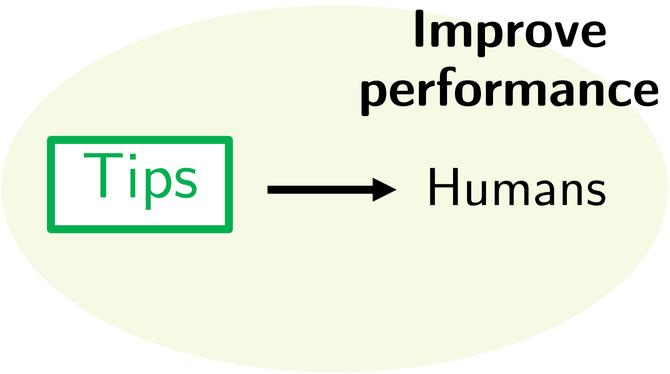


Potential Issues



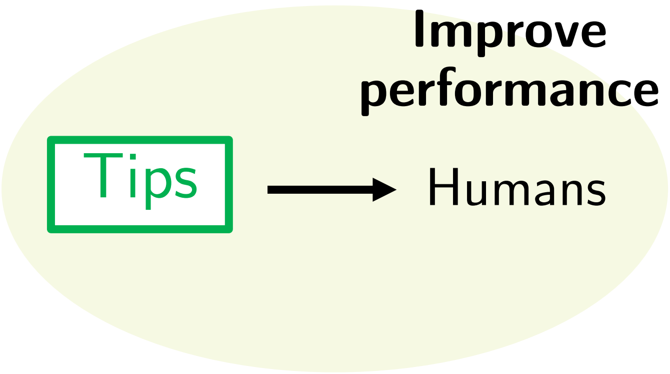
Potential Issues

- Compliance to tips, "algorithm aversion" (e.g., Dietvorst et al 2015)



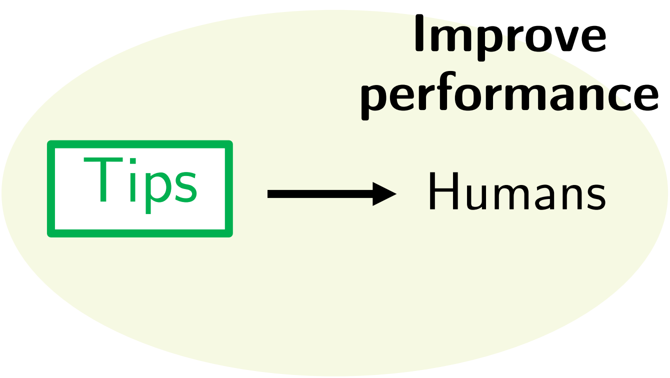
Potential Issues

- Compliance to tips, "algorithm aversion" (e.g., Dietvorst et al 2015)
- Interpretability, inability to precisely implement



Potential Issues

- Compliance to tips, "algorithm aversion" (e.g., Dietvorst et al 2015)
- Interpretability, inability to precisely implement
- Learning curve, spillovers

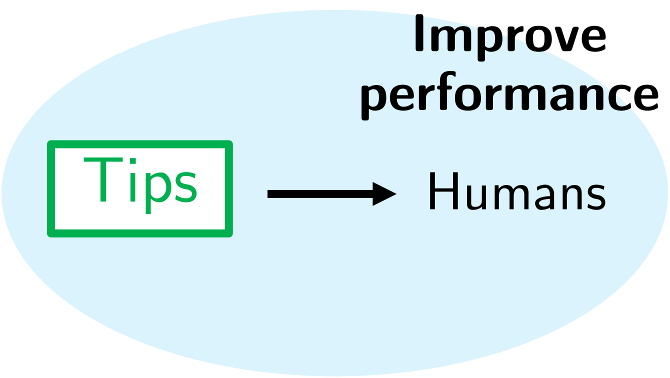


Potential Issues

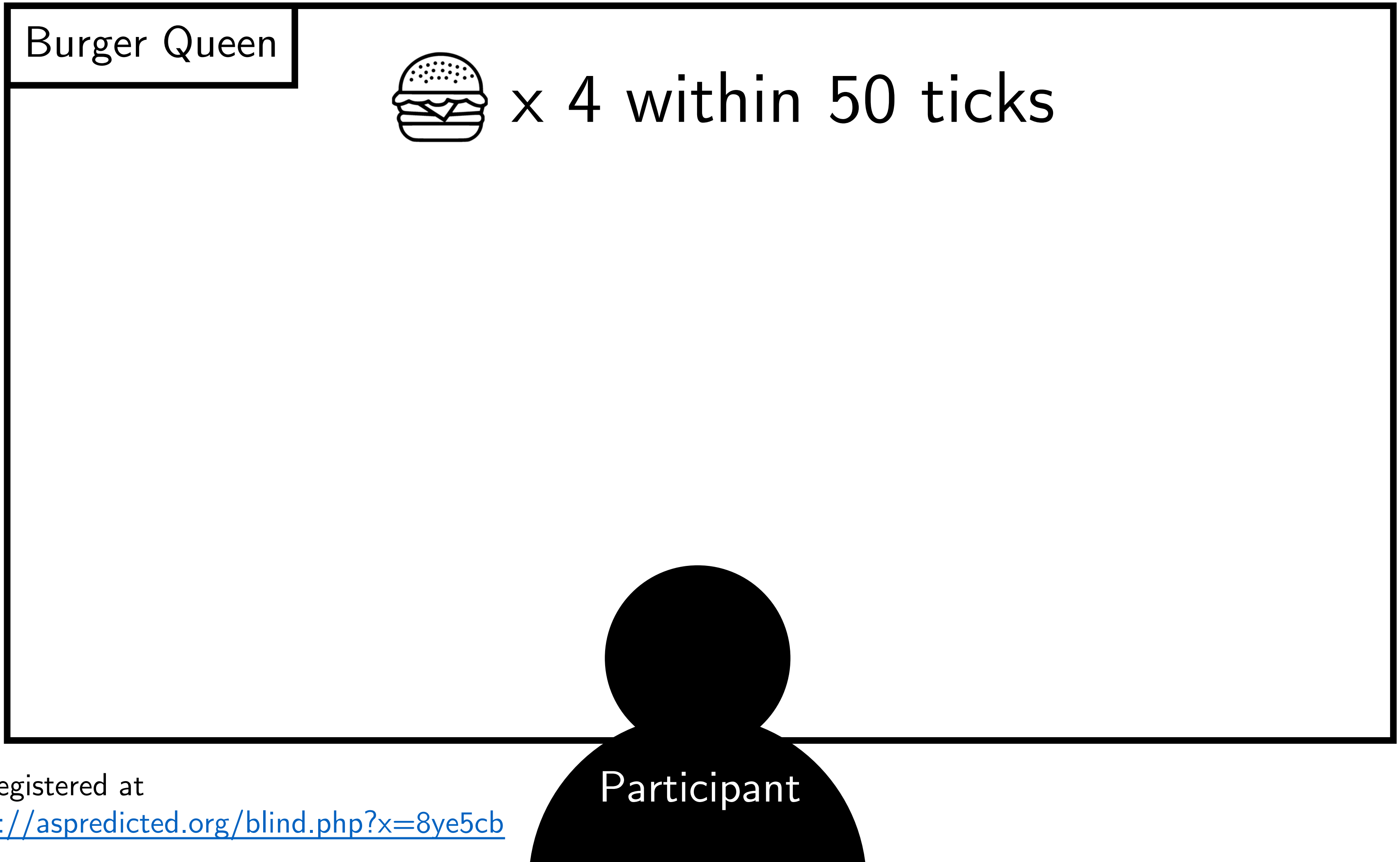
- Compliance to tips, "algorithm aversion" (e.g., Dietvorst et al 2015)
- Interpretability, inability to precisely implement
- Learning curve, spillovers

What We Did:

Controlled environment
to observe human learning
& decision-making

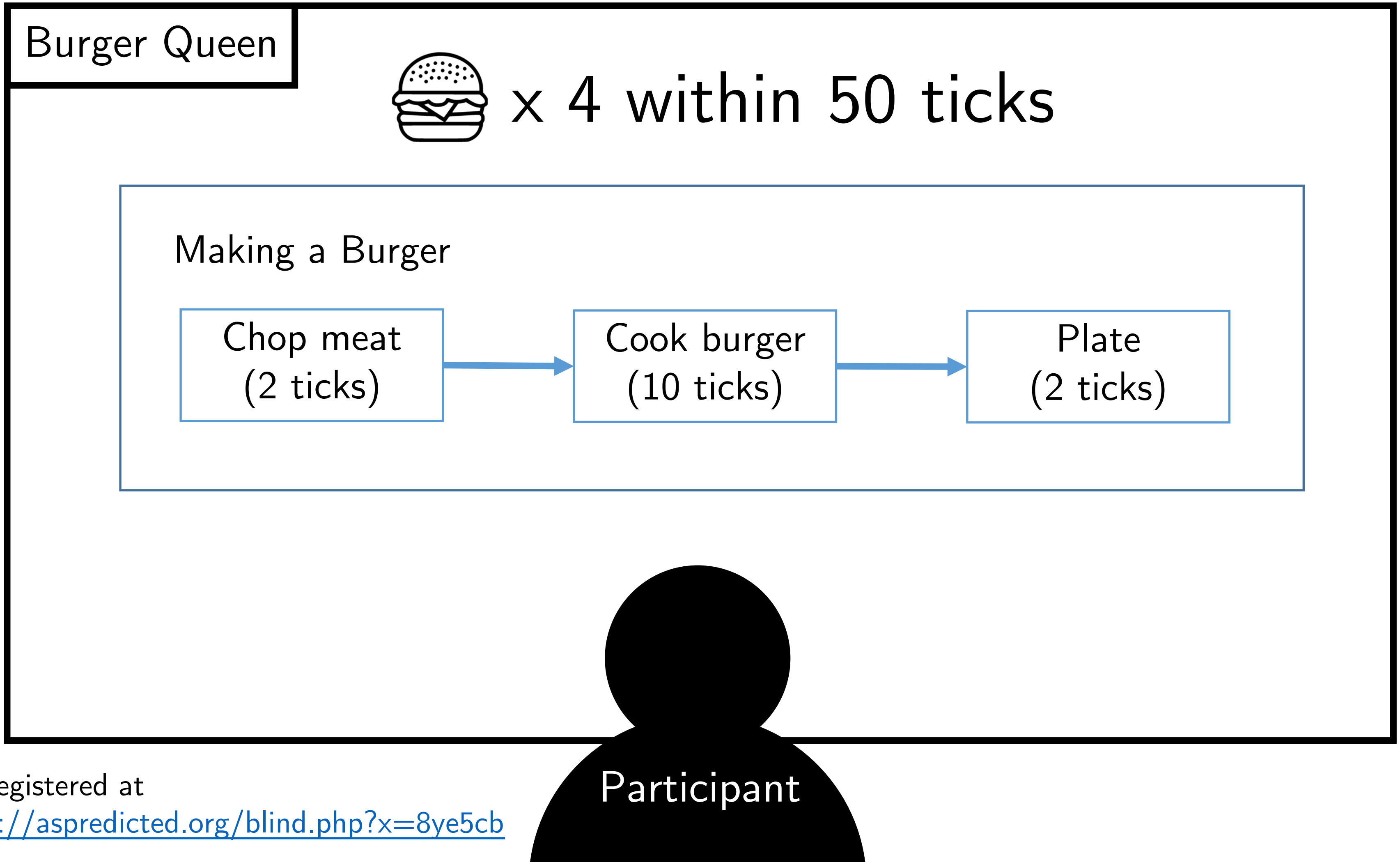


Cooking Game



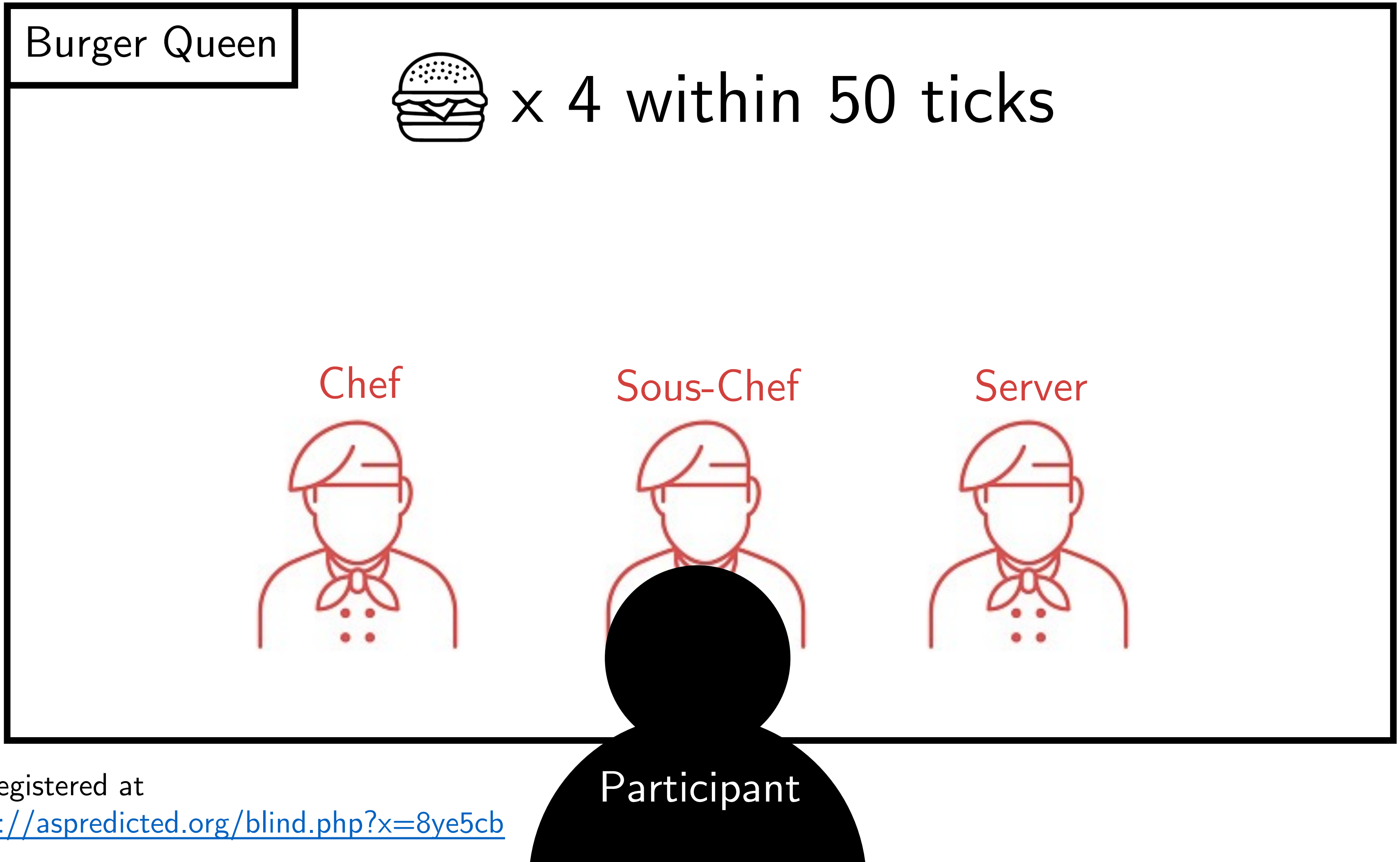
Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

Cooking Game



Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

Cooking Game

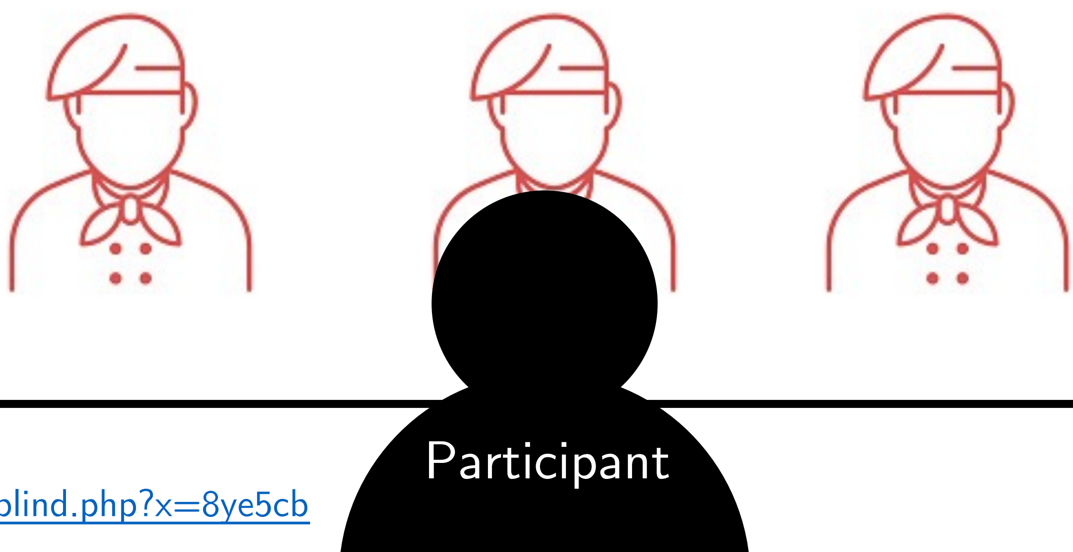


Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

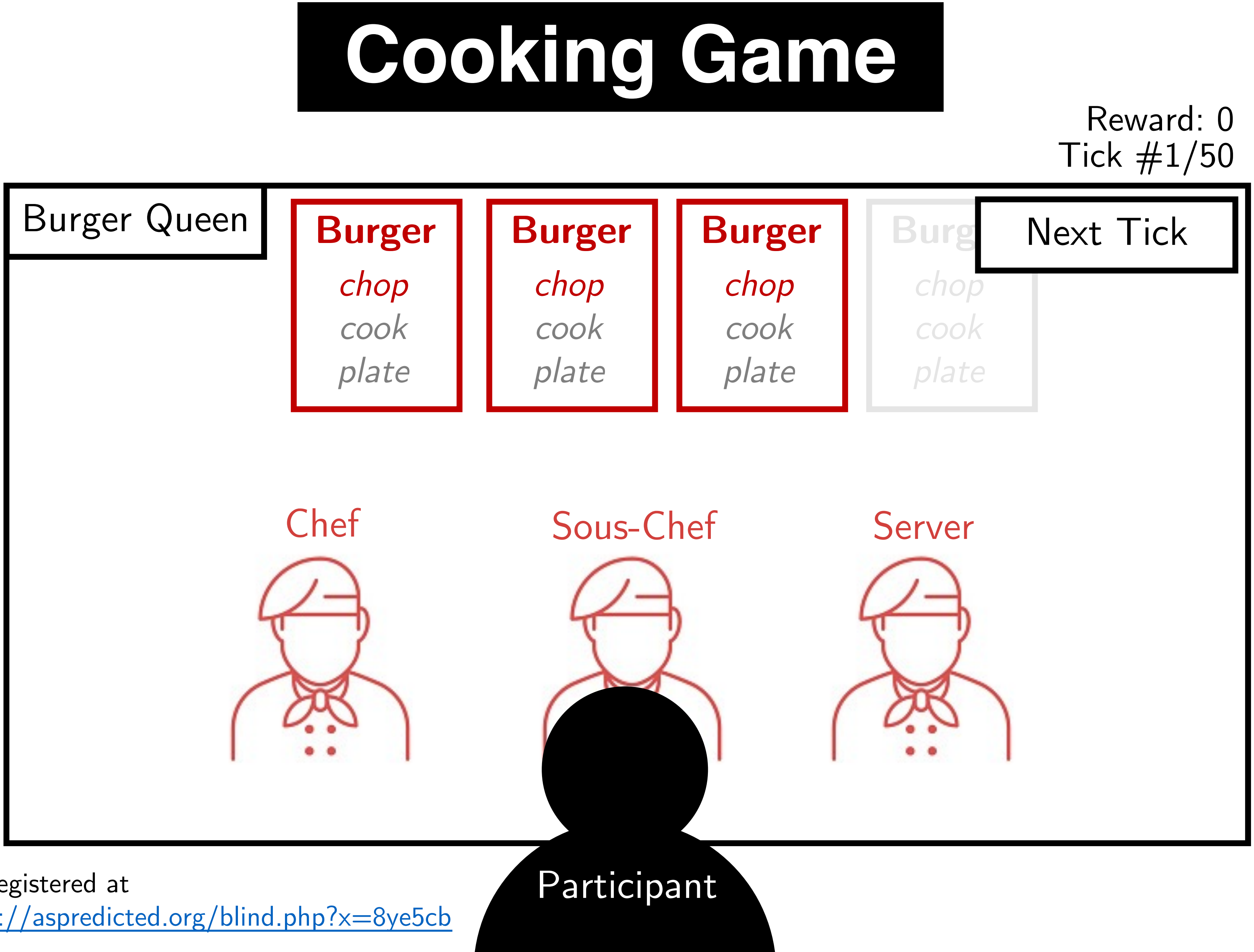
Cooking Game

Burger Queen

Chopping:	Fast	Average	Slow
Cooking:	Fast	Average	Slow
Plating:	Slow	Average	Fast
	Chef	Sous-Chef	Server

