Global Rewards in Restless Multi-Armed Bandits

And some Applications to Food Rescue

Naveen Raman, July 11th

Food Insecurity

Food Insecurity

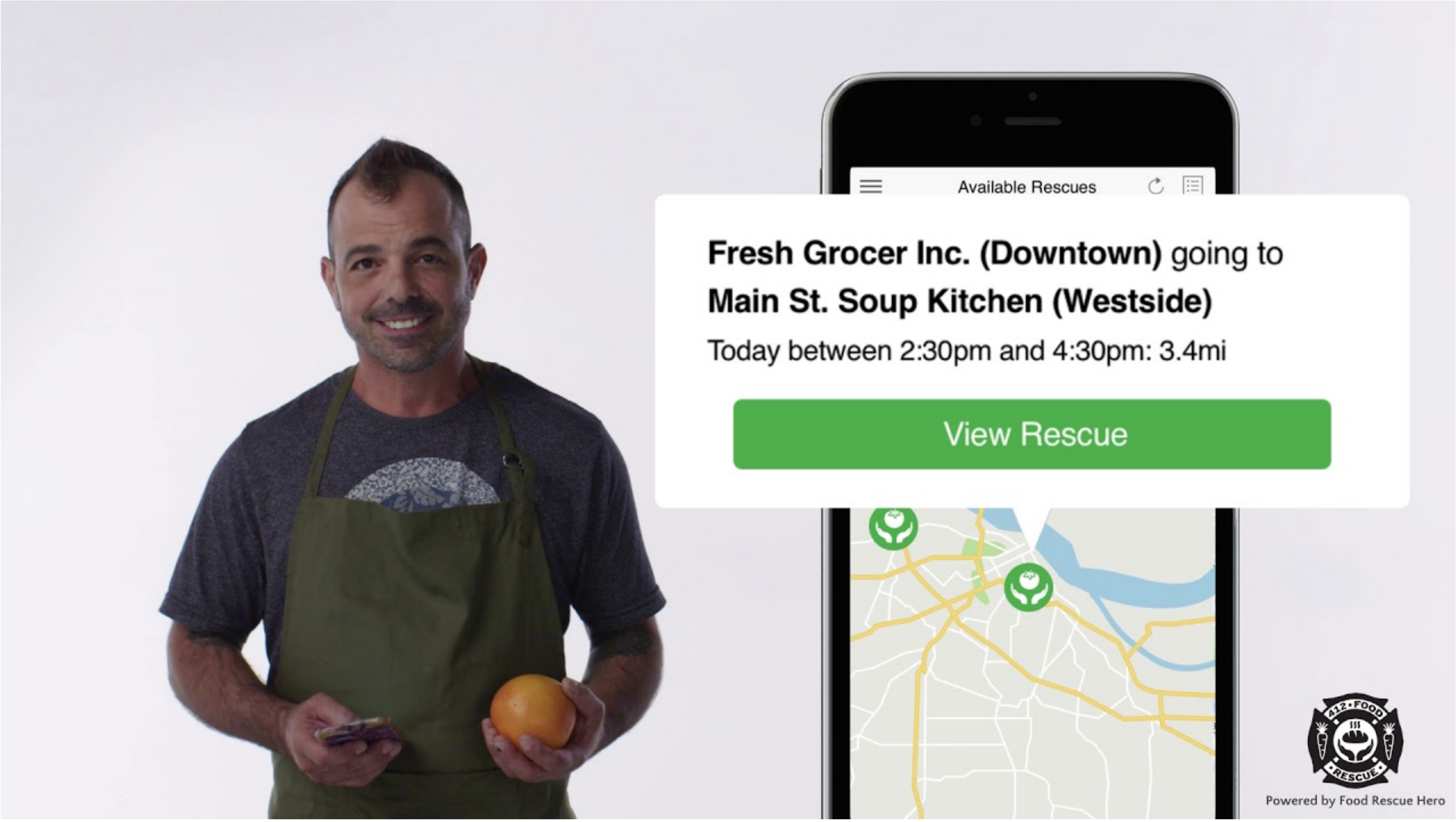
"*Enough food is produced today to feed everyone on the planet, but hunger is on the rise in some parts of the world, and some 821 million people are considered to be "chronically undernourished" - United Nations*