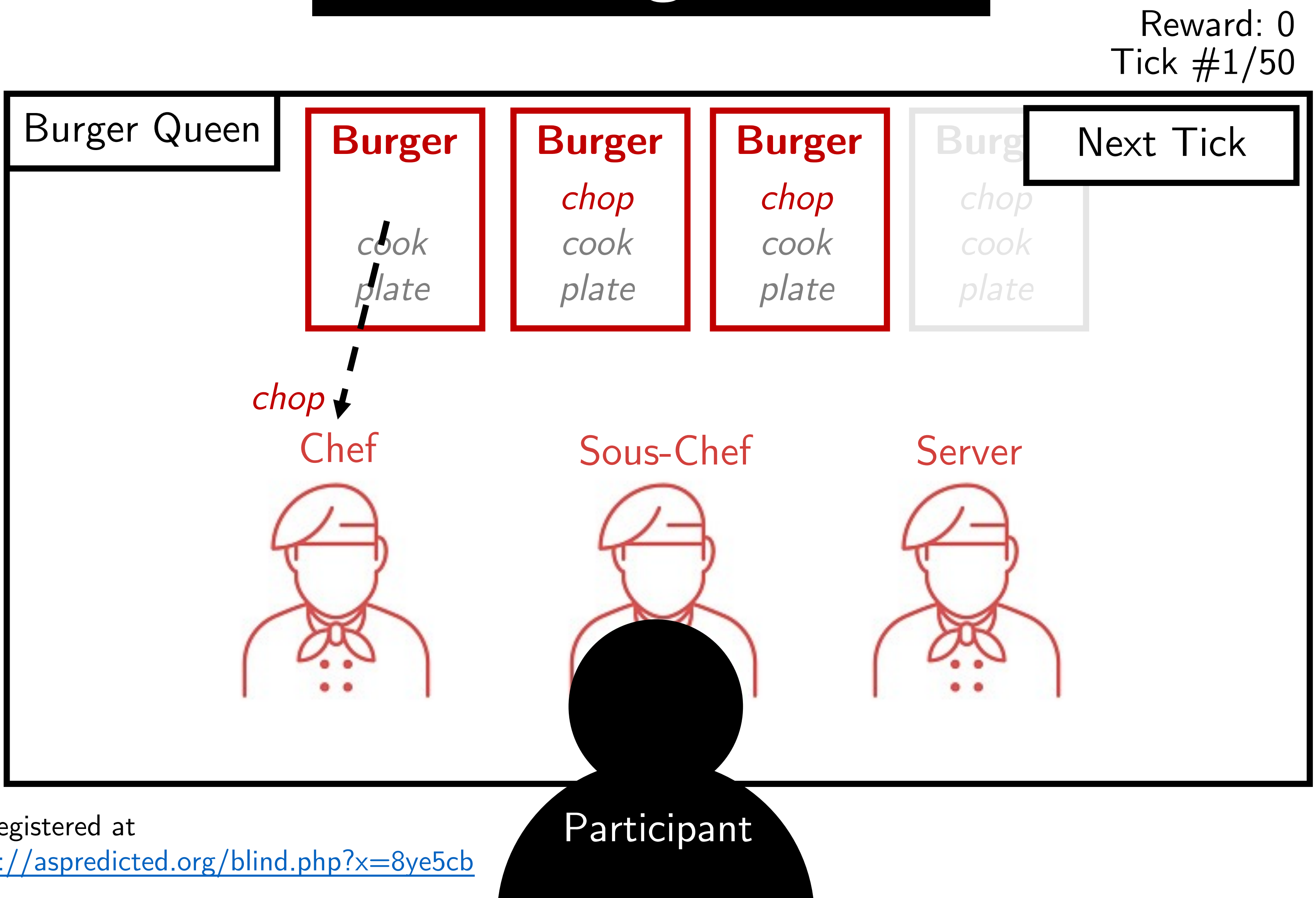


Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb



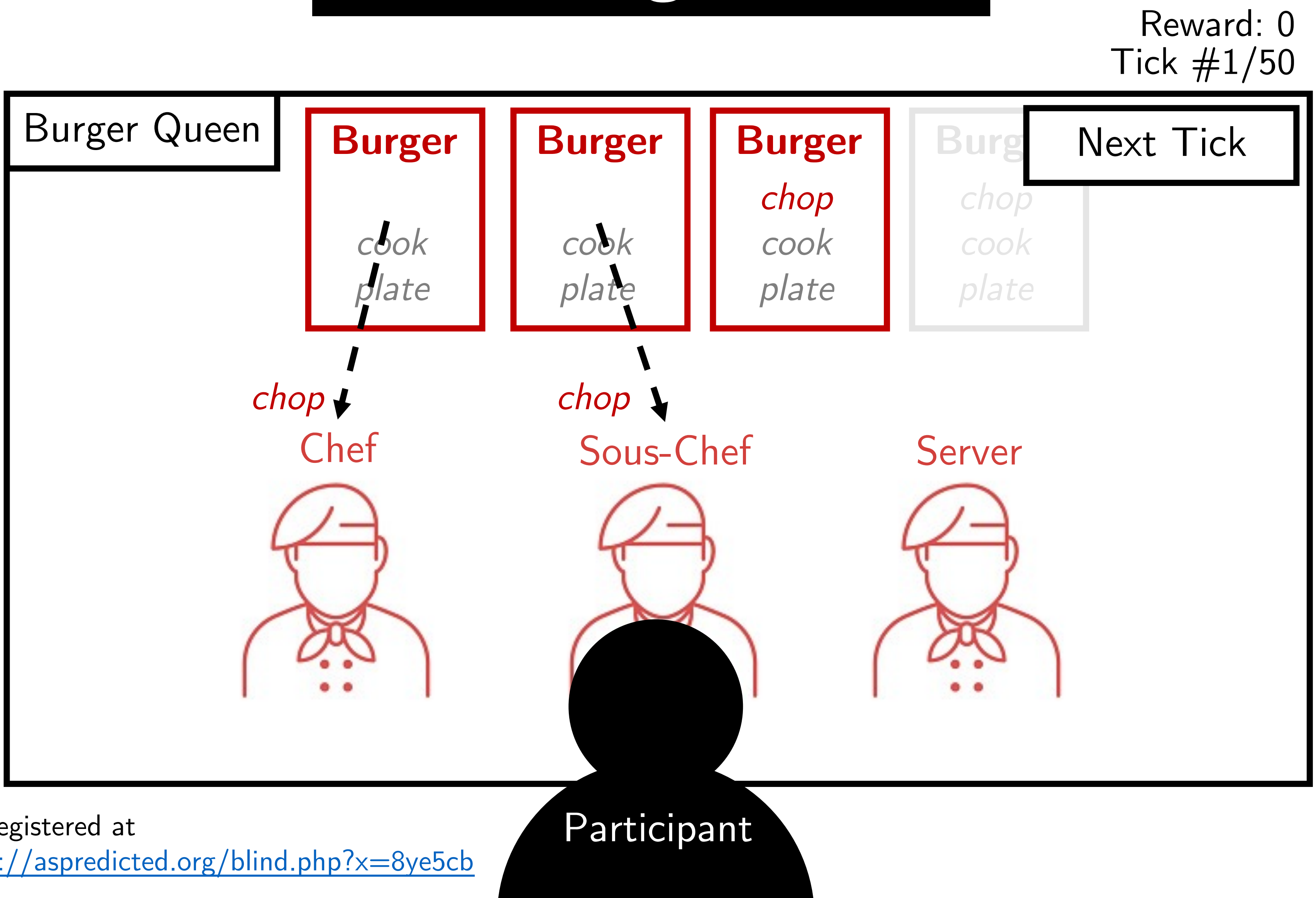
Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

Cooking Game



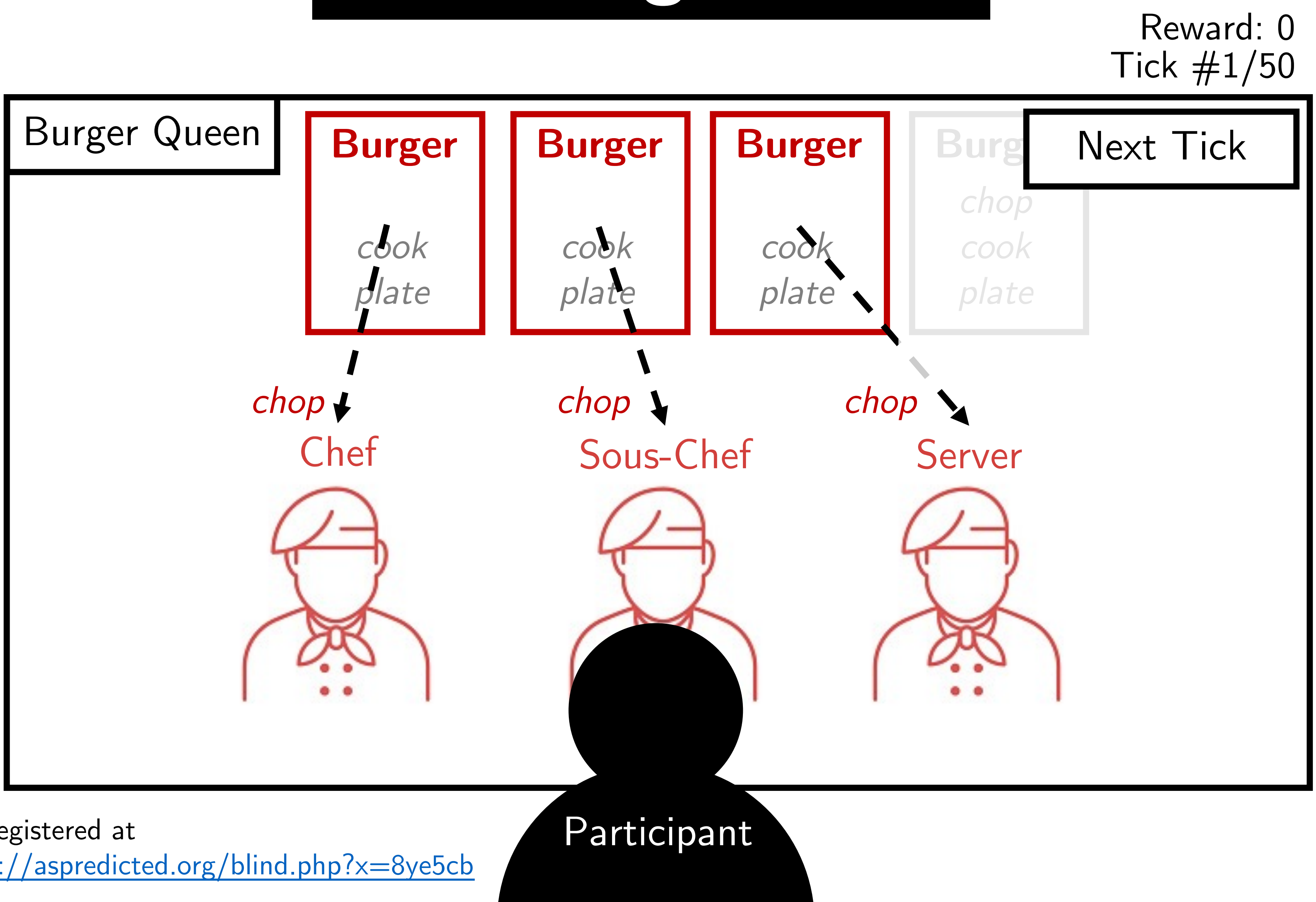
Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

Cooking Game



Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

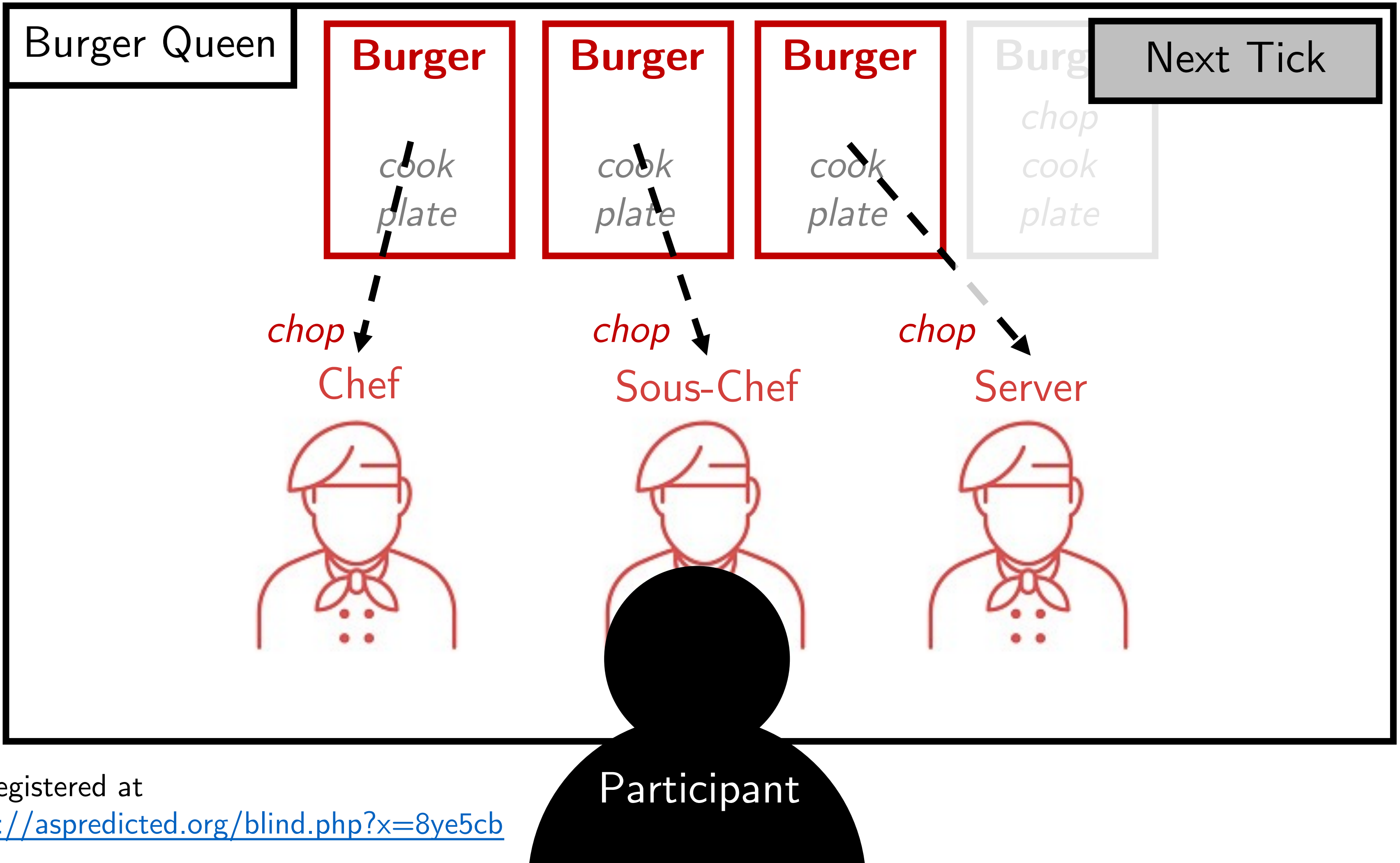
Cooking Game



Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

Cooking Game

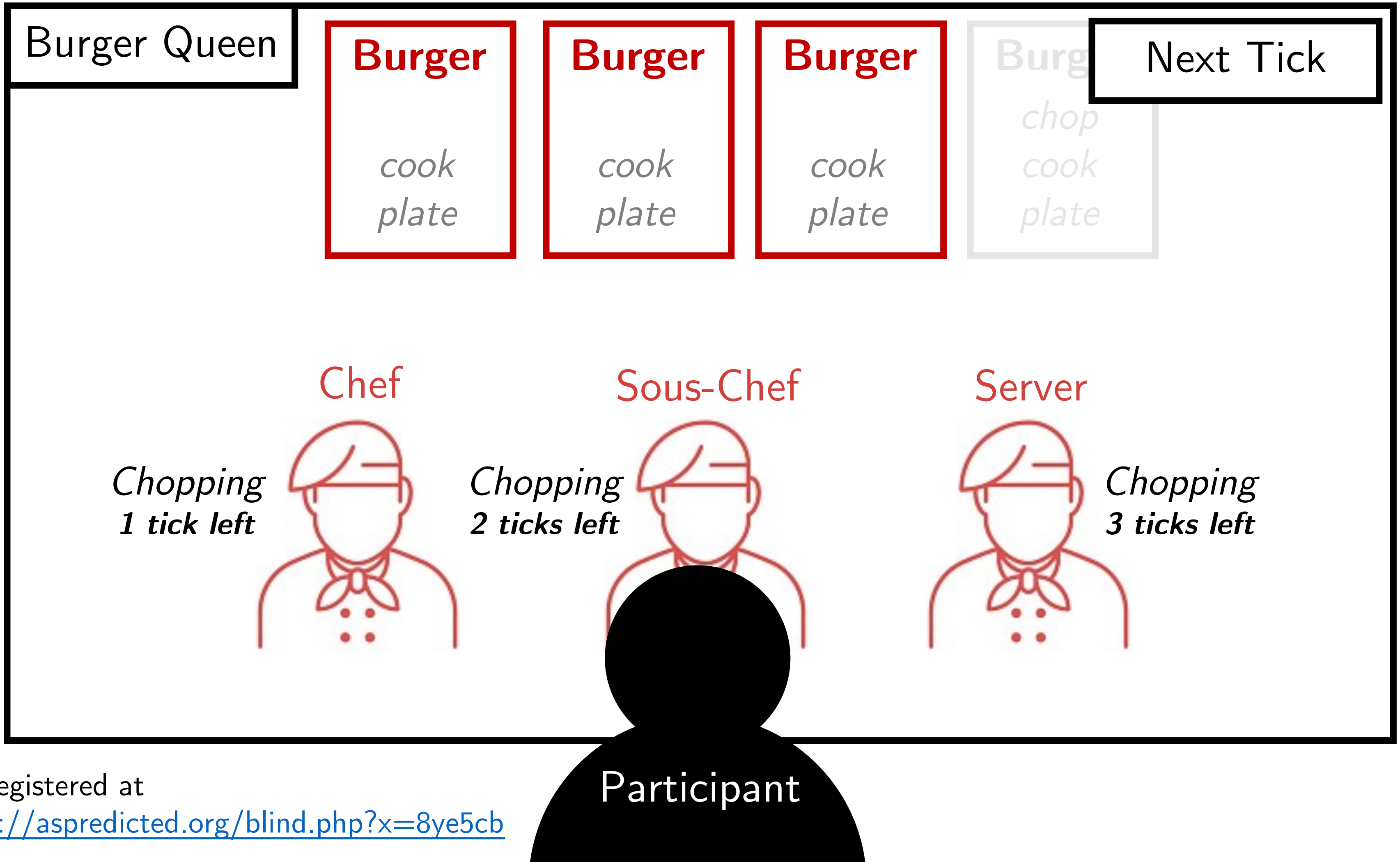
Reward: 0
Tick #1/50



Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

Cooking Game

Reward: 0
Tick #2/50



Pre-registered at
https://aspredicted.org/blind.php?x=8ye5cb

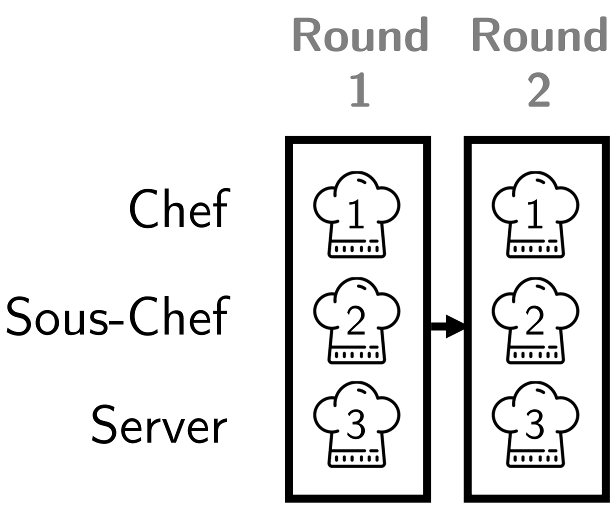
Design Disruption Scenario

🍔 x 4 within 50 ticks

Pre-registered at https://aspredicted.org/blind.php?x=8ye5cb

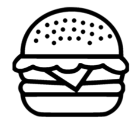
Design Disruption Scenario

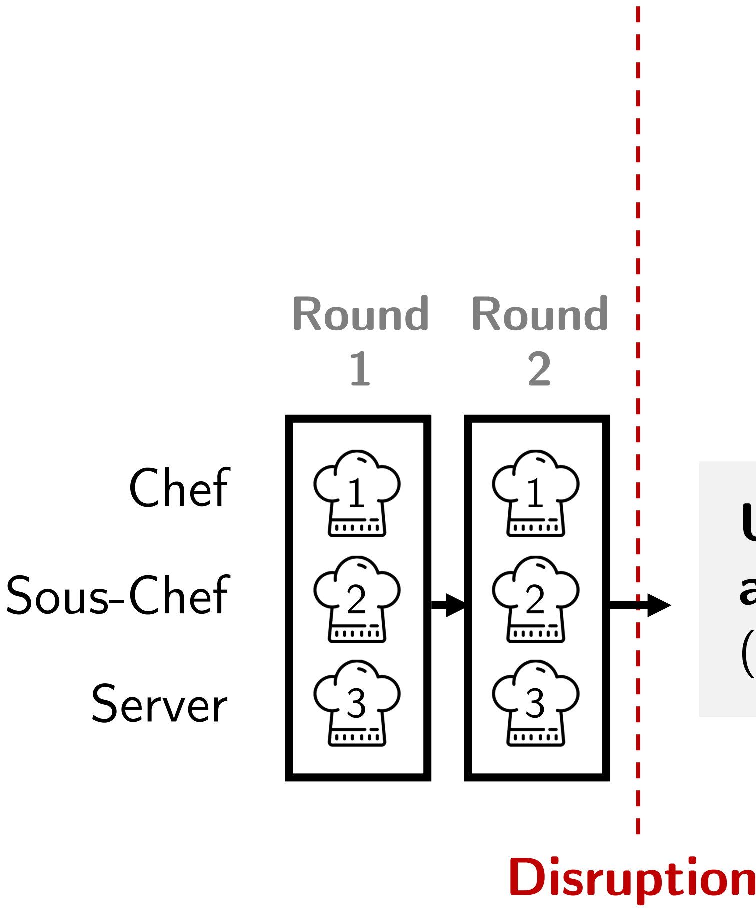
x 4 within 50 ticks



Pre-registered at https://aspredicted.org/blind.php?x=8ye5cb

Design Disruption Scenario

 x 4 within 50 ticks

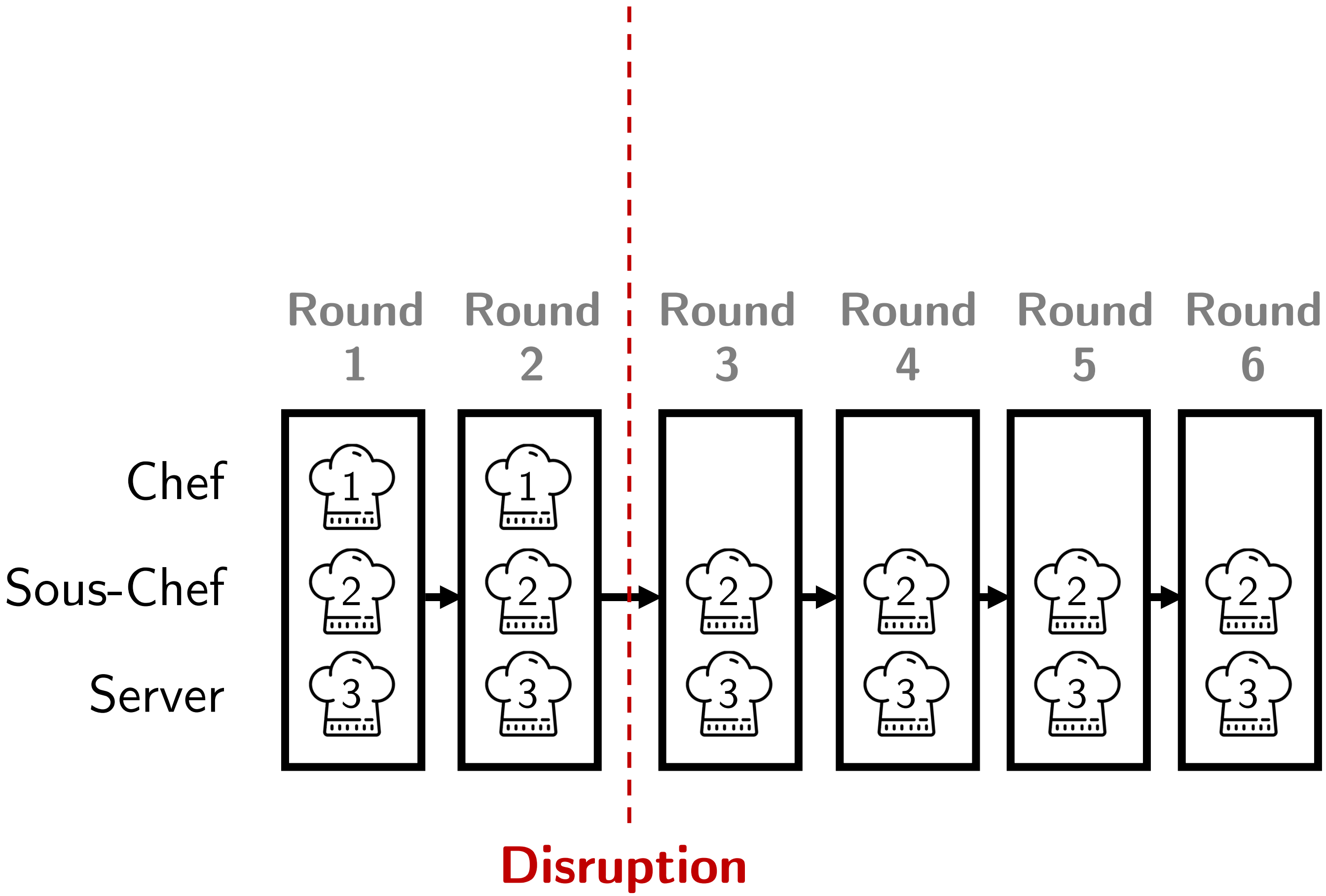


Unfortunately, the Chef is away training for Paris 2024.
(Good luck, Chef!)

Pre-registered at https://aspredicted.org/blind.php?x=8ye5cb

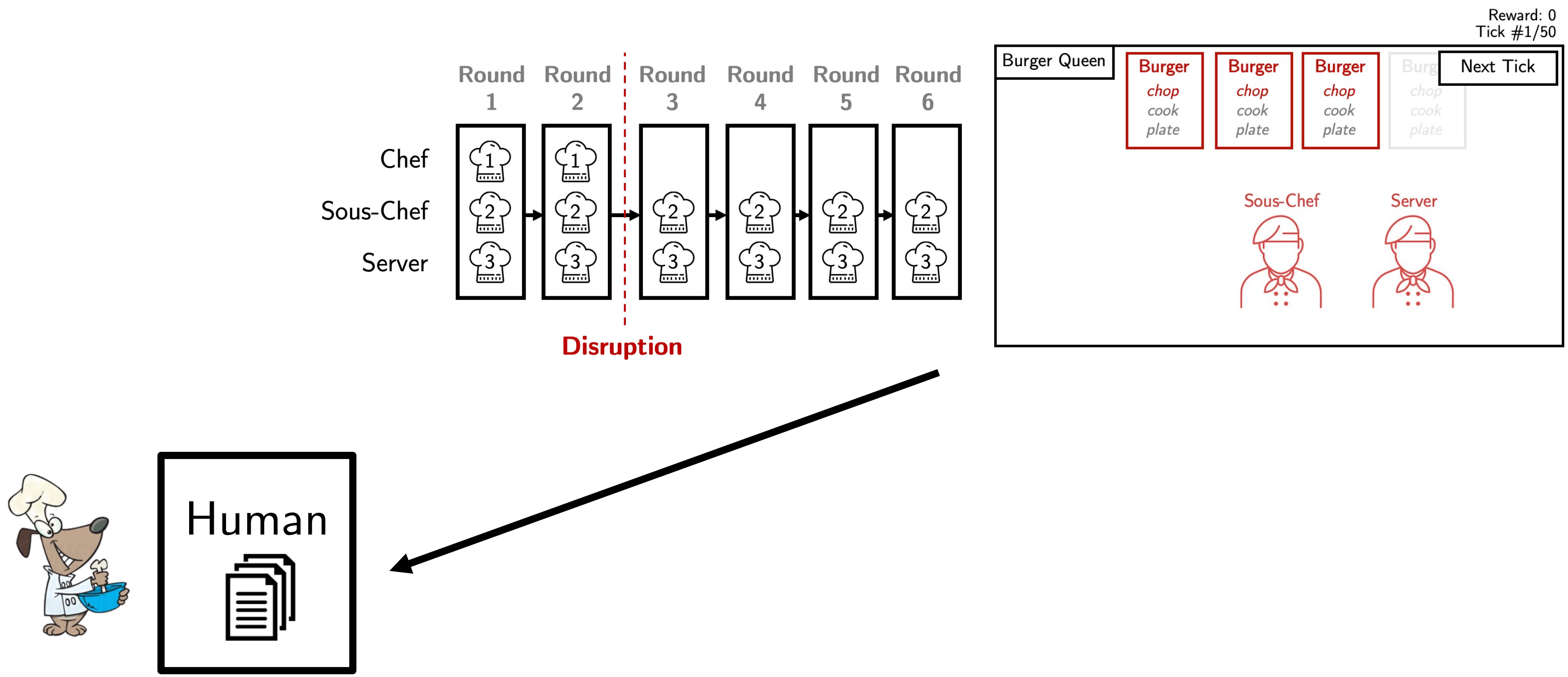
Design Disruption Scenario

x 4 within 50 ticks



Pre-registered at https://aspredicted.org/blind.php?x=8ye5cb

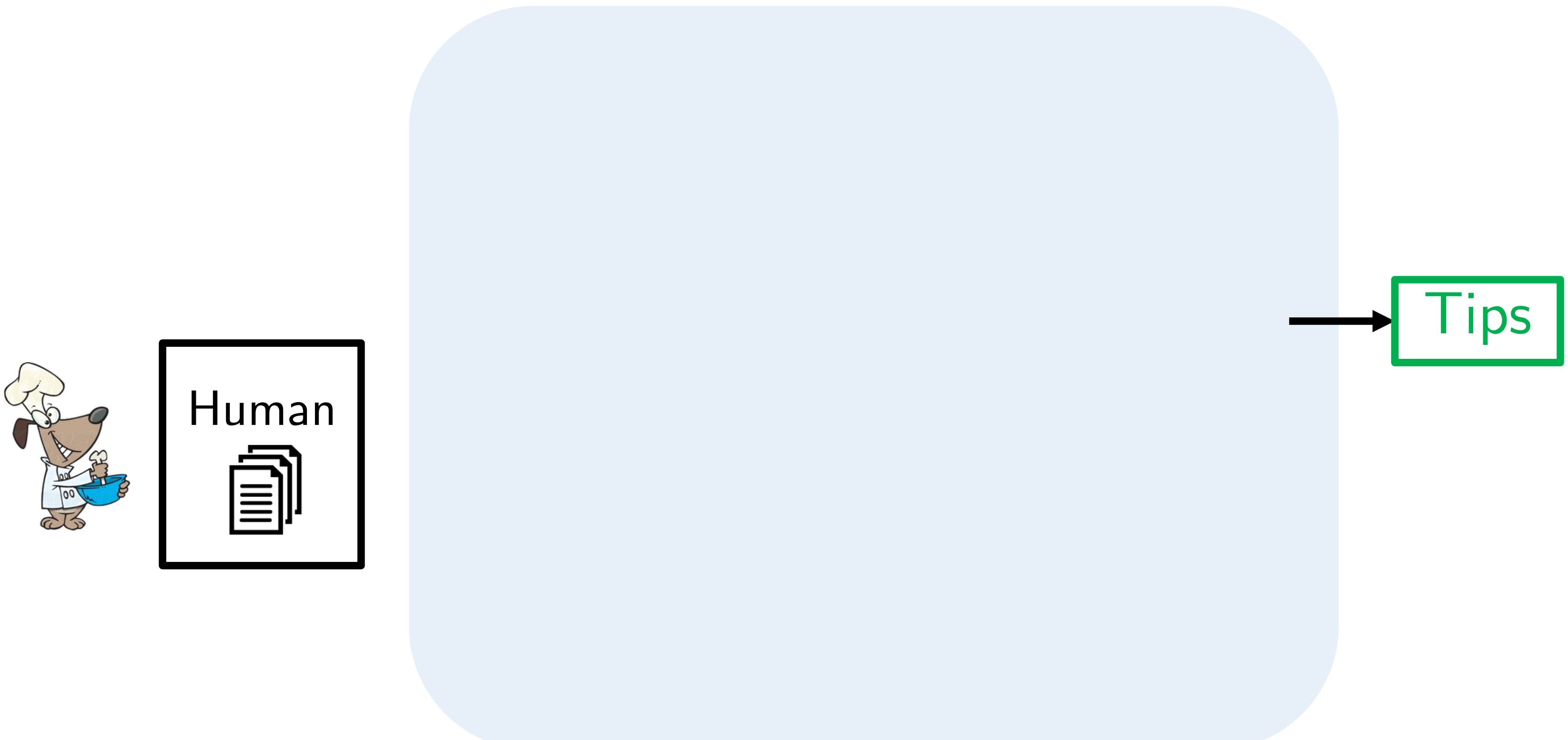
Phase I Collect Trace Data



Amazon Mechanical Turk, N = 172

mean age 36.4, 62% female

Our Approach



Our Approach

MDP: $\mathcal{M} = (S, A, R, P, \gamma)$

Human

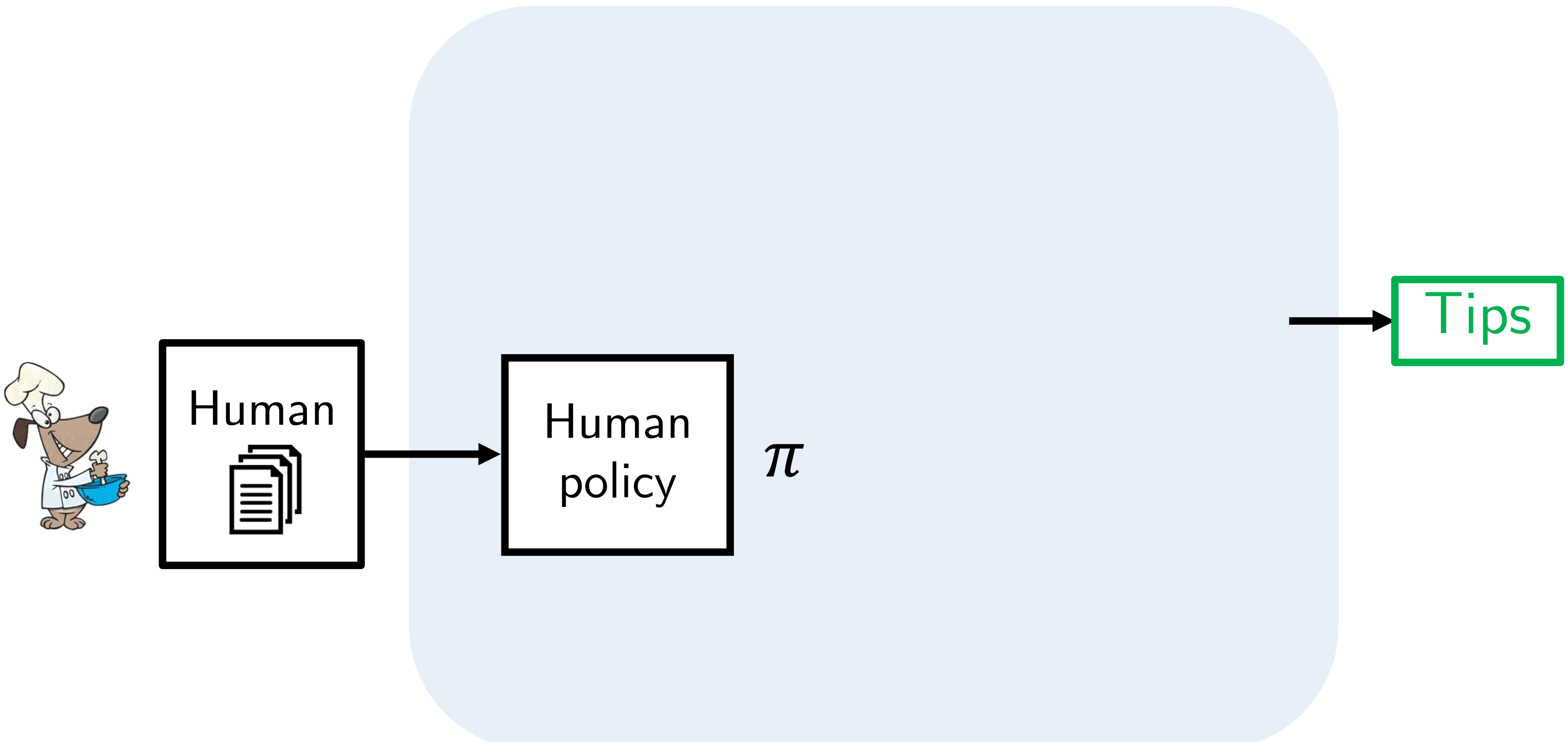
Input:
Trace data $\hat{d}_h$
from human