The Role of Food Rescue

The Role of Food Rescue

The Role of Food Rescue

The Role of Food Rescue



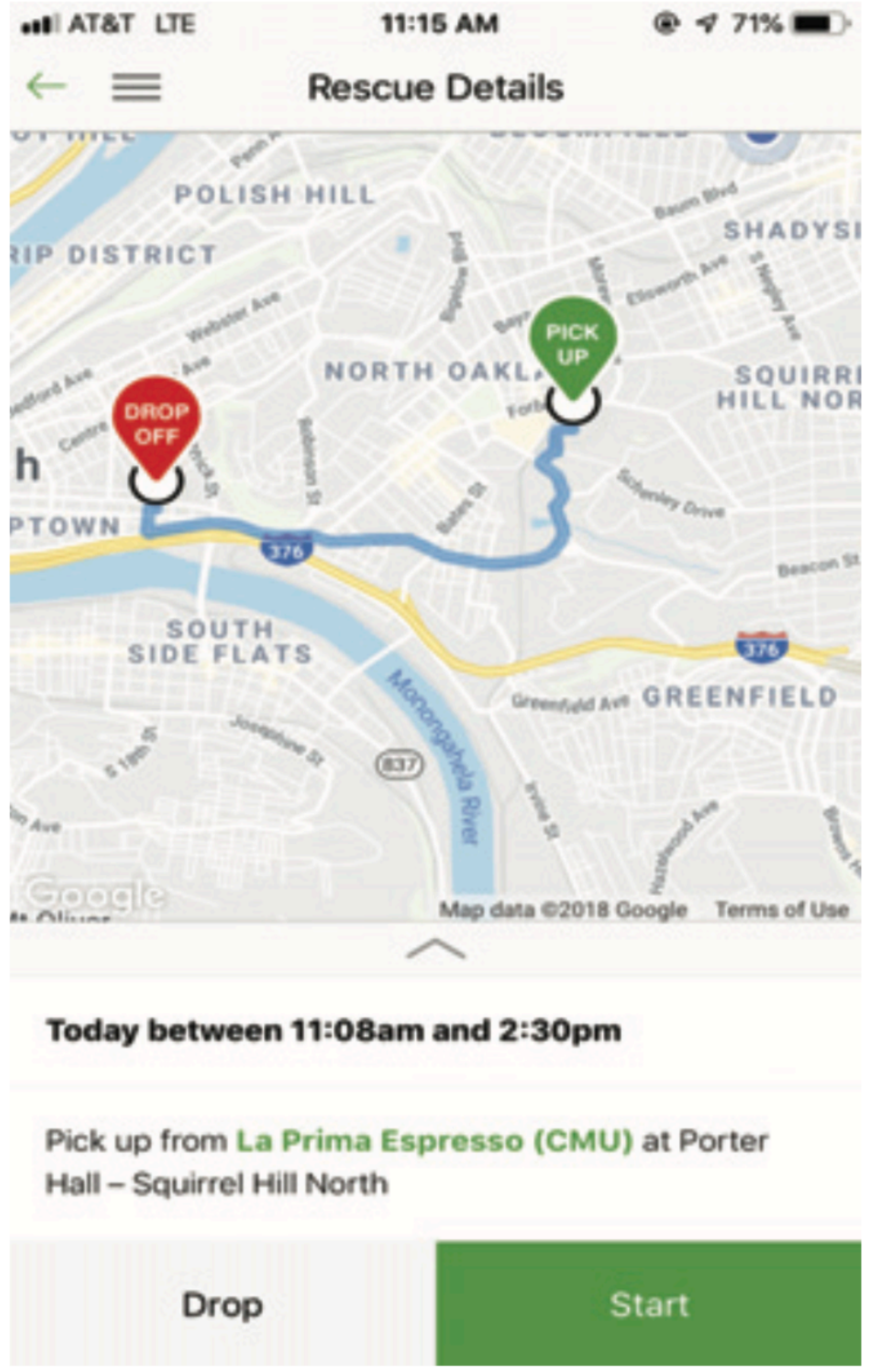
Notifications in Food Rescue

Notifications in Food Rescue

Trip Notification

Notifications in Food Rescue

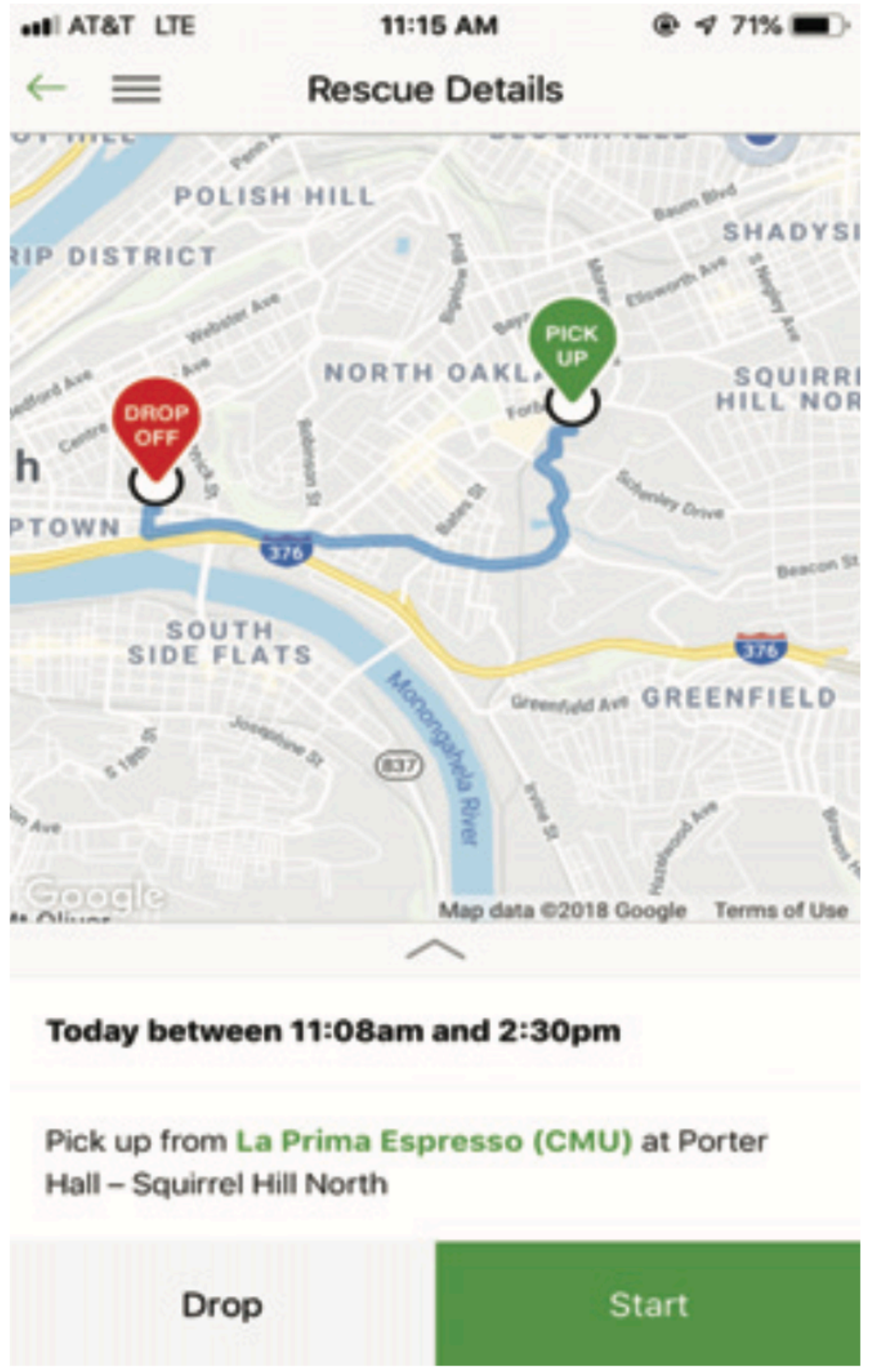
Trip Notification



Trip Acceptance

Notifications in Food Rescue

Trip Notification

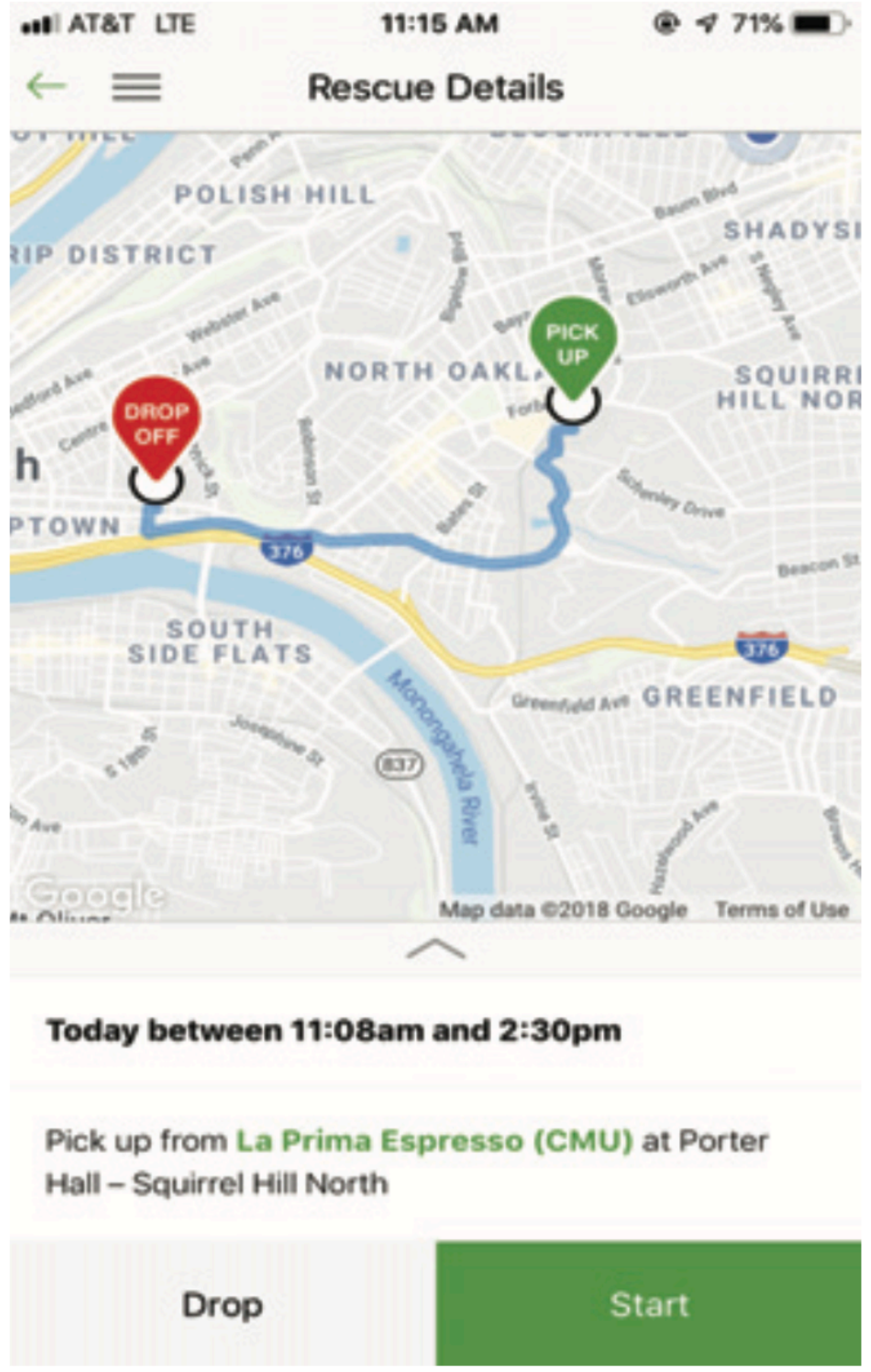


Trip Acceptance

Trip Completion

Notifications in Food Rescue

Trip Notification



Trip Acceptance

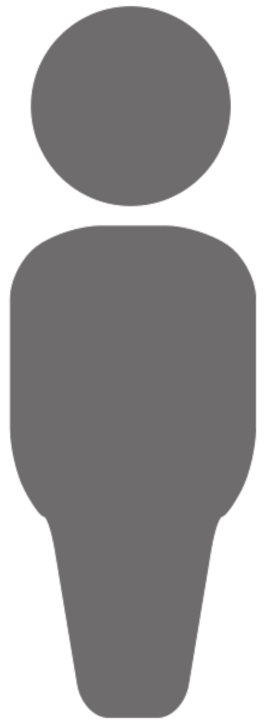
Trip Completion

How can we notify volunteers in Food Rescue to maximize donated food, while keeping volunteers engaged?

A Model of Food Rescue

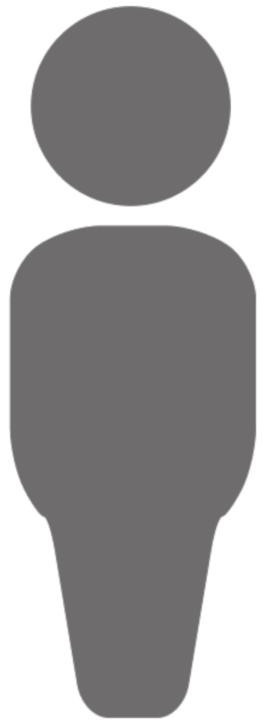