$\{(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots, (s_T, a_T, r_T)\}$

Tips

Our Approach

MDP: $\mathcal{M} = (S, A, R, P, \gamma)$

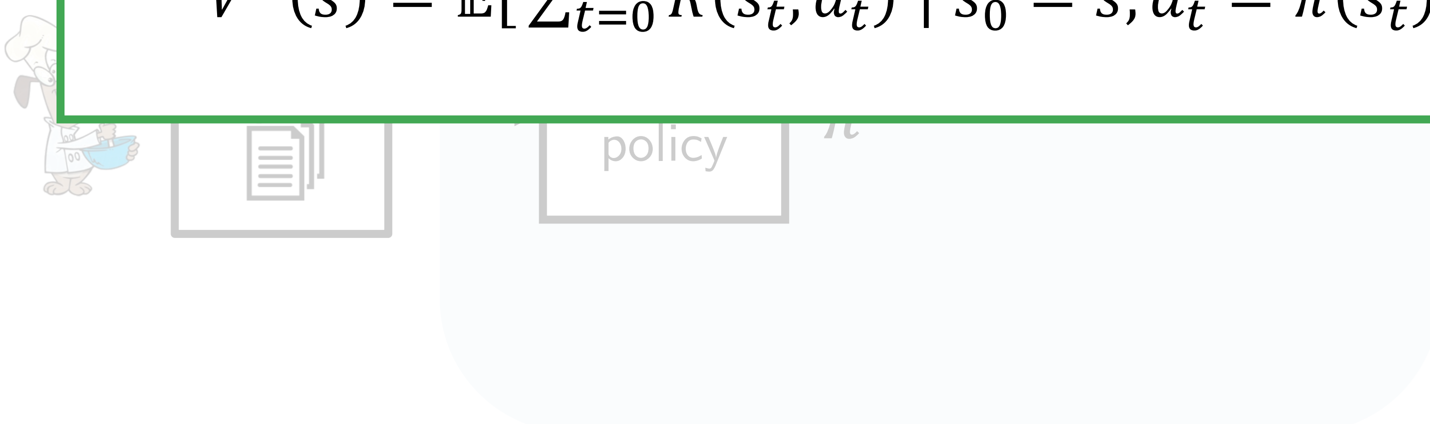


Our Approach

MDP: $\mathcal{M} = (S, A, R, P, \gamma)$

Value function $V^{\pi}(s)$ is the cumulative reward obtained by using policy π from state s

$$V^{\pi}(s) = \mathbb{E}[\, \sum_{t=0}^{T} R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t) \,]$$



Step 1: Q-Learning

MDP: $\mathcal{M} = (S, A, R, P, \gamma)$

Q function $Q^{\pi}(s, a)$ is the reward obtained by taking action a in state s and using policy π thereafter

$$Q^{\pi}(s, a) = \mathbb{E}_{s' \sim p(s'|s,a)}[V^{\pi}(s')]$$

- Watkins & Dayan 1992

Step 1: Q-Learning

MDP: $\mathcal{M} = (S, A, R, P, \gamma)$

Q function $Q^{\pi}(s,a)$ is the reward obtained by taking action a in state s and using policy π thereafter

$$Q^{\pi}(s,a) = \mathbb{E}_{s' \sim p(s'|s,a)}[V^{\pi}(s')]$$

- Watkins & Dayan 1992

- Learn using supervised learning on trace data obtained using π

$$\hat{Q}^{\pi}_{\theta}(s,a) \approx Q^{\pi}(s,a)$$

Our Approach

MDP: $\mathcal{M} = (S, A, R, P, \gamma)$

Optimal policy π^*

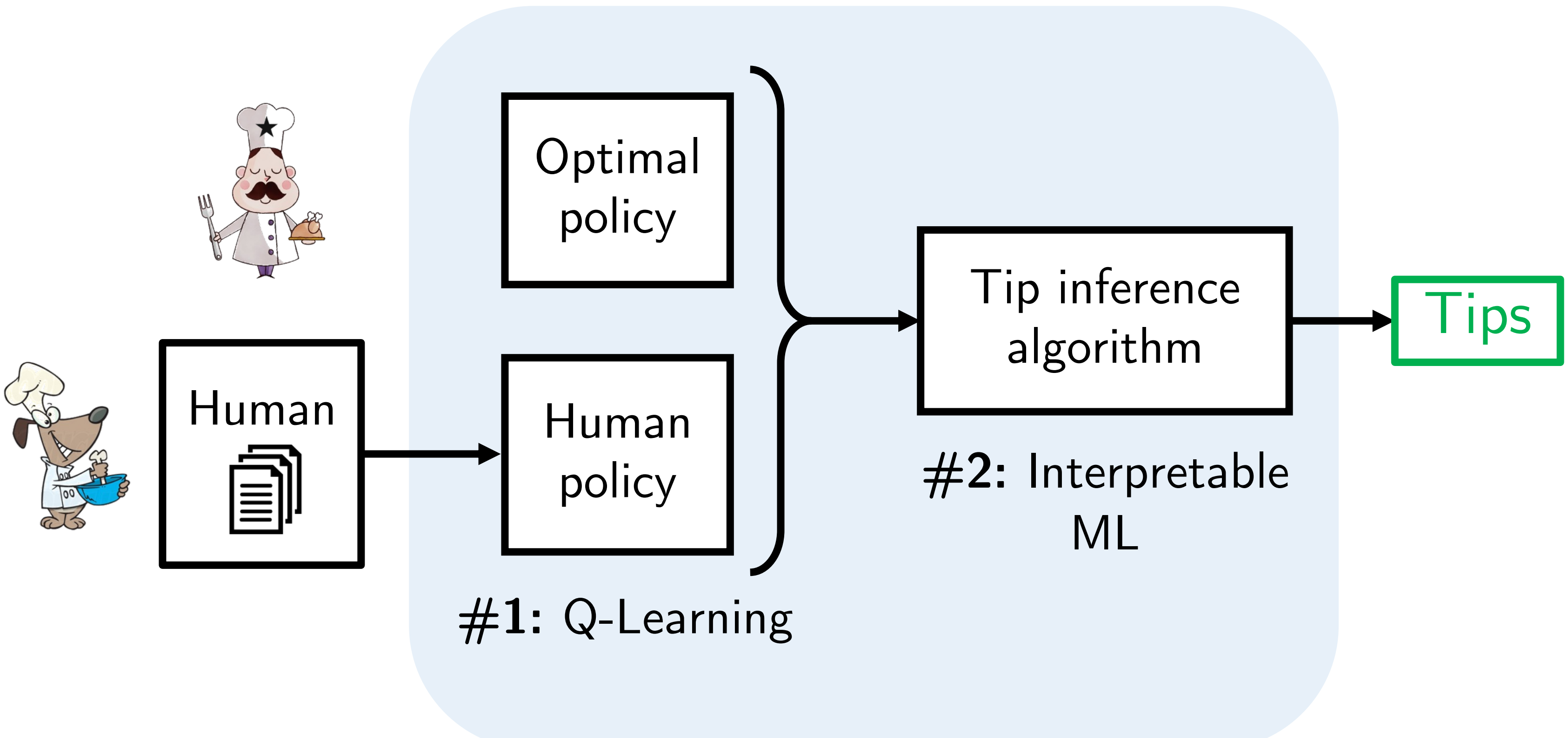
Human

Human policy π

#1: Q-Learning

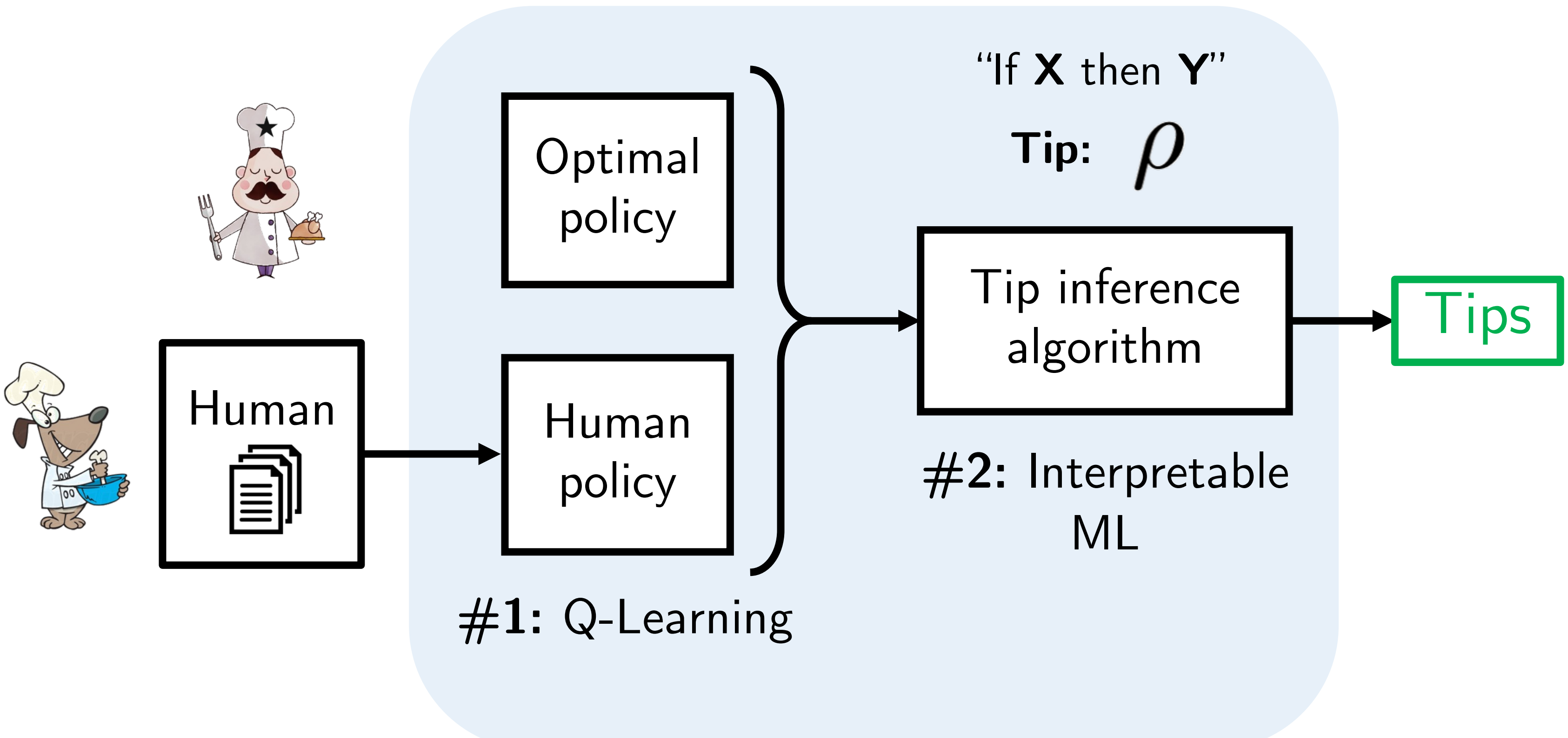
Tips

Our Approach



Caruana et al. 2015, Letham et al. 2015

Our Approach



Caruana et al. 2015, Letham et al. 2015

Step 2: Tip Inference

Cumulative reward for a given policy

$$J(\pi) = \mathbb{E}_{\zeta \sim D(\pi)} \left[\sum_{t=1}^{T} r_t \right]$$

Step 2: Tip Inference

Cumulative reward for a given policy

$$J(\pi) = \mathbb{E}_{\zeta \sim D(\pi)} \left[\sum_{t=1}^{T} r_t \right]$$

- **Algorithm:** Choose tip ρ that maximizes the objective

$$\underbrace{J(\pi_H \oplus \rho)}_{\text{Human policy + tip}} - \underbrace{J(\pi_H)}_{\text{Only human policy}}$$

  - $\pi_\mathrm{h} \oplus \rho$ denotes overriding the human policy with tip ρ.

Step 2: Tip Inference

Cumulative reward for a given policy

$$J(\pi) = \mathbb{E}_{\zeta \sim D(\pi)}\left[\sum_{t=1}^{T} r_t\right]$$

- **Algorithm:** Choose tip ρ that maximizes the objective

$$J(\pi_H \oplus \rho) - J(\pi_H)$$

Human policy + tip **Only human policy**

  - $\pi_{\mathrm{h}} \oplus \rho$ denotes overriding the human policy with tip ρ.

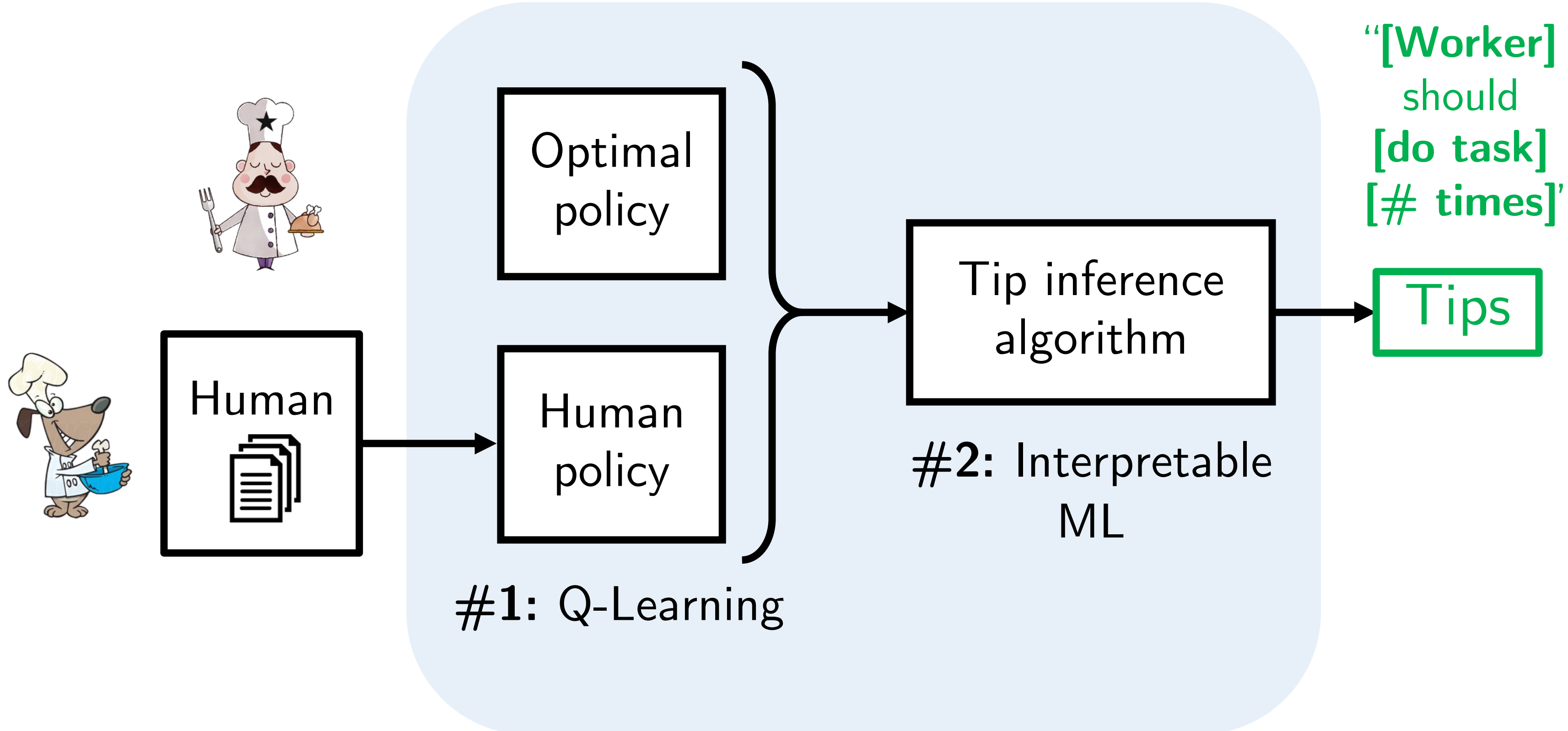
- **Lemma:** $J(\pi_H \oplus \rho) - J(\pi_H) \approx$

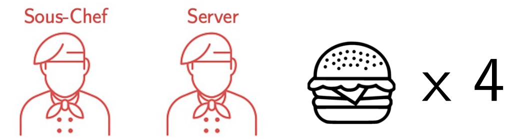
$$\mathbb{E}_{\zeta \sim D(\pi_H)}\left[\sum_{t=1}^{T} Q_t^*(s_t, \pi_H \oplus \rho(s_t)) - Q_t^*(s_t, \pi_H(s_t))\right]$$

Indirect effect of distribution shift is small; use observed data

Q-network we learned previously!