A Model of Food Rescue

Volunteers



A Model of Food Rescue

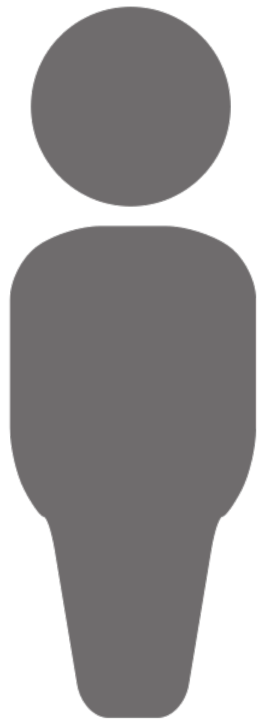
Volunteers



$$s = \{\text{not engaged}, \text{engaged}\}$$
$$a = \{\text{not notified}, \text{notified}\}$$

A Model of Food Rescue

Volunteers

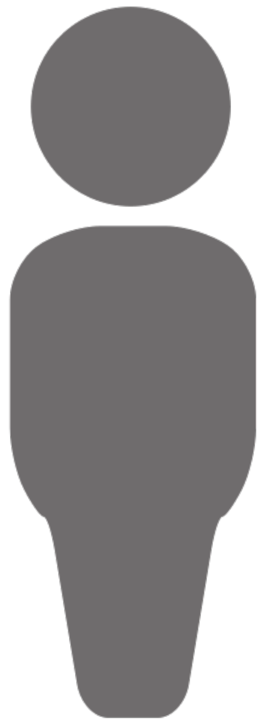


$s = \{\text{not engaged}, \text{engaged}\}$
$a = \{\text{not notified}, \text{notified}\}$
$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

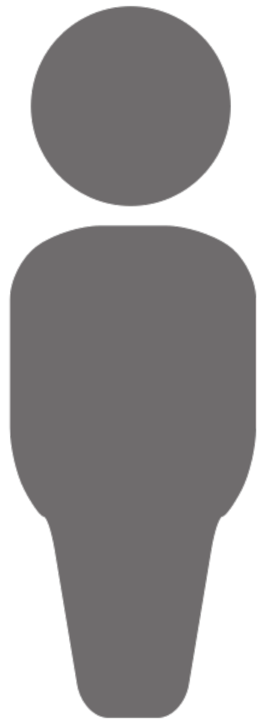
$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$ Match Probability

$P \in [0,1]^{2\times2\times2}$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

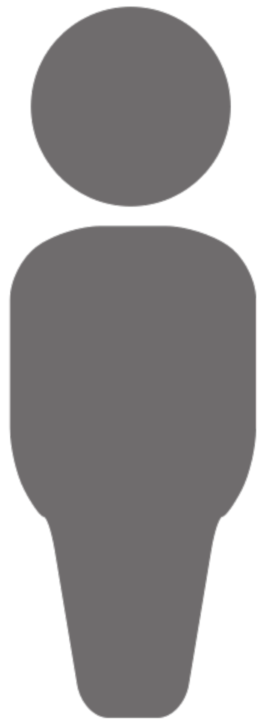
$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