Our Approach



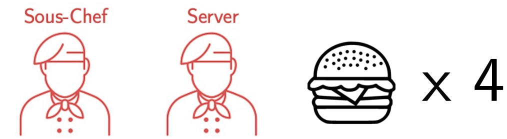


Phase I Inferred Tips

Algorithm

Server
should cook twice

Amazon Mechanical Turk, N = 172
mean age 36.4, 62% female



Phase I Inferred Tips

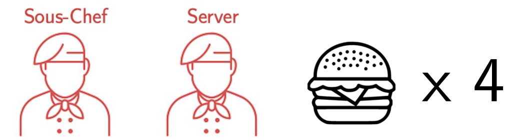
Algorithm

Server
should cook twice

Human

Most frequent tip
chosen by participants

Amazon Mechanical Turk, N = 172
mean age 36.4, 62% female

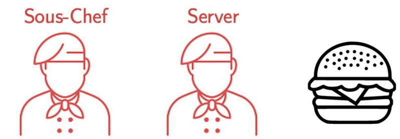


Phase I Inferred Tips

Algorithm	Human
Server should cook twice	Server should cook once

Most frequent tip chosen by participants

Amazon Mechanical Turk, N = 172
mean age 36.4, 62% female



x 4

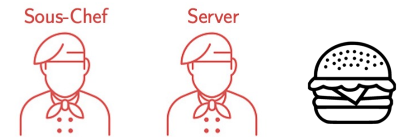
Phase I Inferred Tips

Algorithm	Human	Baseline
Server should cook twice	Server should cook once	
	Most frequent tip chosen by participants	
		Most frequent s-a deviation b/w optimal and trainee policies

Amazon Mechanical Turk, N = 172

mean age 36.4, 62% female

Phase I Inferred Tips



x 4

Algorithm	Human	Baseline
Server should cook twice	Server should cook once	Sous-Chef should plate twice
	Most frequent tip chosen by participants	*Most frequent s-a deviation b/w optimal and trainee policies*

Amazon Mechanical Turk, N = 172
mean age 36.4, 62% female

Phase II Comparing Tips

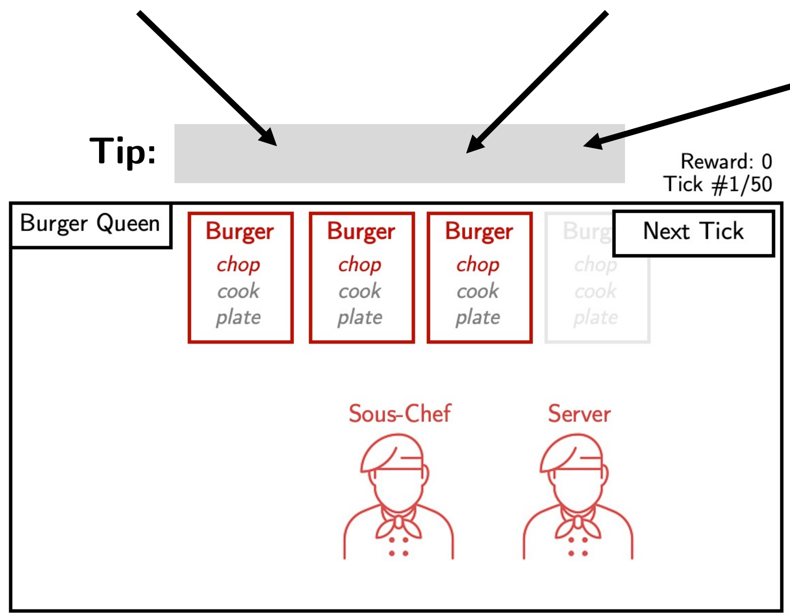
Control	Algorithm	Human	Baseline
- No tip -	Server should cook twice	Server should cook once	Sous-Chef should plate twice

Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Phase II Comparing Tips

Control	Algorithm	Human	Baseline
- No tip -	Server should cook twice	Server should cook once	Sous-Chef should plate twice



Amazon Mechanical Turk, N = 1,011

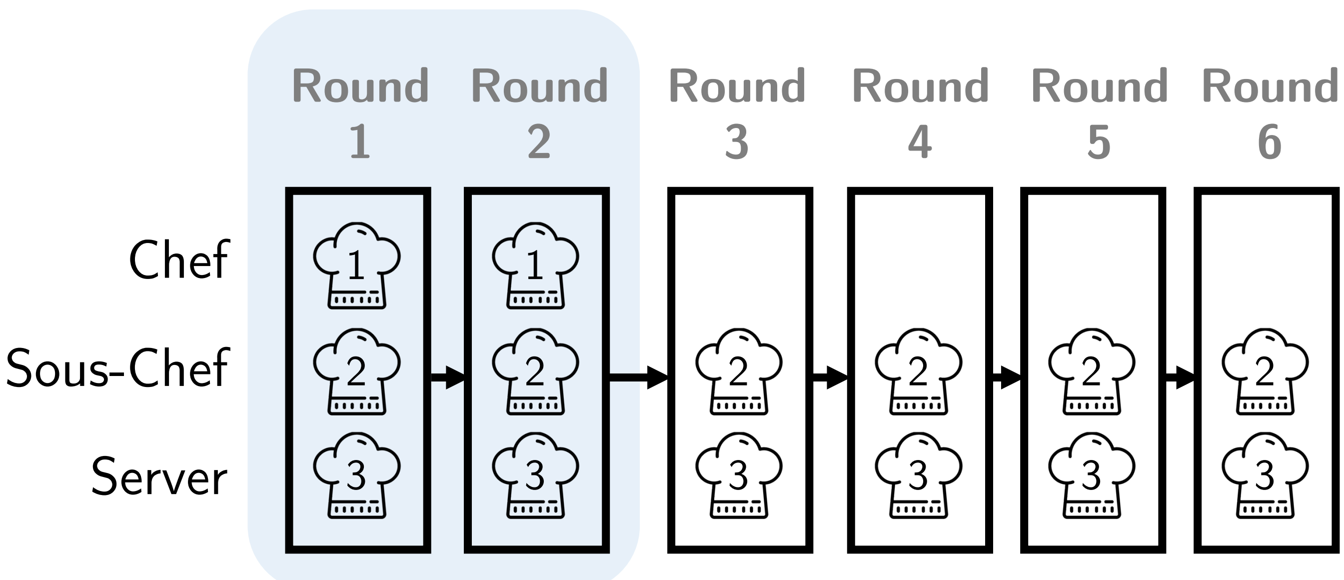
mean age 34.9, 60% female

Algorithm vs Human

Algorithm	Human
Server should cook twice	Server should cook once

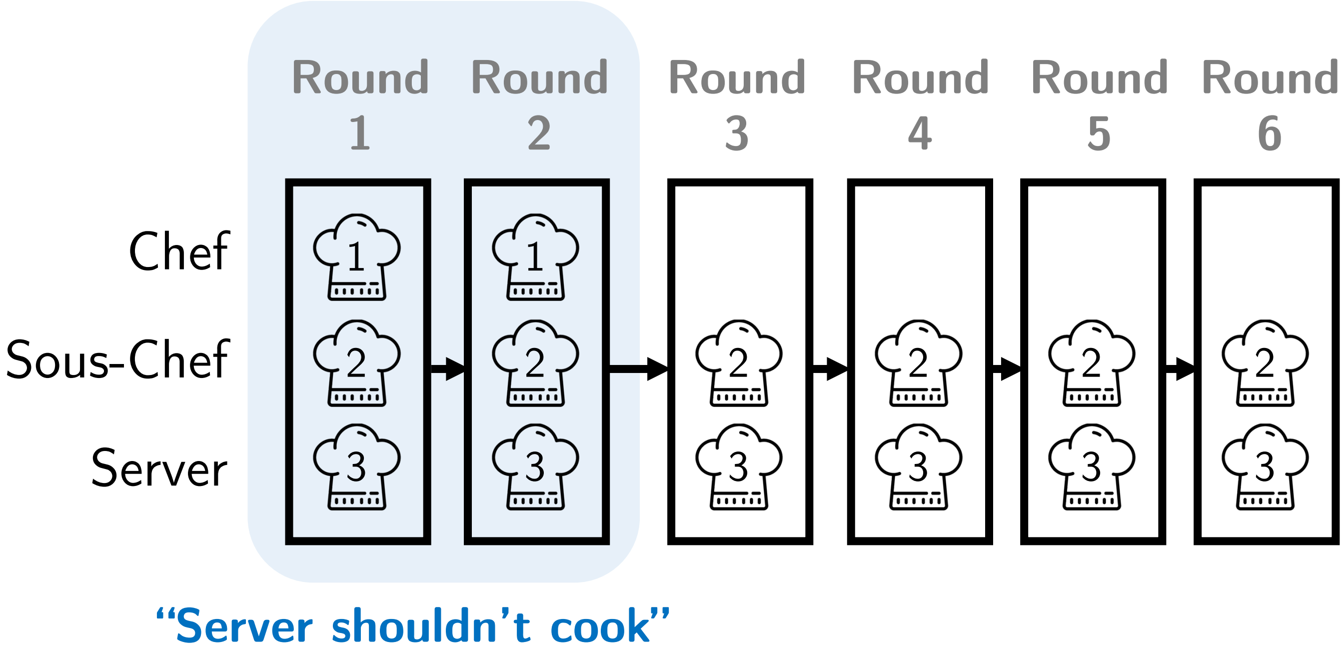
Algorithm vs Human

Algorithm	Human
Server should cook twice	Server should cook once



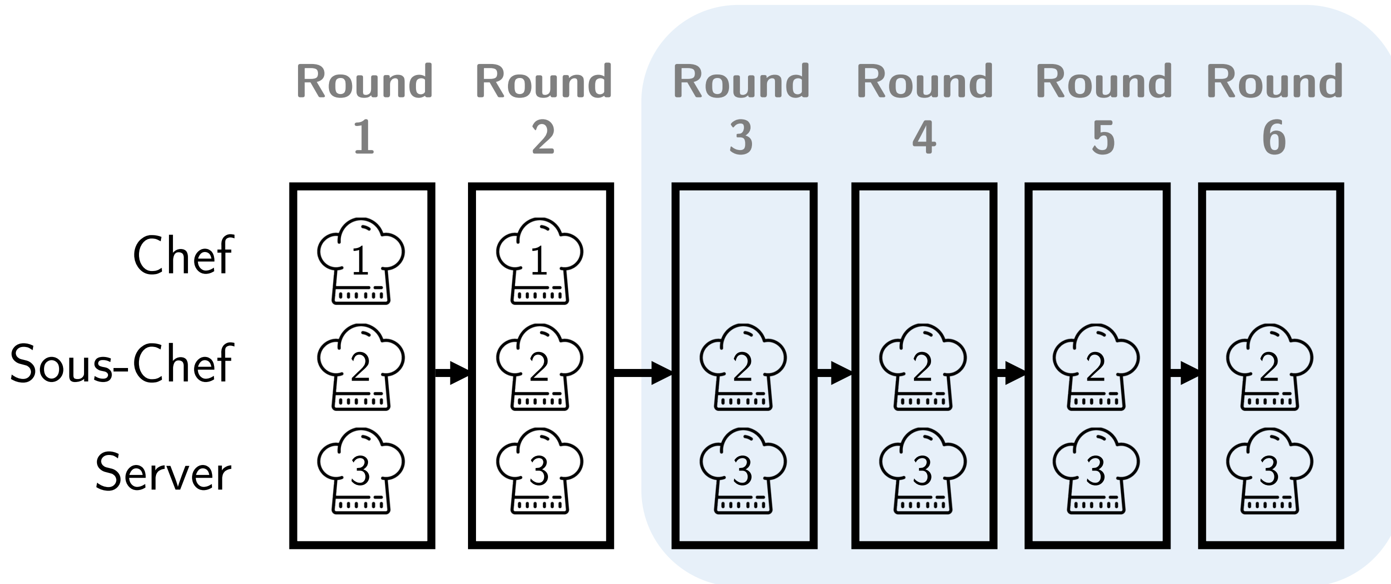
Algorithm vs Human

Algorithm	Human
Server should cook twice	Server should cook once



Algorithm vs Human

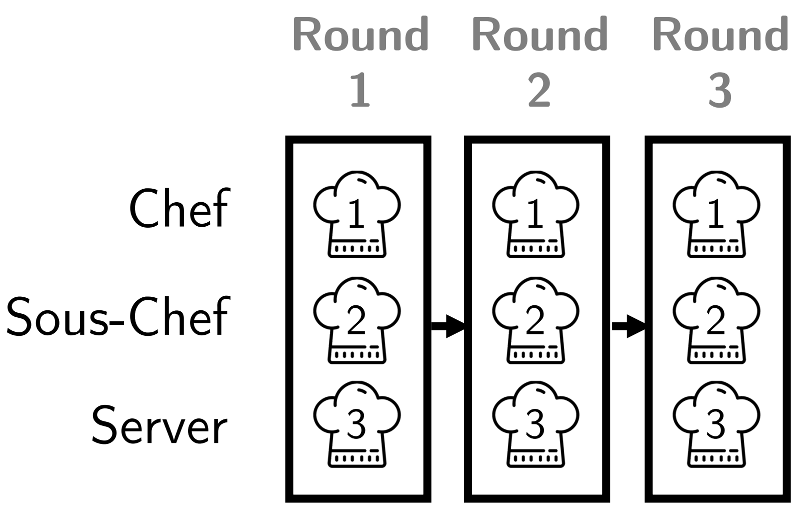
Algorithm	Human
Server should cook twice	Server should cook once



"Server shouldn't cook"

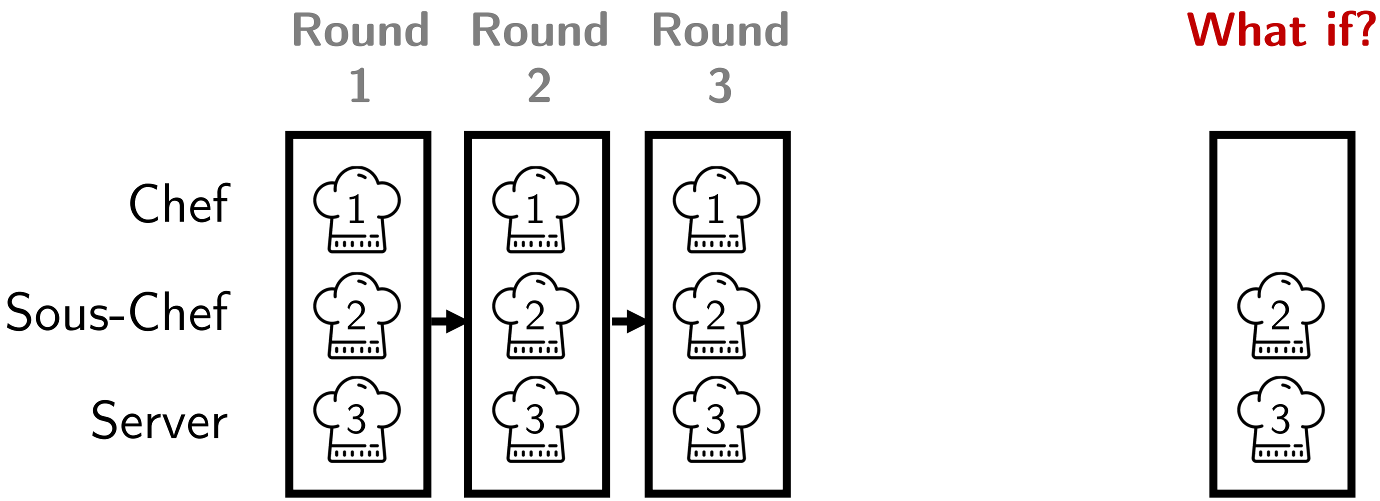
Algorithm vs Human

Algorithm	Human
Server should cook twice	Server should cook once



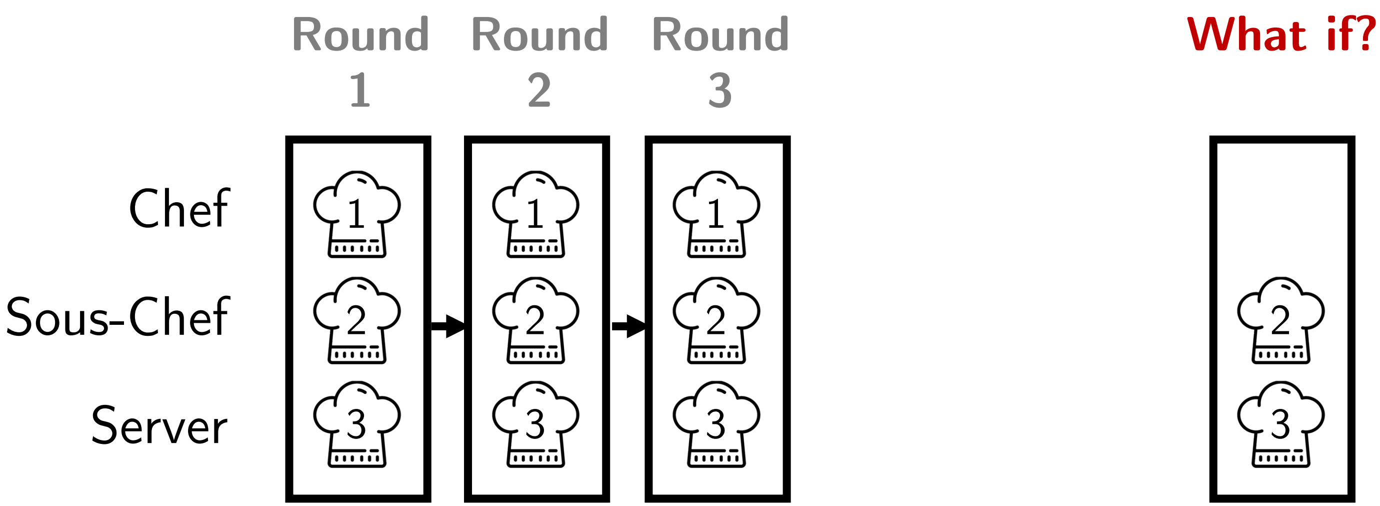
Algorithm vs Human

Algorithm	Human
Server should cook twice	Server should cook once



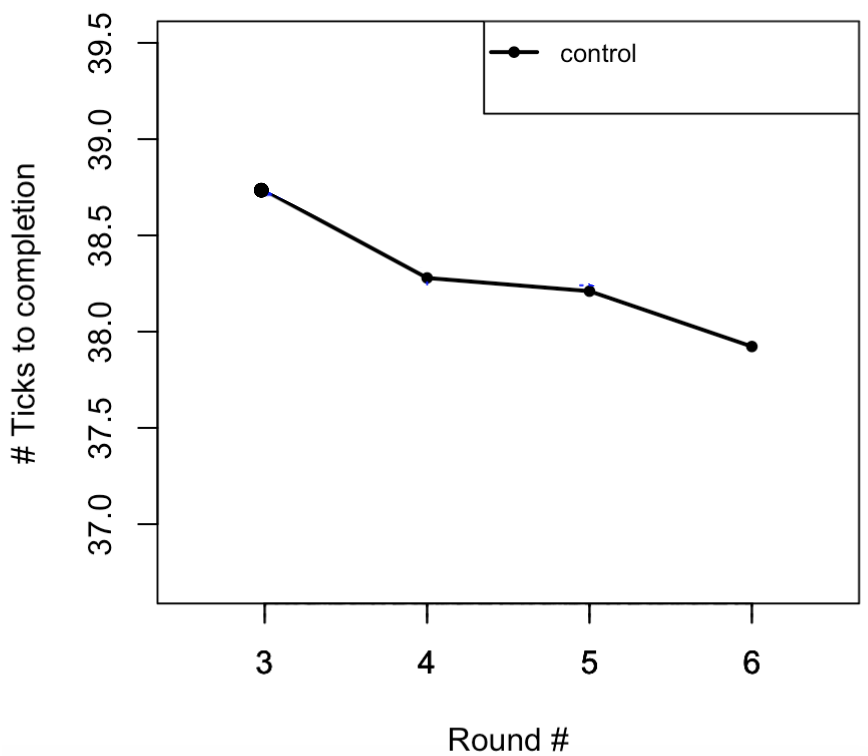
Algorithm vs Human

Algorithm	Human	Hypothetical
Server should cook twice	Server should cook once	Server shouldn't cook



Results People Improve Over Time

Ticks to completion

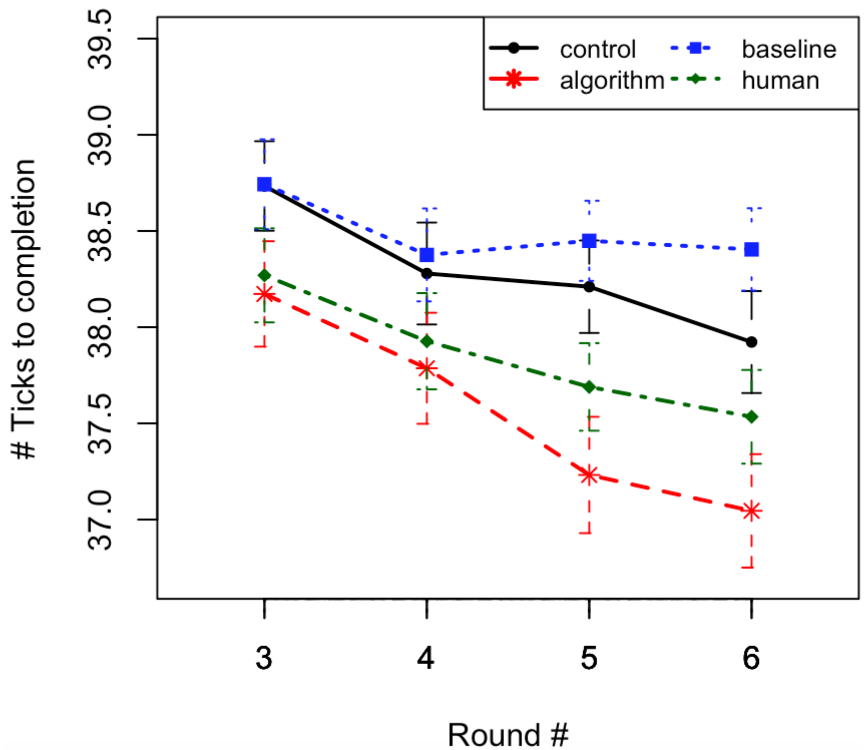


Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Results Our Tip Improves Performance

Ticks to completion



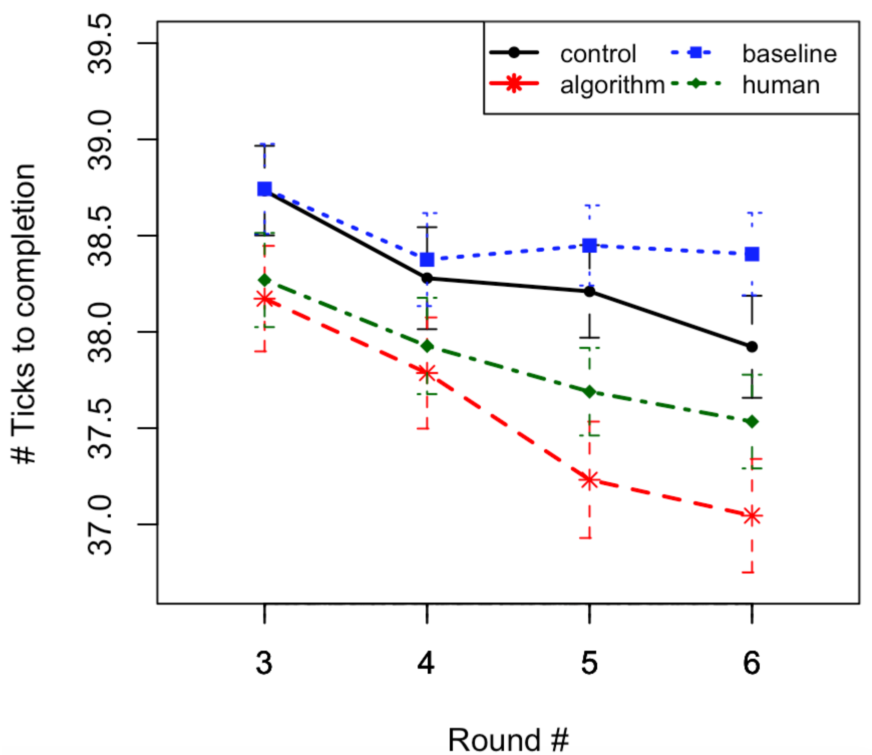
One-sided T-Tests

Algorithm *beats* Control ($p = 0.000008$)
Algorithm *beats* Human ($p = 0.006$)
Algorithm *beats* Baseline ($p < 1e\text{-}12$)

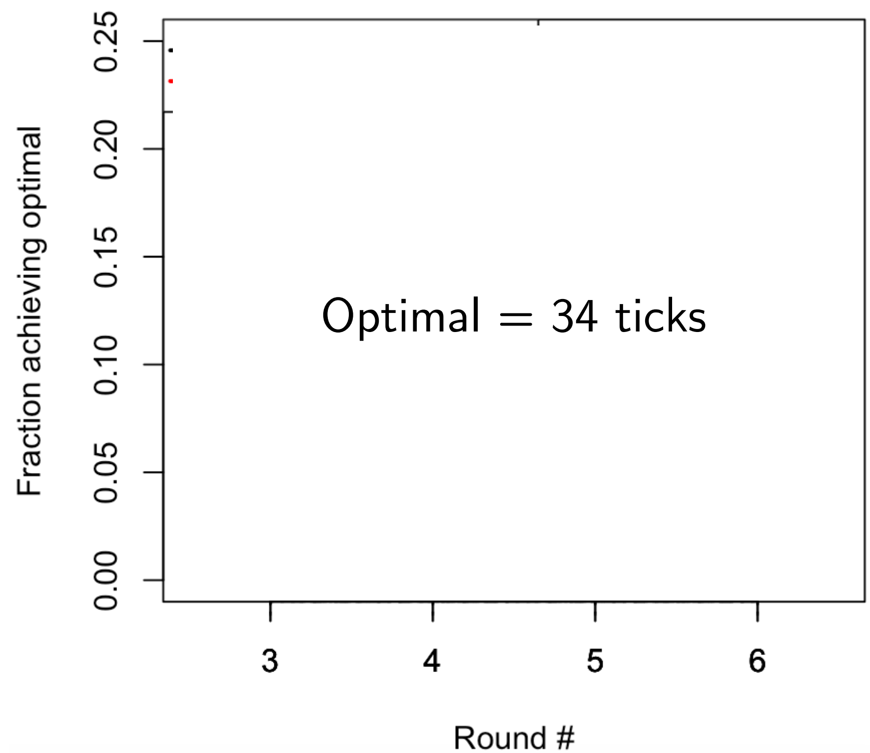
Amazon Mechanical Turk, N = 1,011
mean age 34.9, 60% female

Results

Ticks to completion



Fraction achieving optimal

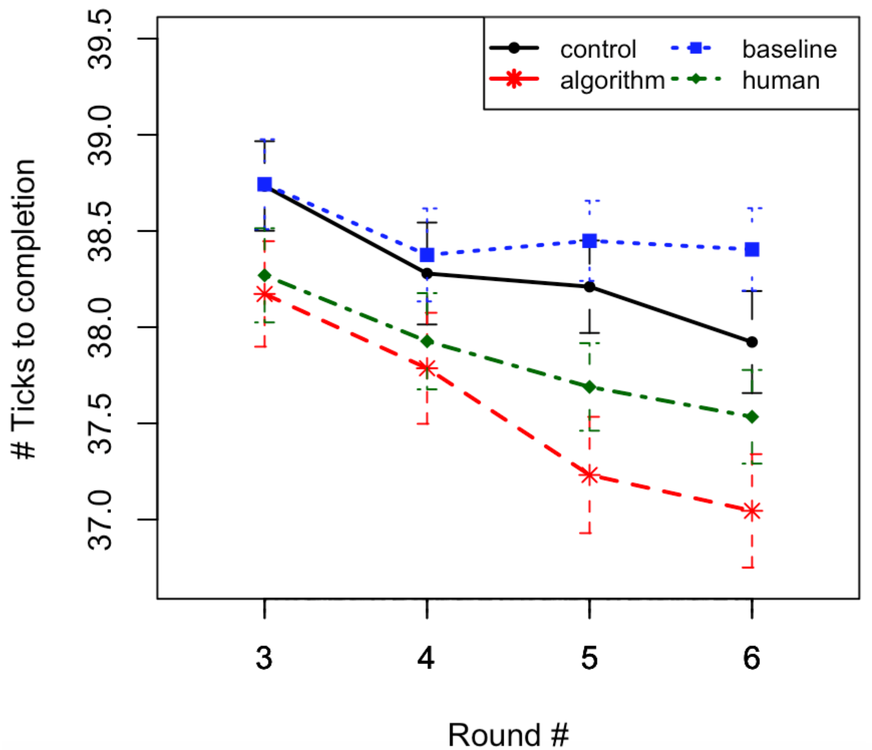


Amazon Mechanical Turk, N = 1,011

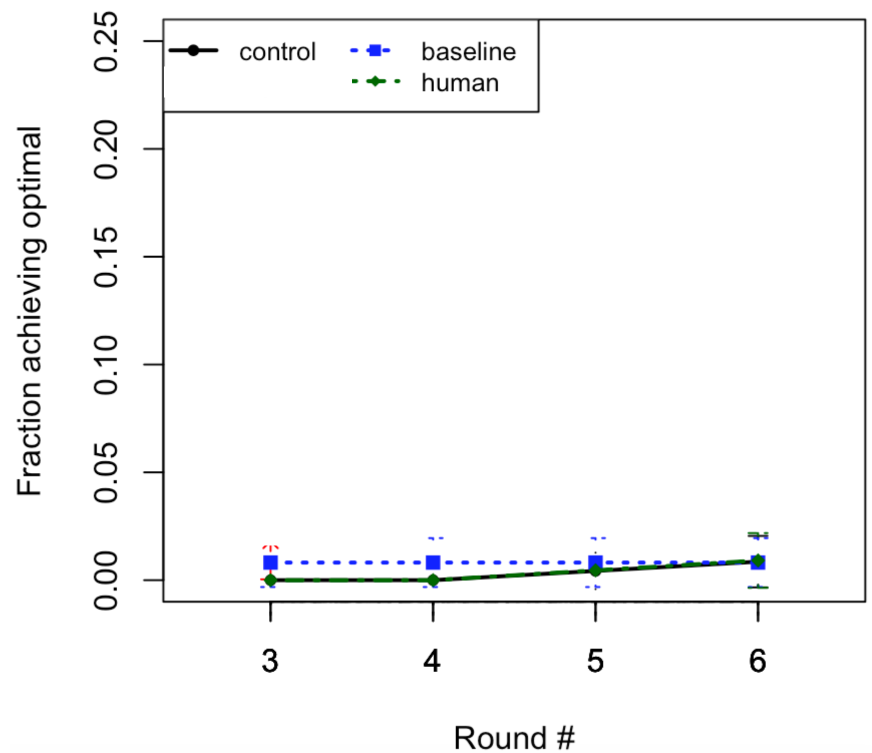
mean age 34.9, 60% female

Results Difficult to Reach Optimal

Ticks to completion



Fraction achieving optimal

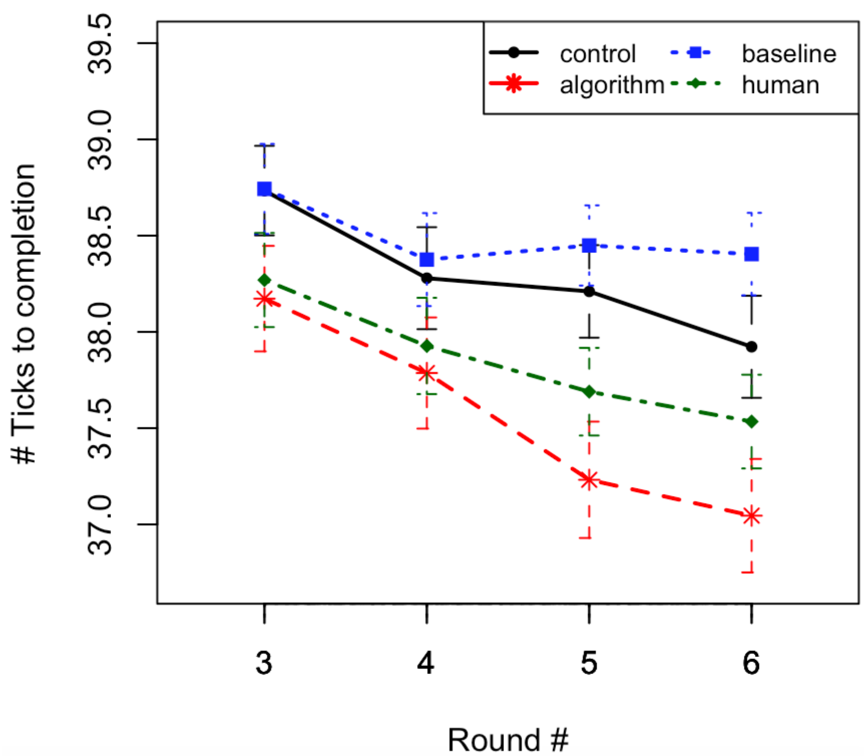


Amazon Mechanical Turk, N = 1,011

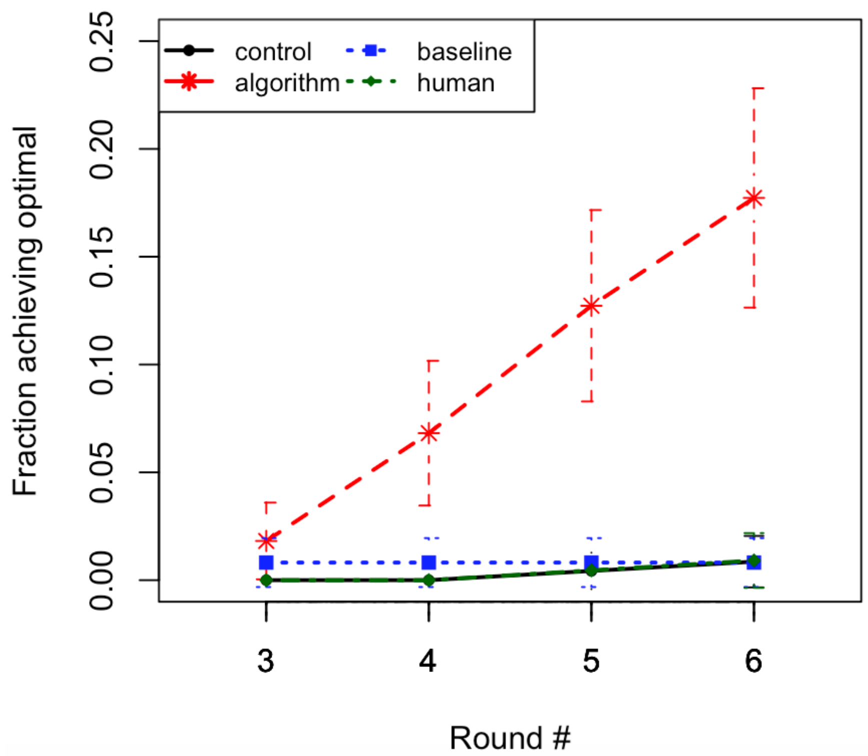
mean age 34.9, 60% female

Results Our Tip Helps Reach Optimal

Ticks to completion



Fraction achieving optimal

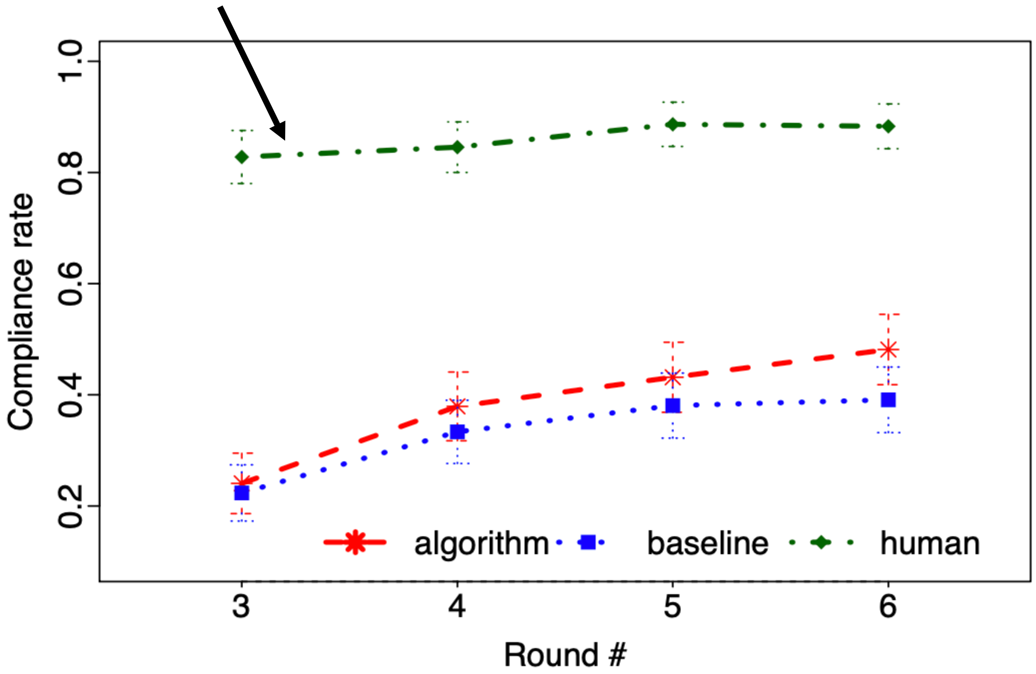


Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Results Complying with Intuitive Tip

Human "Server cooks once"



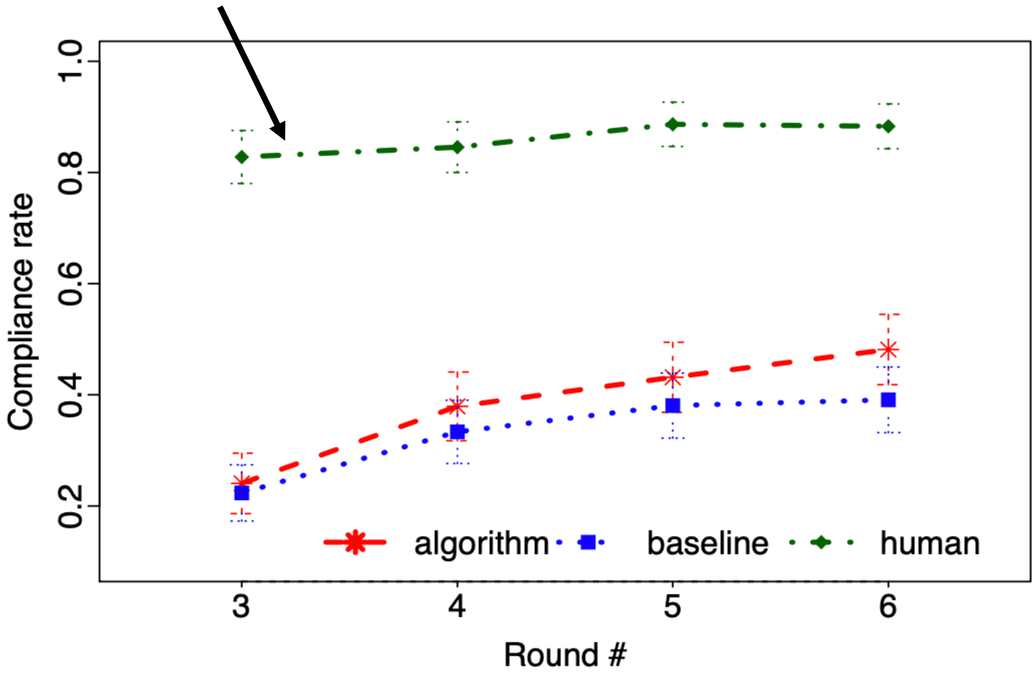
Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Results Complying with Intuitive Tip

26% Positive, 17% Negative

Human “Server cooks once”



“I felt that tip was **valid**.”