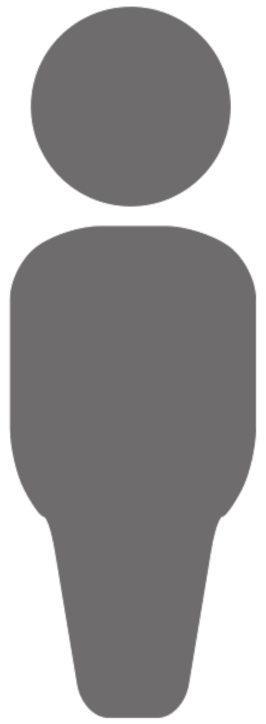
$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2 \times 2 \times 2}$ Transition Matrix

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

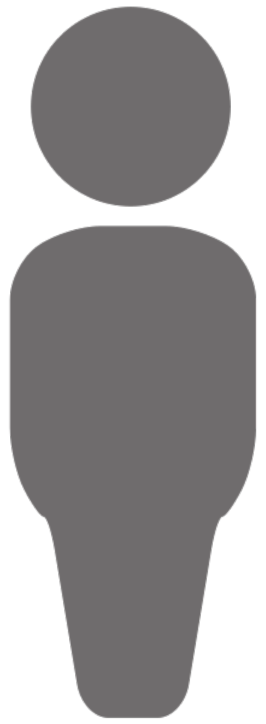
$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

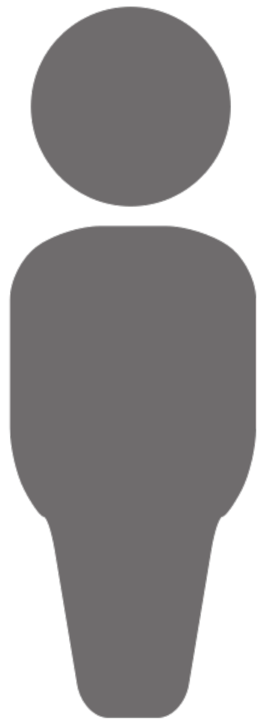
$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

Platform

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

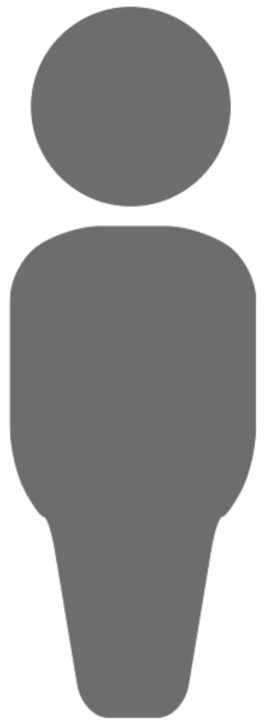
Platform

$\mathbf{s} \in \{0,1\}^N$

$\mathbf{a} \in \{0,1\}^N$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times 2\times 2}$

Platform

$\mathbf{s} \in \{0,1\}^N$

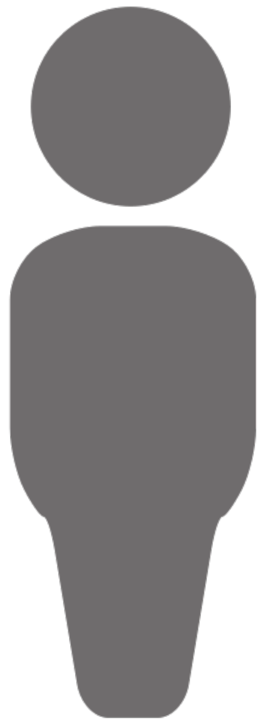
$\mathbf{a} \in \{0,1\}^N$

$$R(\mathbf{s}, \mathbf{a}) = 1 - \prod_{i=1}^{N} (1 - p_i s_i a_i)$$

$$\sum_{i=1}^{N} a_i \leq K$$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

Platform

$\mathbf{s} \in \{0,1\}^N$

$\mathbf{a} \in \{0,1\}^N$

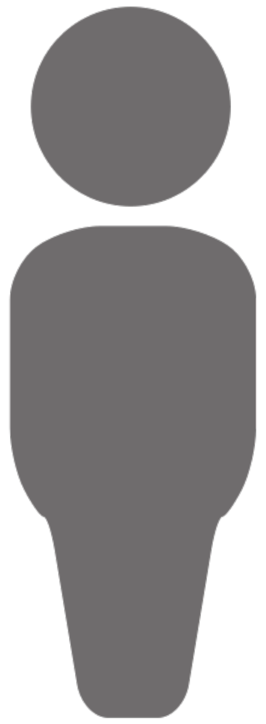
Any volunteer matches

$$R(\mathbf{s}, \mathbf{a}) = 1 - \prod_{i=1}^{N} (1 - p_i s_i a_i)$$

$$\sum_{i=1}^{N} a_i \leq K$$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

Platform

$\mathbf{s} \in \{0,1\}^N$

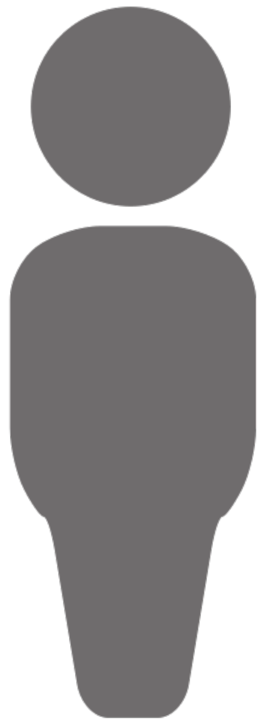
$\mathbf{a} \in \{0,1\}^N$

$$R(\mathbf{s}, \mathbf{a}) = 1 - \prod_{i=1}^{N} (1 - p_i s_i a_i)$$

$$\sum_{i=1}^{N} a_i \leq K$$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

Platform

$\mathbf{s} \in \{0,1\}^N$

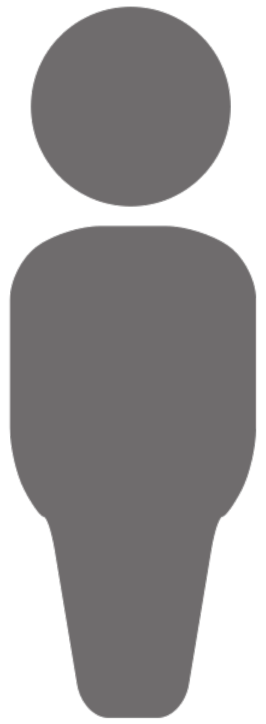
$\mathbf{a} \in \{0,1\}^N$

$$R(\mathbf{s}, \mathbf{a}) = 1 - \prod_{i=1}^{N} (1 - p_i s_i a_i)$$

Budget constraint $\rightarrow$ $$\sum_{i=1}^{N} a_i \leq K$$

A Model of Food Rescue

Volunteers



$s = \{\text{not engaged}, \text{engaged}\}$

$a = \{\text{not notified}, \text{notified}\}$

$p \in [0,1]$

$P \in [0,1]^{2\times2\times2}$

Platform

$\mathbf{s} \in \{0,1\}^N$

$\mathbf{a} \in \{0,1\}^N$

$$R(\mathbf{s}, \mathbf{a}) = 1 - \prod_{i=1}^{N} (1 - p_i s_i a_i)$$

$$\sum_{i=1}^{N} a_i \leq K$$

Optimizing Notifications

Optimizing Notifications

Food Rescue Optimization

$$\boxed{\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1 - \prod_{i=1}^{N}(1 - p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]}$$

Optimizing Notifications

Food Rescue Optimization

$$\boxed{\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1 - \prod_{i=1}^{N}(1 - p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]}$$

Probability any volunteer matches

Optimizing Notifications

Food Rescue Optimization

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1 - \prod_{i=1}^{N}(1 - p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]$$

Probability any volunteer matches

Fraction of engaged volunteers

Optimizing Notifications

Food Rescue Optimization

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1 - \prod_{i=1}^{N}(1 - p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]$$

Probability any volunteer matches

Fraction of engaged volunteers

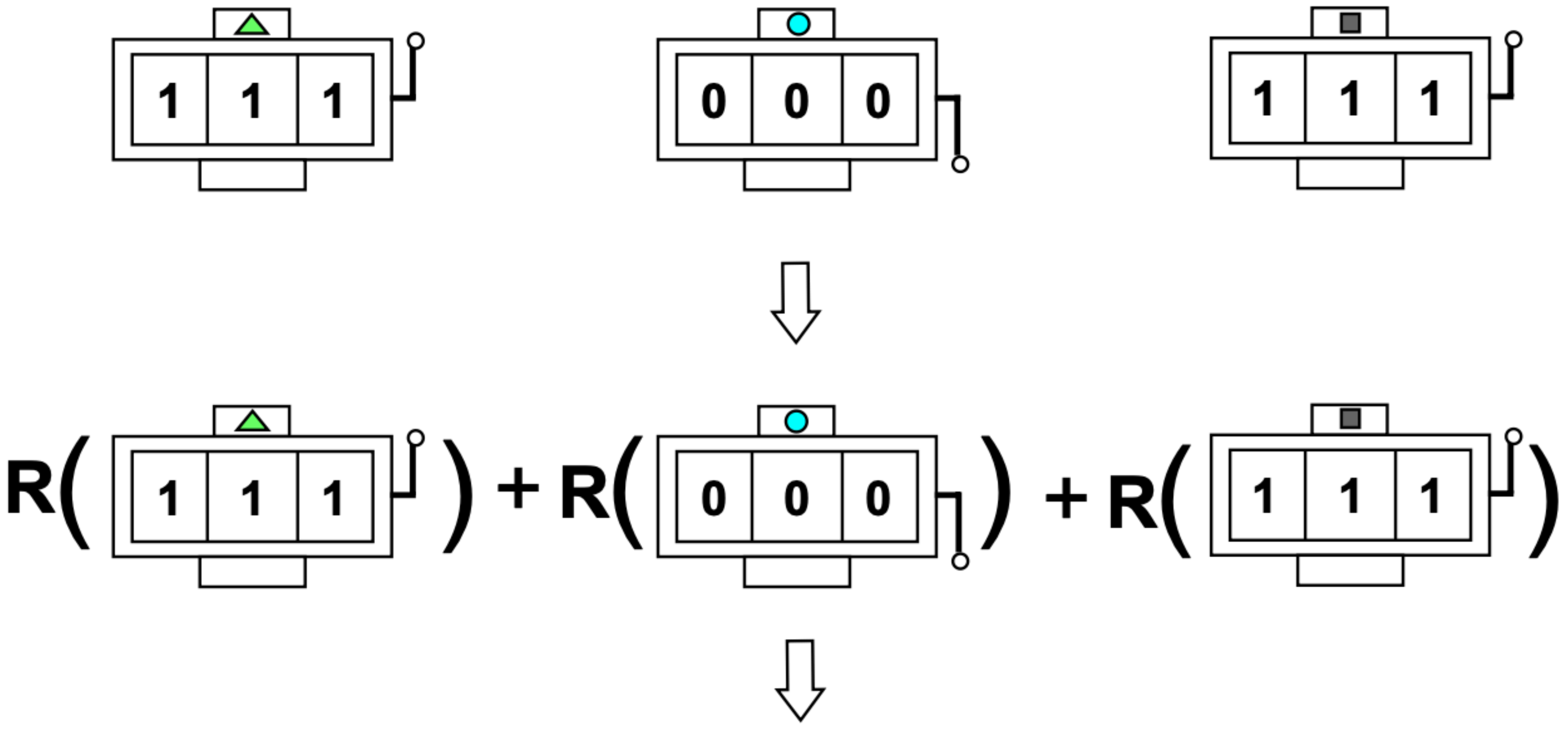
Generalized Problem

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t(R_{\text{glob}}(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) + \sum_{i=1}^{N} R_i(s_i^{(t)}, a_i^{(t)}))]$$