R_1rvkYTwgAjD0z4z

“It helped because she could cook one burger but **any more than that and your ticks would be too high**.”

R_d6YSuigdikyaNdT

“It was **accurate**, and I implemented it.”

R_1pA8wDYgWc9hbIt

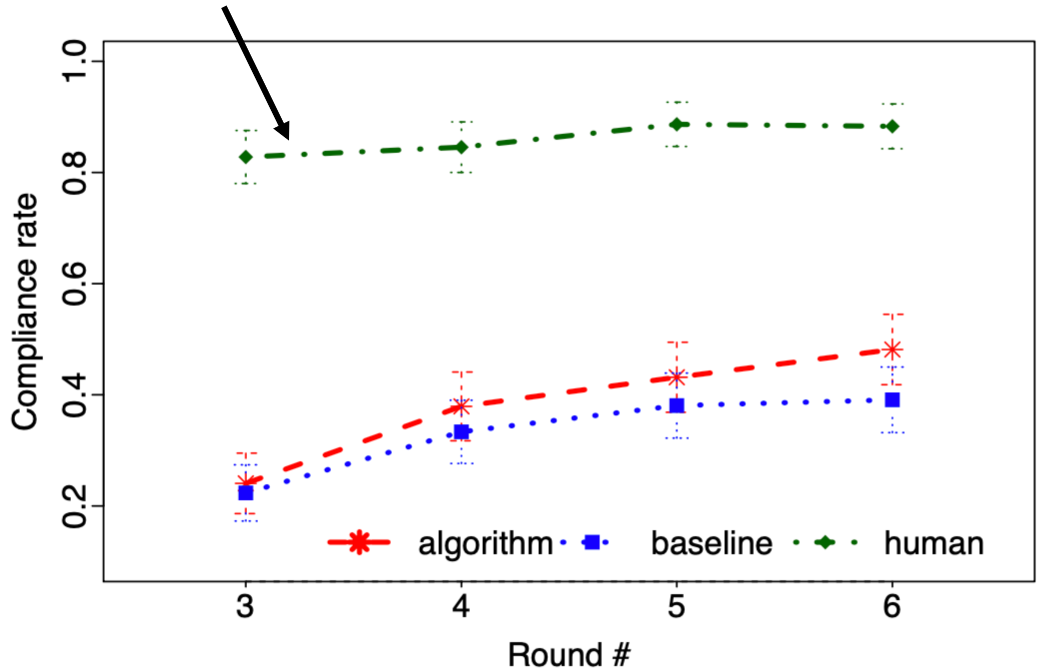
Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Results Complying with Intuitive Tip

26% Positive, 17% Negative

Human "Server cooks once"



Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

"I felt that tip was **valid**."

R_1rvkYTwgAjD0z4z

"It helped because she could cook one burger but **any more than that and your ticks would be too high**."

R_d6YSuigdikyaNdT

"It was **accurate**, and I implemented it."

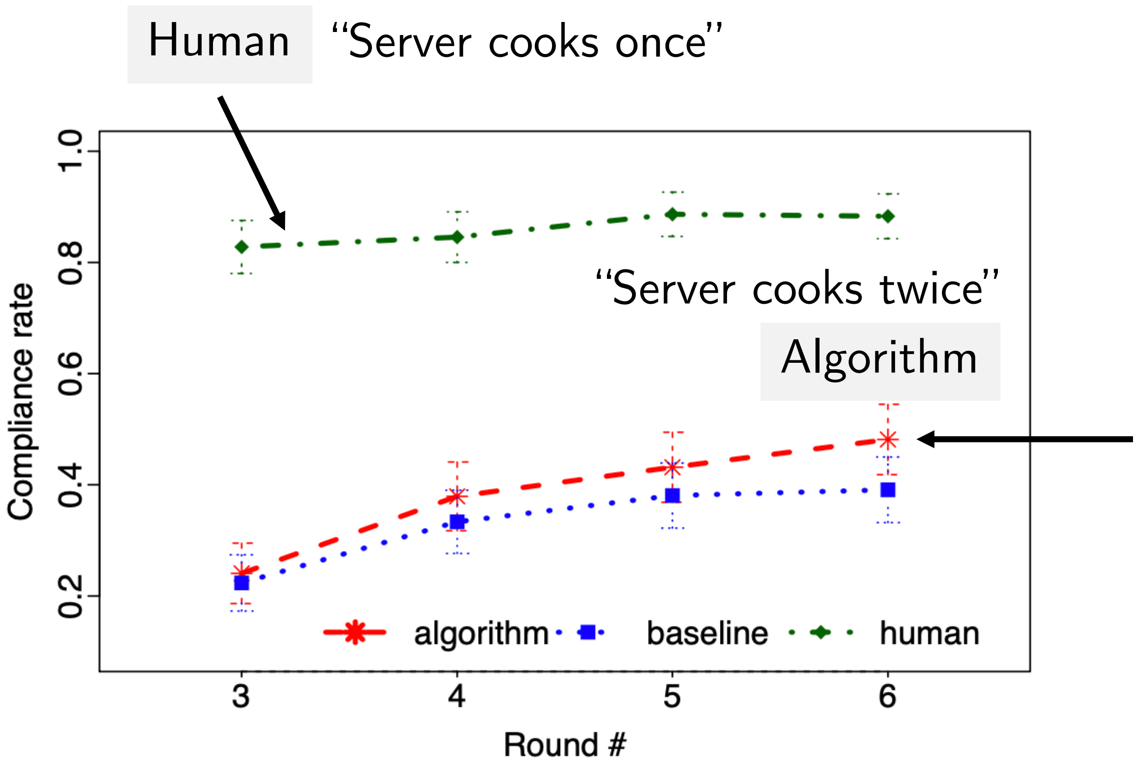
R_1pA8wDYgWc9hblt

"It stunk honestly. **The server takes forever to cook."**

R_beijQ8guDyExa5r

"I used the tip but **I don't think it was helpful.** The server took long to cook."

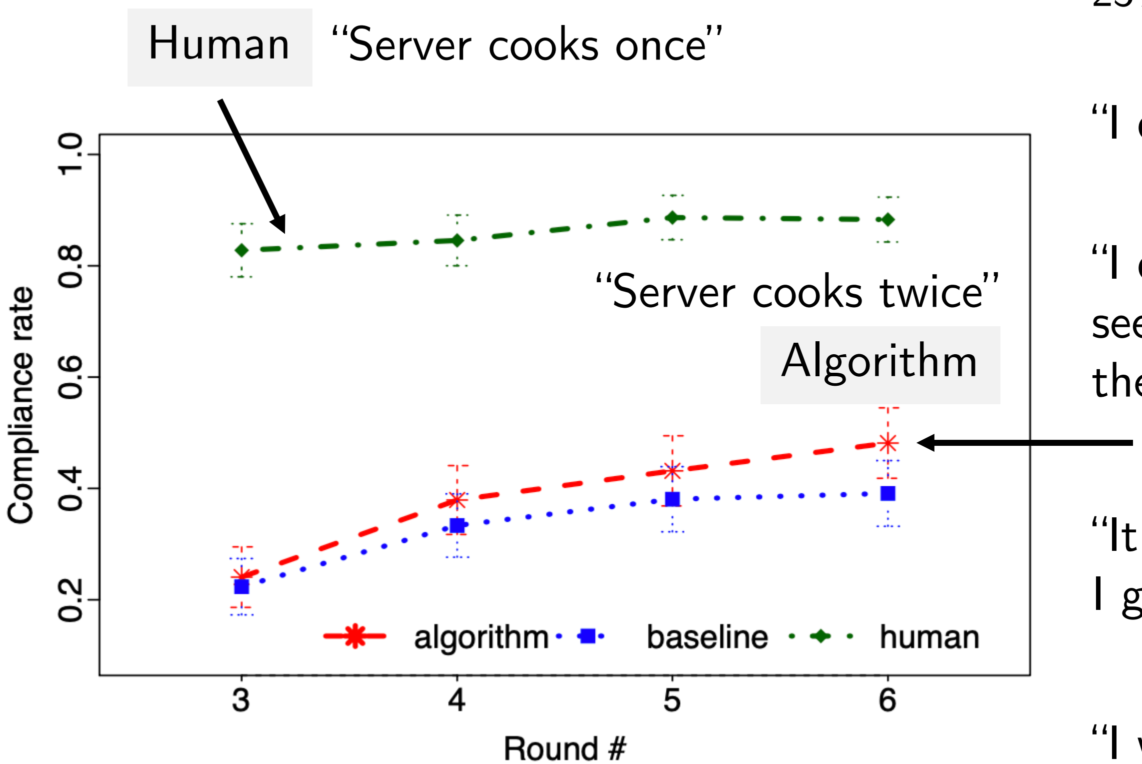
Results Against Counterintuitive Tips



Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Results Against Counterintuitive Tips



23% Positive, **33% Negative**

"I didn't think it was right."
R_3EgrcrQouPcb1fS

"I didn't follow it because it seemed counter intuitive since they're slow."
R_10HkPUkR6o0qDFT

"It didn't make sense and in fact I got worse trying to use it,"
R_2YD5x6BL7mhCYEP

"I wasn't sure how to use it."
R_2s0UA1omAifrFgx

Amazon Mechanical Turk, N = 1,011
mean age 34.9, 60% female

Results Learning Beyond Tips

Structure of Optimal Policy

		Chop	Cook	Plate	
Sous-Chef	2	3	2	2	times
Server	3	1	2	2	times

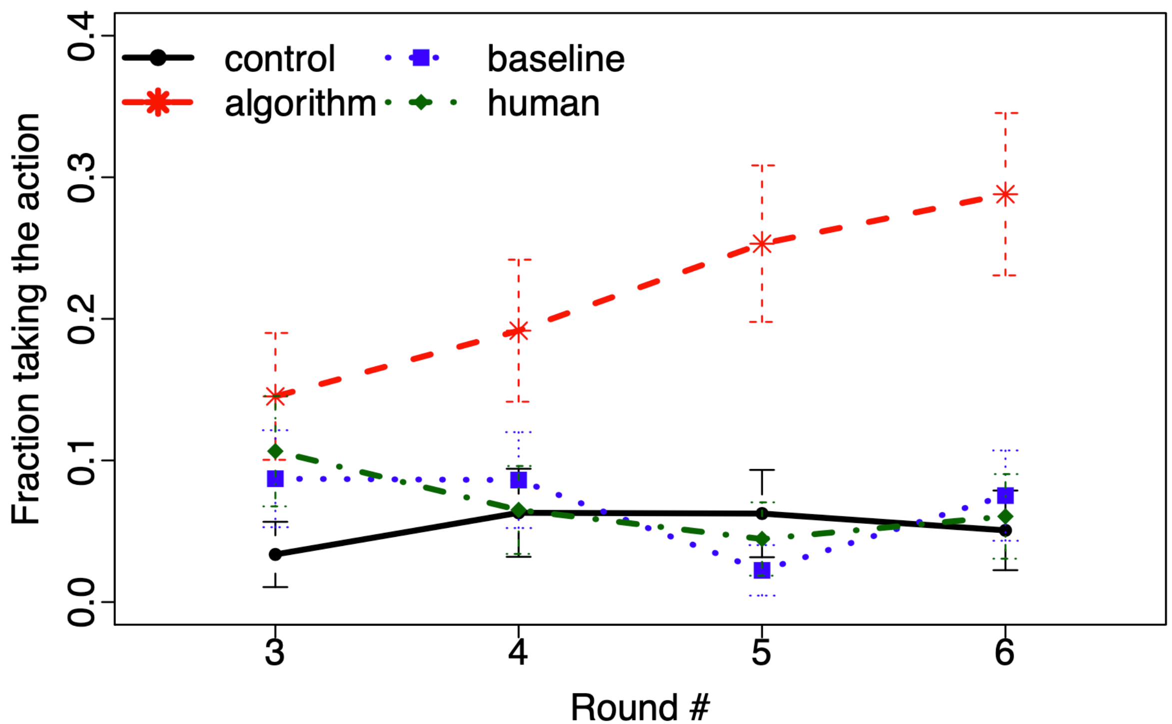
Algorithm Baseline

Results Learning Beyond Tips

Our tip effectively led people to the states they can learn other optimal strategies

Sous-Chef chops 3 times

Part of optimal policy but not stated in any of the tips



Amazon Mechanical Turk, N = 1,011

mean age 34.9, 60% female

Summary

ML framework to leverage behavioral trace data to infer simple tips that help humans

Trace data → **Extract best practices** Machine Learning → **Mine simple tips** Tips → **Improve performance** Humans

Our tips improve performance, speed up learning, help humans adapt to disruption, and uncover other optimal strategies

Performance/compliance tradeoff



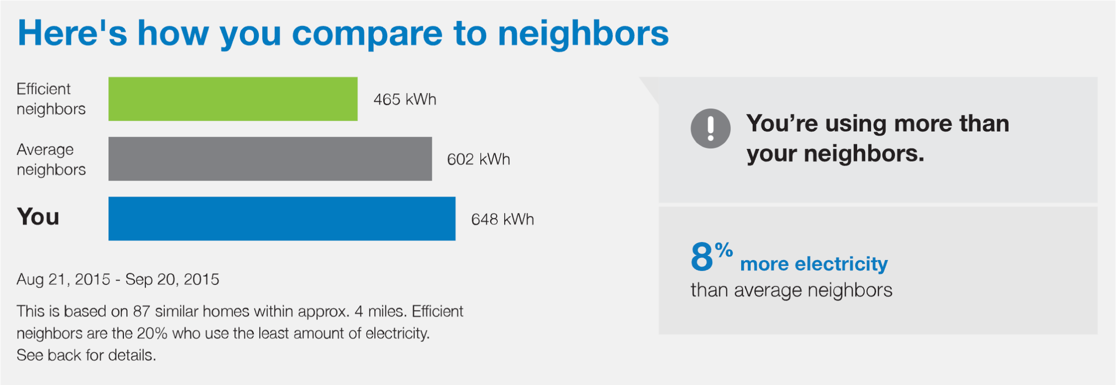
Feedback (+ tips) very welcome!

Hamsa Bastani, Osbert Bastani, **Park Sinchaisri** (parksinchaisri@berkeley.edu / parksinchaisri.github.io)
Minor Revision at *Management Science* Paper available at: https://bit.ly/tipspaper

Improving Compliance?