Restless Multi-Armed Bandits with Global Reward

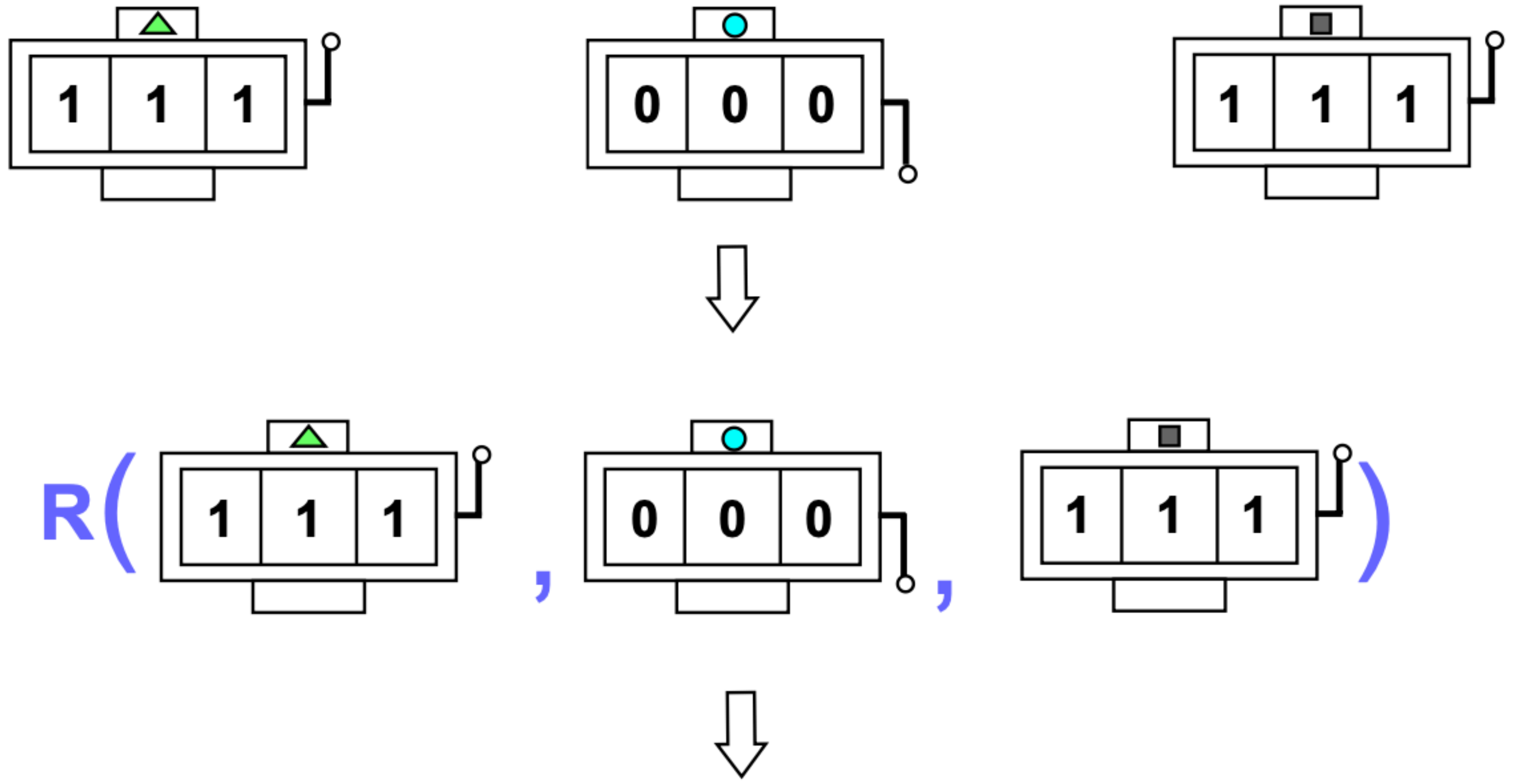
Restless Multi-Armed Bandits with Global Reward

A **Restless Multi-Armed Bandit (RMAB)**



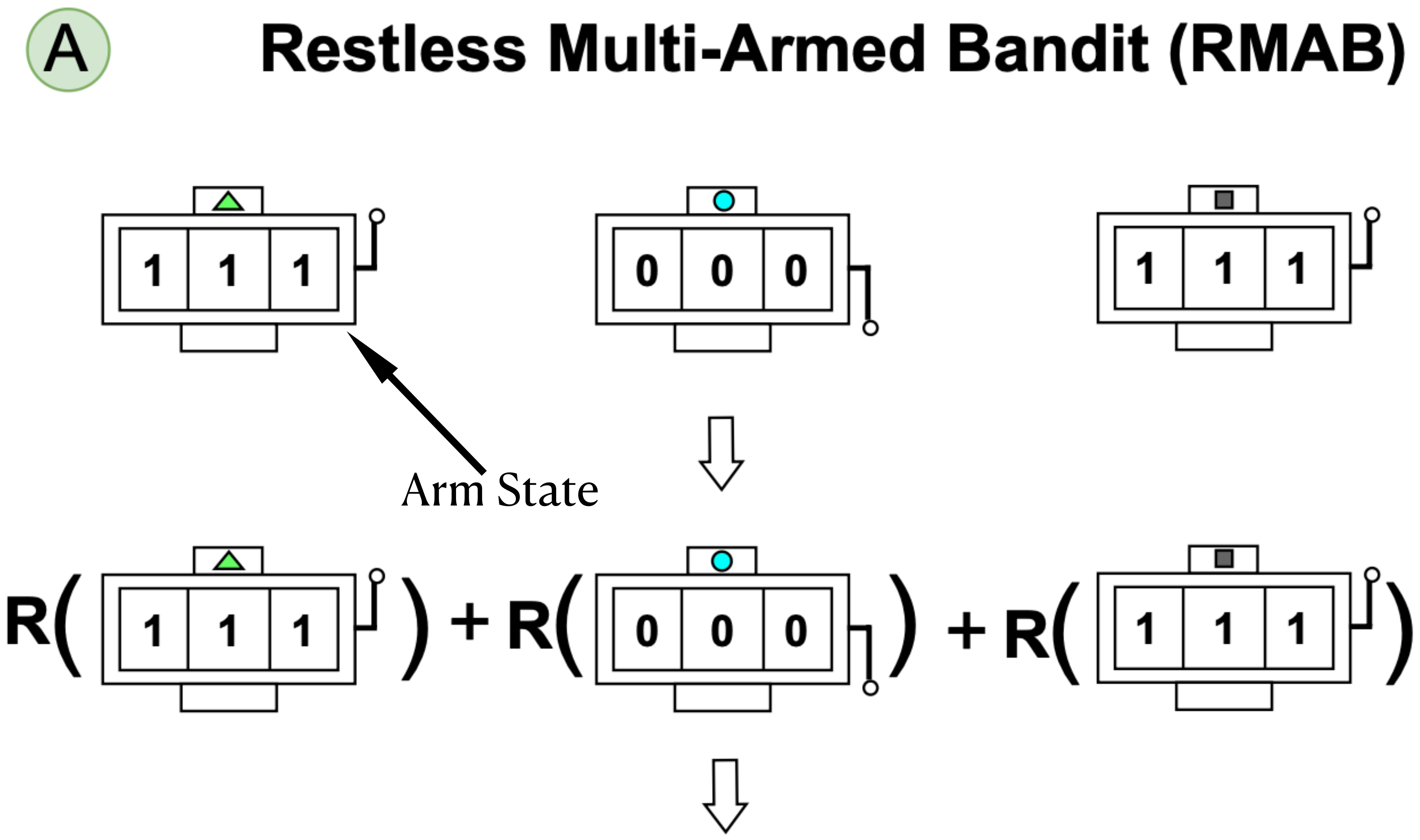
Pulling arms results in separable rewards

B **Restless Multi-Armed Bandit with Global Rewards (RMAB-G)**

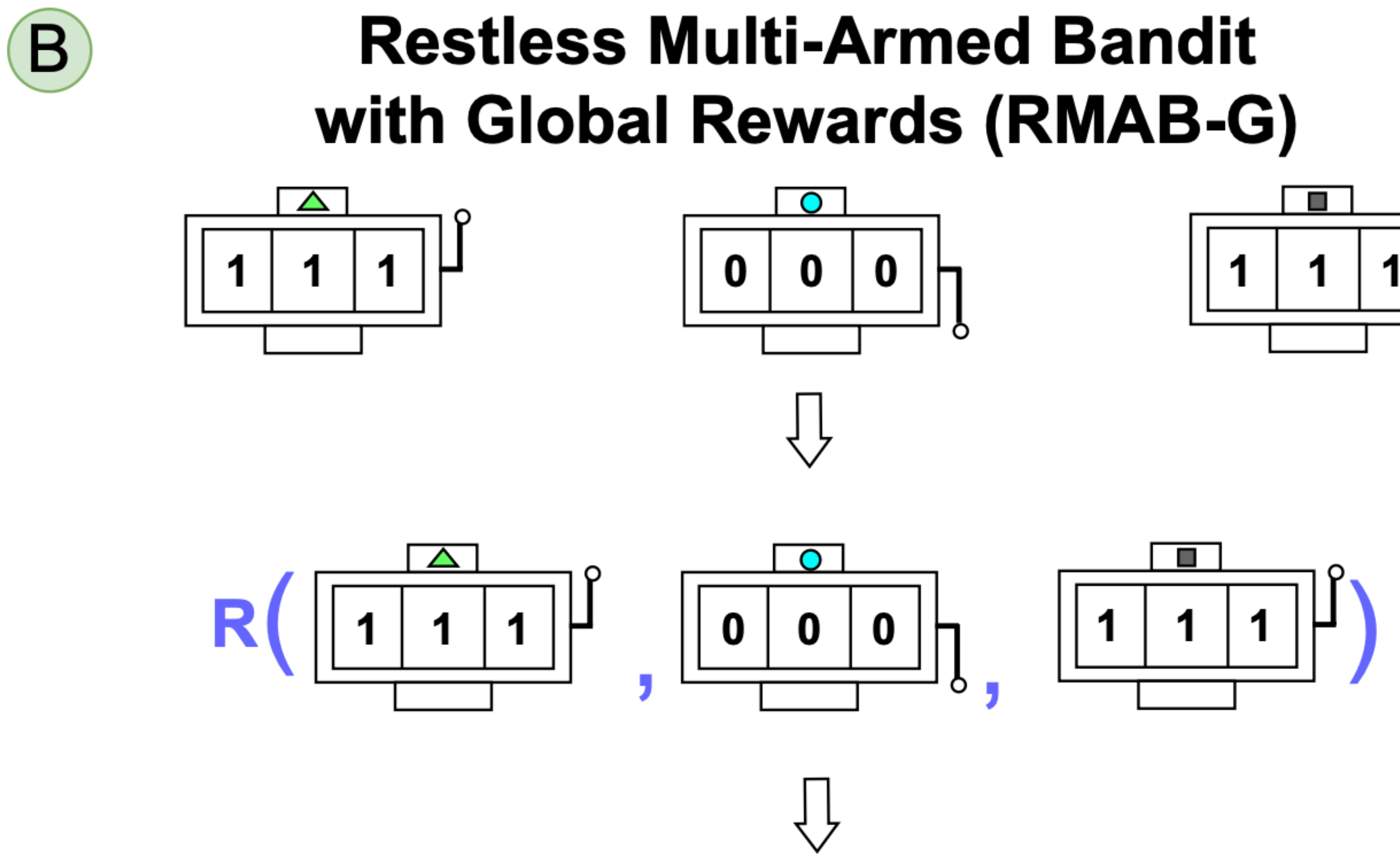


Pulling arms results in **a global reward**

Restless Multi-Armed Bandits with Global Reward

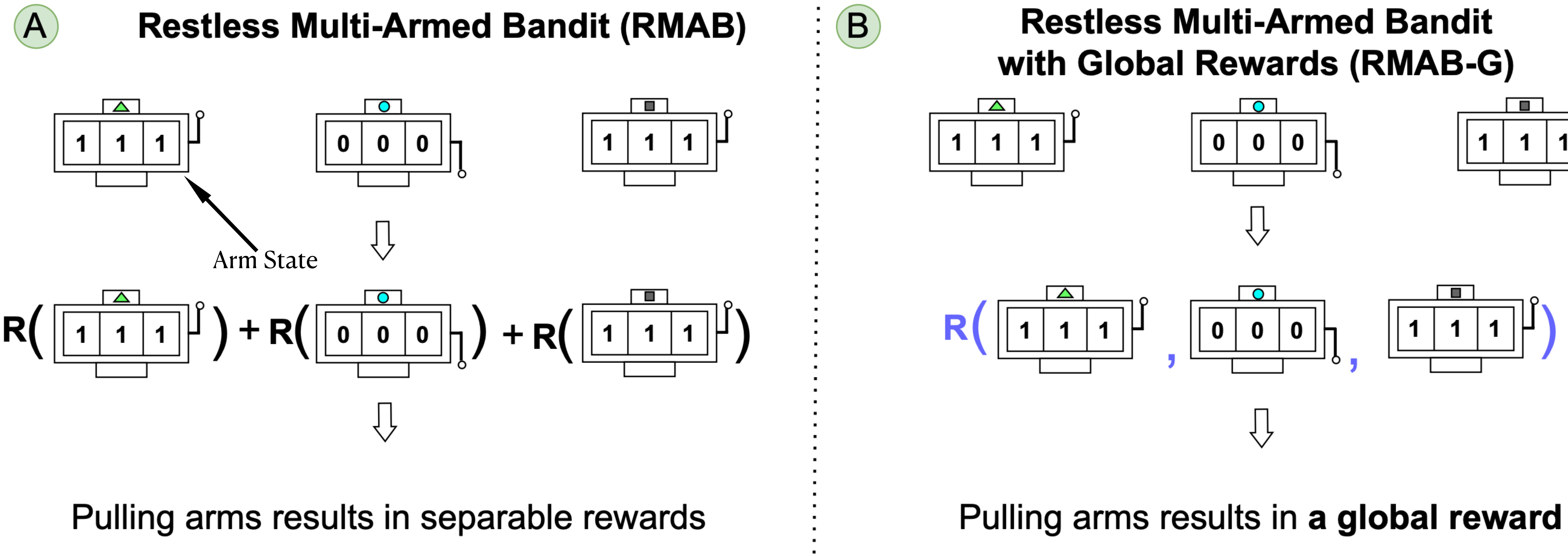


Pulling arms results in separable rewards



Pulling arms results in **a global reward**

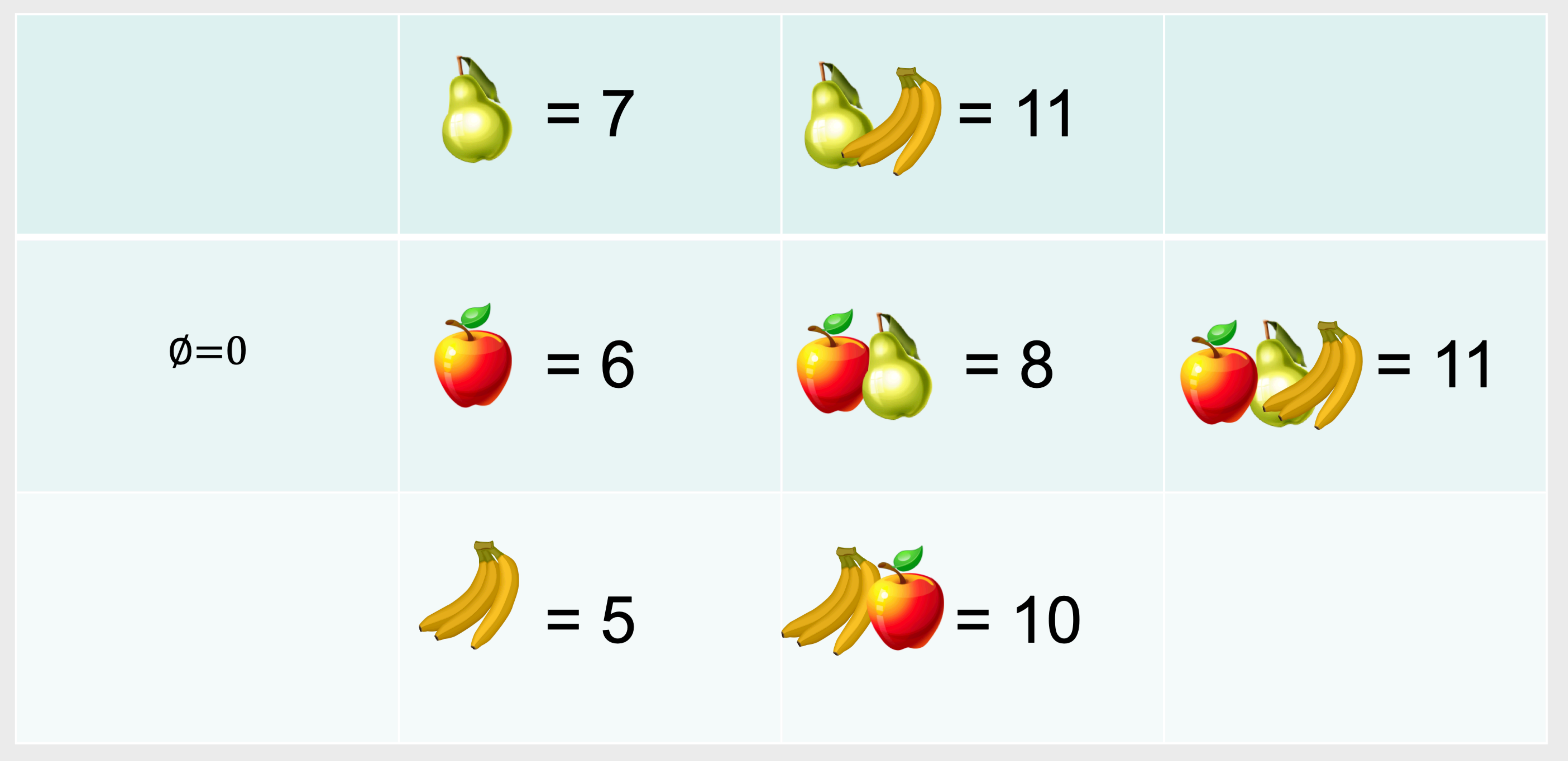
Restless Multi-Armed Bandits with Global Reward



How can we optimize the restless bandits with a global reward?

Submodular Monotonic Functions

Let R_{glob} be **submodular:** Pulling extra arms gives diminishing returns and **monotonic:** Pulling extra arms improves reward



Submodular Monotonic Functions are quickly optimizable and ubiquitous

Recall our Goal

Recall our Goal

Food Rescue Optimization

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1-\prod_{i=1}^{N}(1-p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]$$

Generalized Problem (RMAB-Global)

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t(R_{\text{glob}}(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) + \sum_{i=1}^{N} R_i(s_i^{(t)}, a_i^{(t)}))]$$

Recall our Goal

Food Rescue Optimization

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1-\prod_{i=1}^{N}(1-p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]$$

Generalized Problem (RMAB-Global)

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t(R_{\text{glob}}(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) + \sum_{i=1}^{N} R_i(s_i^{(t)}, a_i^{(t)}))]$$

What are existing solutions, without the global reward?

Why Applying Whittle Indices is Difficult

Why Applying Whittle Indices is Difficult

Whittle Indices: Optimal policy for Restless Bandit

Pulls the arms with the largest value for some index, computed as

$$w_i(s_i) = \min_w \{w \;\; Q_{i,w}(s_i,0) > Q_{i,w}(s_i,1)\}$$

$$Q_{i,w}(s_i, a_i) = -wa_i + R_i(s_i, a_i) + \gamma \sum_{s'} P_i(s_i, a_i, s')V_{i,w}(s'), V_{i,w}(s') = \max_a Q_{i,w}(s', a)$$