Improving Compliance

Social information



Allcott 2011, *Journal of Public Economics*

Improving Compliance

Social information

> "The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

Improving Compliance

Social information

"The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay" – incentive to try

Improving Compliance

Social information

> "The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay" – incentive to try

> "You'll earn the maximum bonus if server cooks twice in this round."

in rounds 3-4, back to original scheme in rounds 5-6

Improving Compliance

Social information

> "The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay" – incentive to try

> "You'll earn the maximum bonus if server cooks twice in this round."

in rounds 3-4, back to original scheme in rounds 5-6

"Curriculum" – pacing learning

Improving Compliance

Social information

"The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay" – incentive to try

"You'll earn the maximum bonus if server cooks twice in this round."

in rounds 3-4, back to original scheme in rounds 5-6

"Curriculum" – pacing learning

Algorithm	Human	Hypothetical
Server should cook twice	Server should cook once	Server shouldn't cook

Improving Compliance

Social information

> "The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay" – incentive to try

> "You'll earn the maximum bonus if server cooks twice in this round."

in rounds 3-4, back to original scheme in rounds 5-6

"Curriculum" – pacing learning

Algorithm	Human	Hypothetical
Server should cook twice	Server should cook once	Server shouldn't cook

in round 3

Improving Compliance

Social information

"The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay" – incentive to try

"You'll earn the maximum bonus if server cooks twice in this round."

in rounds 3-4, back to original scheme in rounds 5-6

"Curriculum" – pacing learning

Algorithm		Human		Hypothetical
Server should cook twice	←	Server should cook once	←	Server shouldn't cook
in rounds 4-6		in round 3		

Improving Compliance

Social information

> "The majority of best players adopted this rule [Server Cook Twice], enabling them to achieve the optimal performance of 34 ticks."

in all 4 disrupted rounds (3-6)

"Pay + Social"

"Pay" – incentive to try

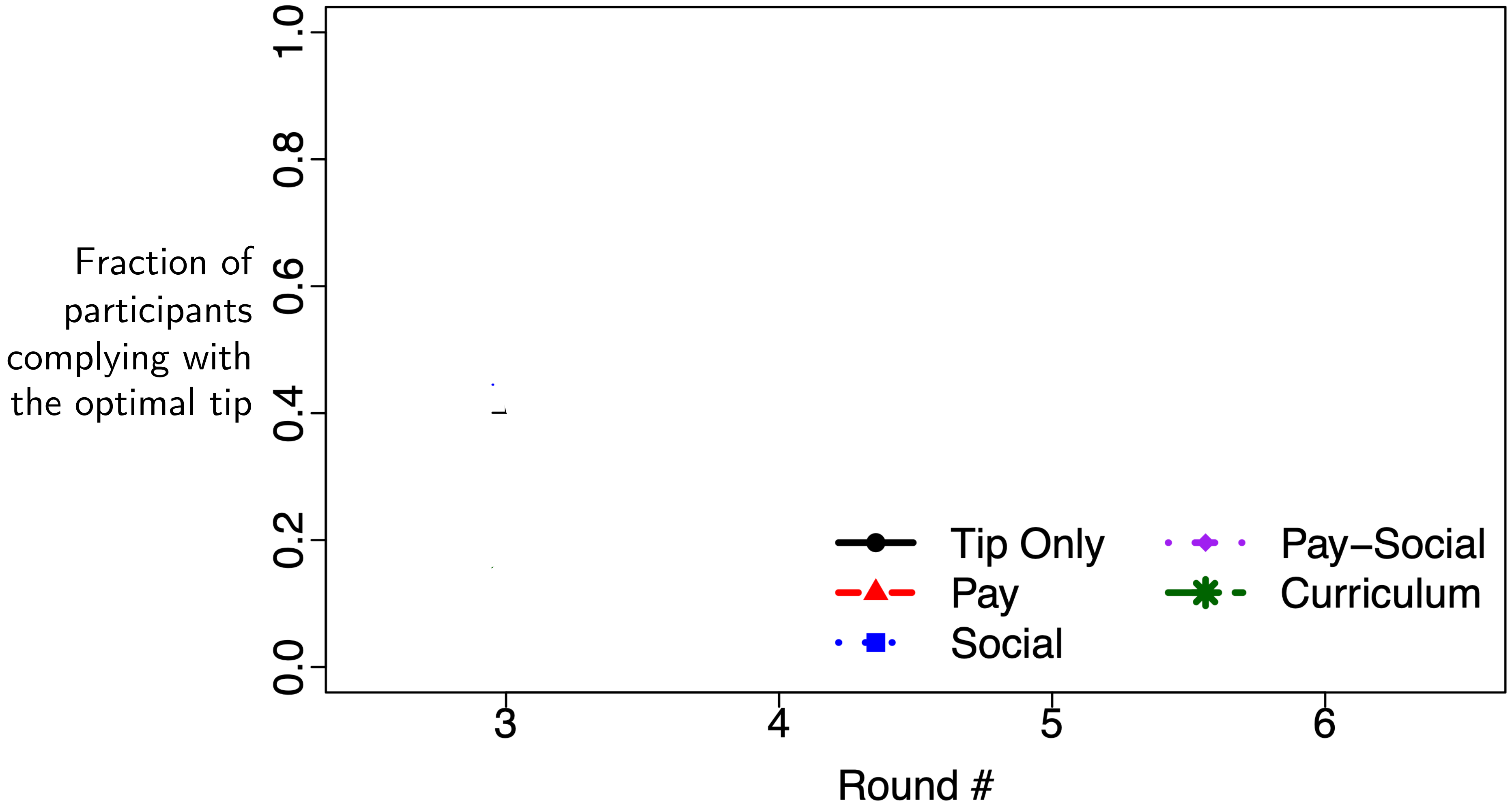
> "You'll earn the maximum bonus if server cooks twice in this round."

in rounds 3-4, back to original scheme in rounds 5-6

"Curriculum" – pacing learning

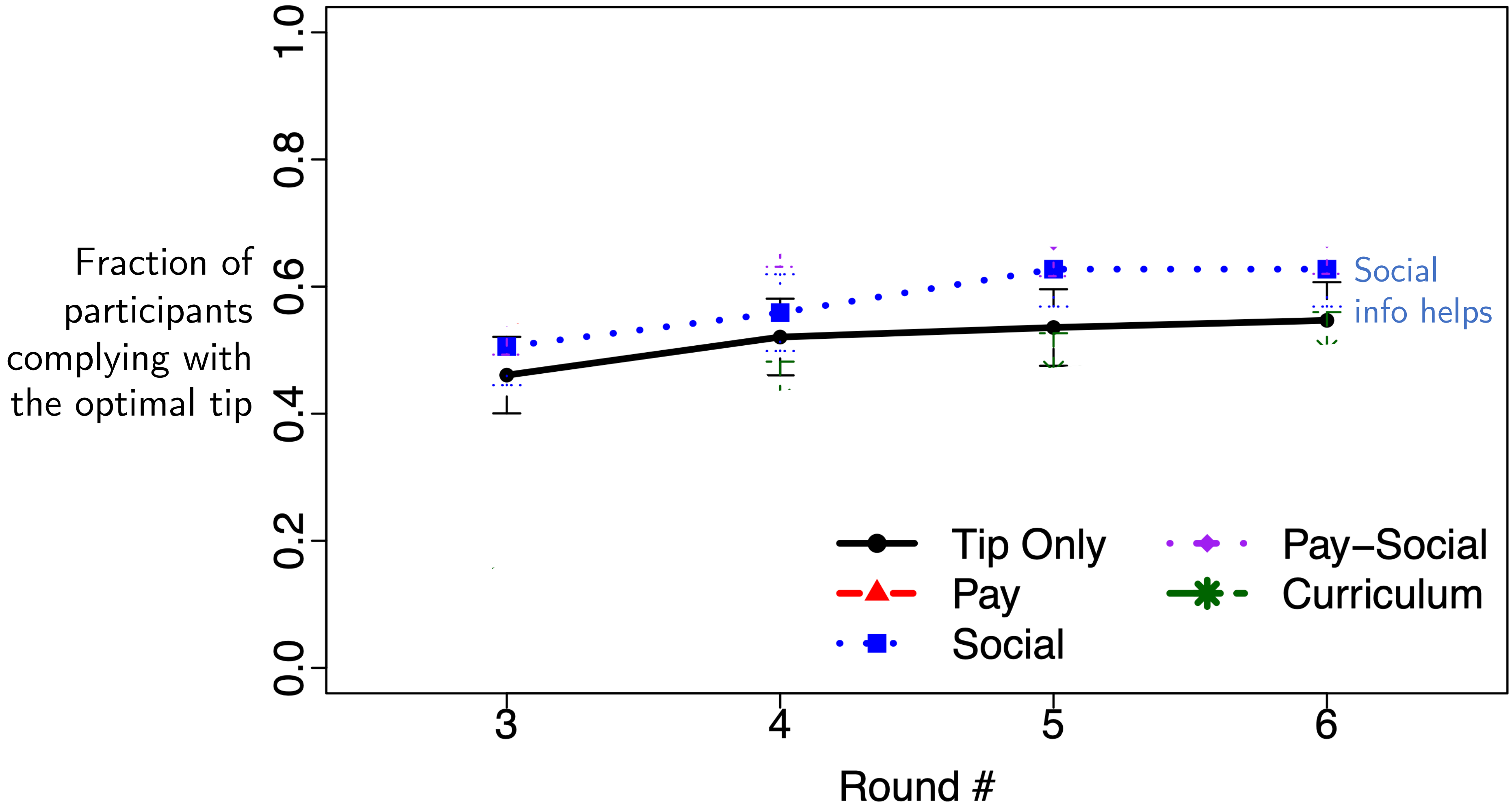
Algorithm		Human		Hypothetical
Server should cook twice	←	Server should cook once	←	Server shouldn't cook
in rounds 4-6		in round 3		

Improving Compliance



Amazon Mechanical Turk, N = 1,416

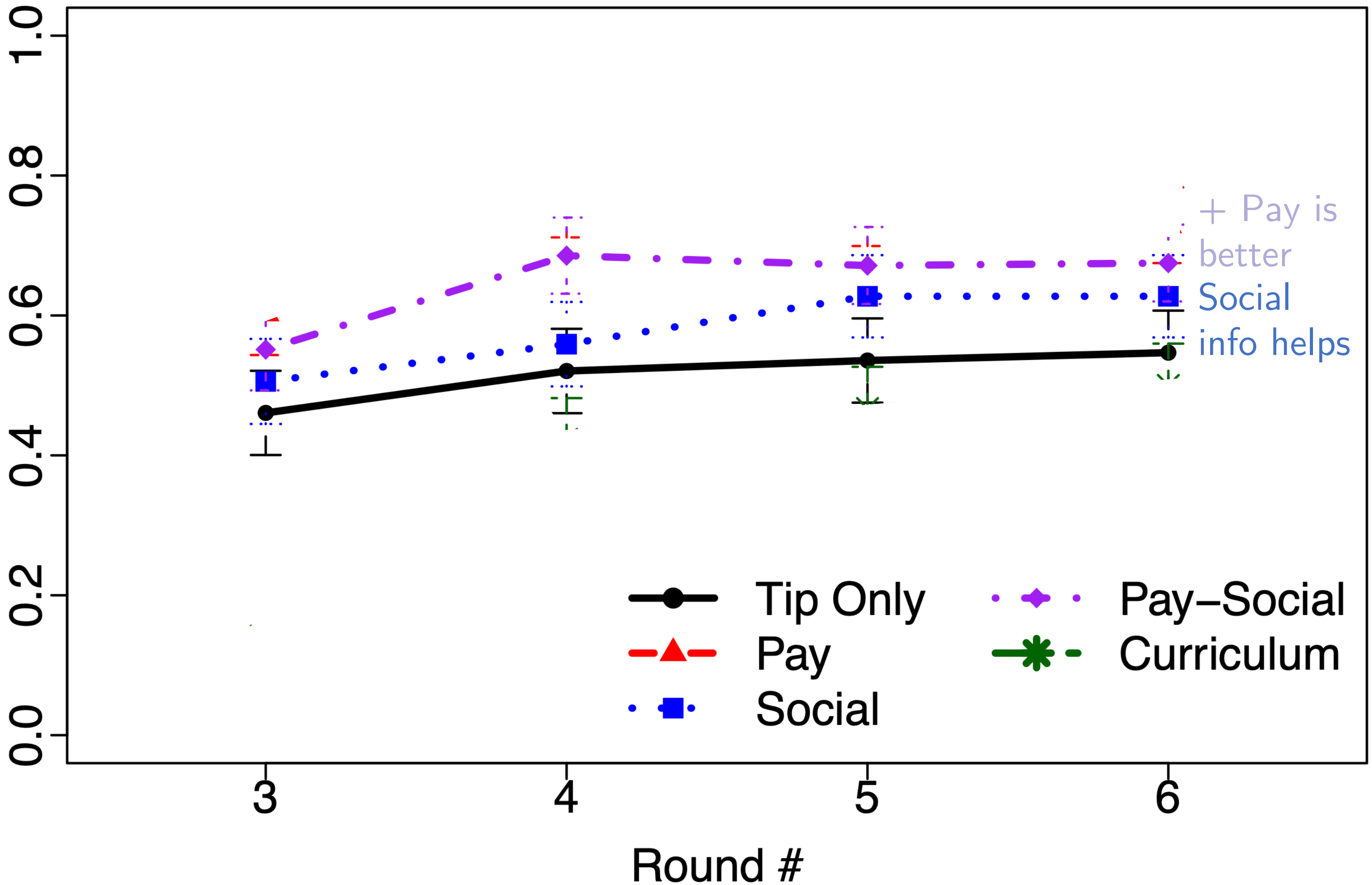
Improving Compliance



Amazon Mechanical Turk, N = 1,416

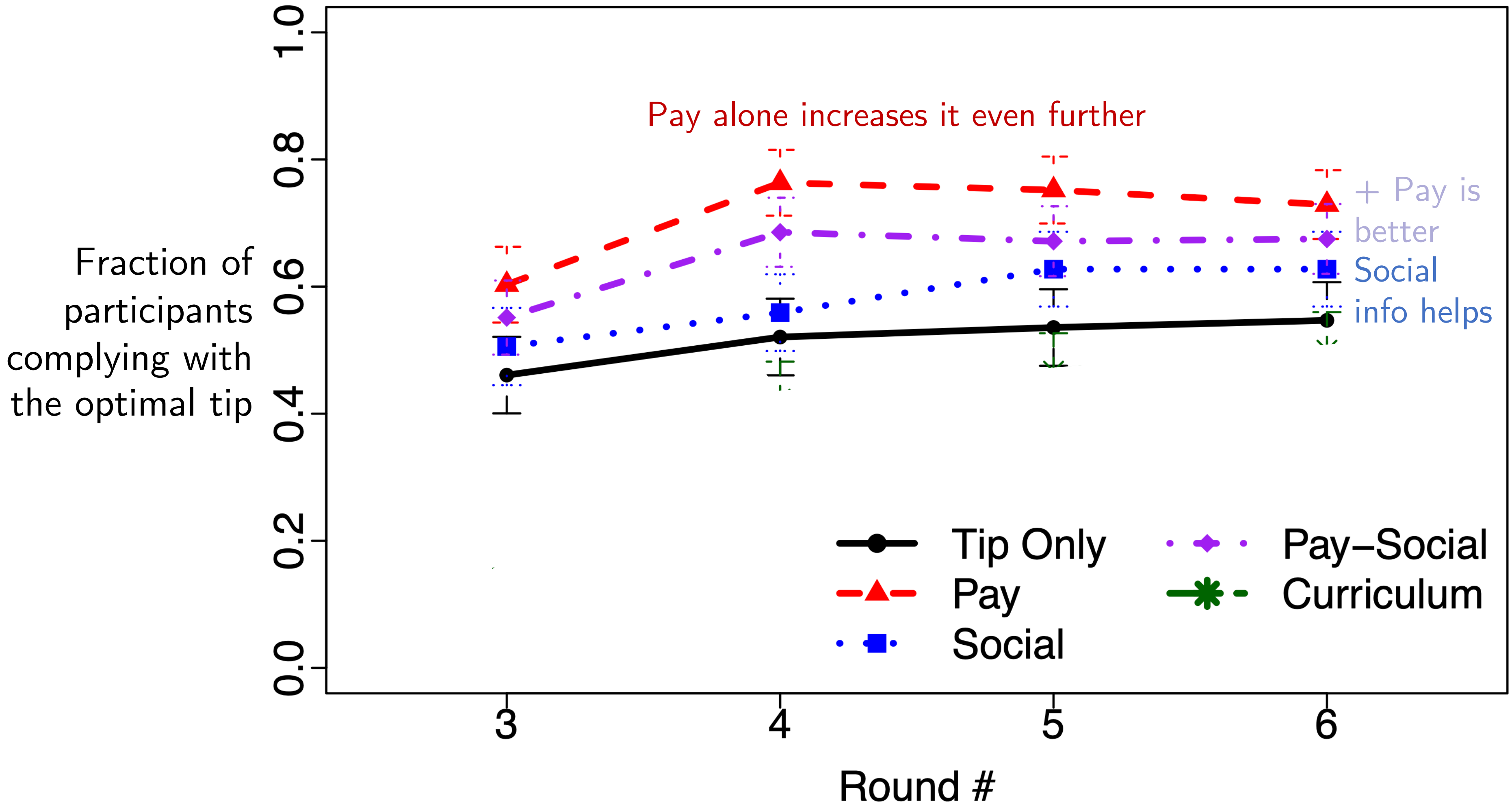
Improving Compliance

Fraction of participants complying with the optimal tip



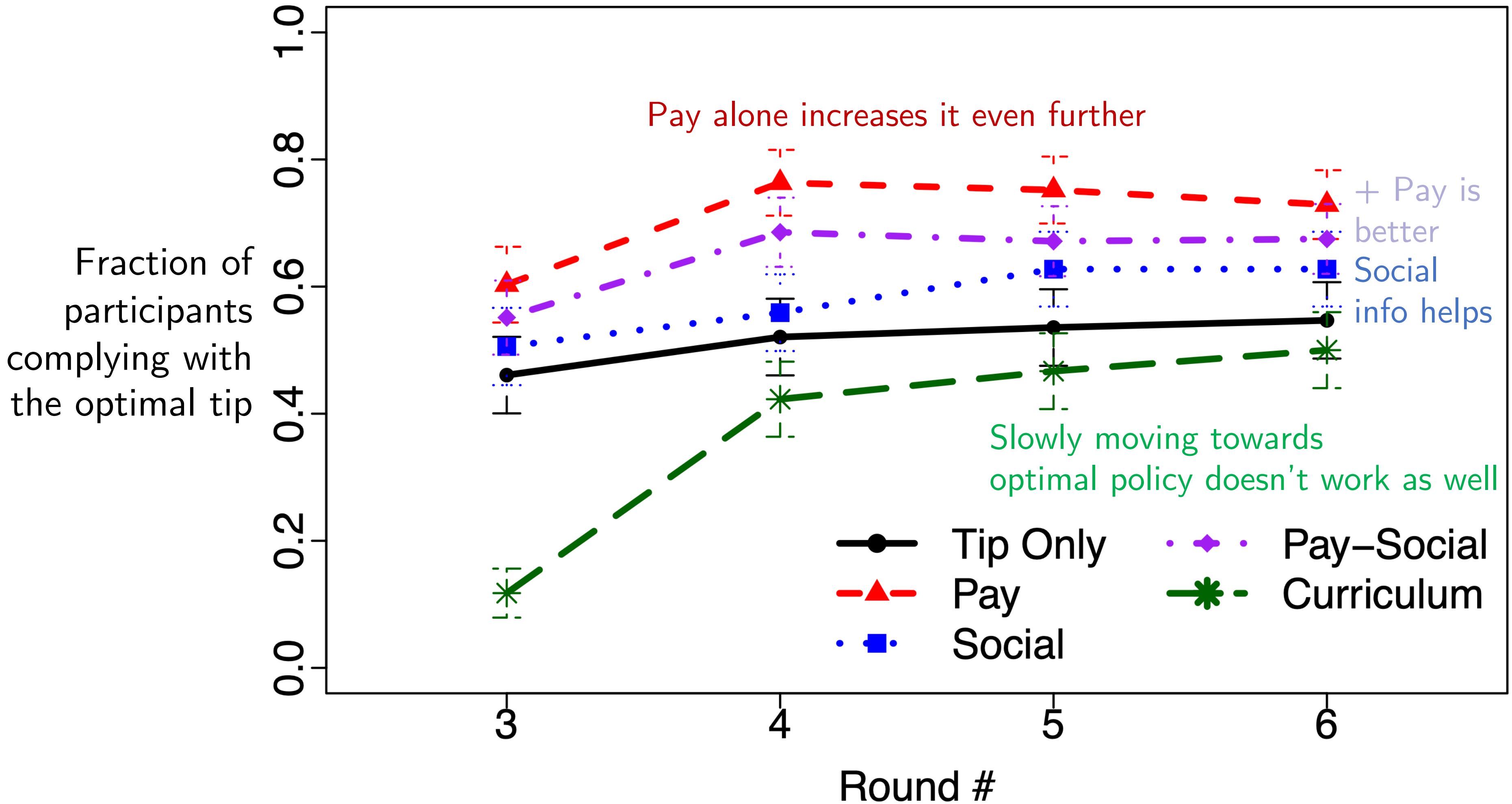
Amazon Mechanical Turk, N = 1,416

Improving Compliance



Amazon Mechanical Turk, N = 1,416

Improving Compliance



Amazon Mechanical Turk, N = 1,416