Why Applying Whittle Indices is Difficult

Whittle Indices: Optimal policy for Restless Bandit

Pulls the arms with the largest value for some index, computed as

$$w_i(s_i) = \min_{w} \{ w \;\; Q_{i,w}(s_i, 0) > Q_{i,w}(s_i, 1) \}$$

$$Q_{i,w}(s_i, a_i) = -wa_i + R_i(s_i, a_i) + \gamma \sum_{s'} P_i(s_i, a_i, s')V_{i,w}(s'), \; V_{i,w}(s') = \max_{a} Q_{i,w}(s', a)$$

Q-value with penalty w

Why Applying Whittle Indices is Difficult

Whittle Indices: Optimal policy for Restless Bandit

Pulls the arms with the largest value for some index, computed as

Penalty where pulling = not pulling

$$w_i(s_i) = \min_w \{ w \quad Q_{i,w}(s_i,0) > Q_{i,w}(s_i,1) \}$$

$$Q_{i,w}(s_i, a_i) = -wa_i + R_i(s_i, a_i) + \gamma \sum_{s'} P_i(s_i, a_i, s')V_{i,w}(s'), \; V_{i,w}(s') = \max_a Q_{i,w}(s', a)$$

Q-value with penalty w

Why Applying Whittle Indices is Difficult

Whittle Indices: Optimal policy for Restless Bandit

Pulls the arms with the largest value for some index, computed as

Penalty where pulling = not pulling

$$w_i(s_i) = \min_w \{w \quad Q_{i,w}(s_i,0) > Q_{i,w}(s_i,1)\}$$

$$Q_{i,w}(s_i, a_i) = -wa_i + R_i(s_i, a_i) + \gamma \sum_{s'} P_i(s_i, a_i, s')V_{i,w}(s'), V_{i,w}(s') = \max_a Q_{i,w}(s', a)$$

Q-value with penalty w

Applying Whittle Indices requires **separable reward function**, which we don't have

Why Using Reinforcement Learning is Difficult

Why Using Reinforcement Learning is Difficult

State Space Size: 2^N

Action Space Size: $\binom{N}{K}$

Why Using Reinforcement Learning is Difficult

State Space Size: 2^N

Action Space Size: $\binom{N}{K}$

$\boxed{\text{Learning on such large state and action spaces is difficult, even approximately}}$

Why Using Reinforcement Learning is Difficult

State Space Size: 2^N

Action Space Size: $\binom{N}{K}$

Learning on such large state and action spaces is difficult, even approximately

We verify this later using Deep-Q Networks (DQNs)

Main New Method: Linear- and Shapley-Whittle

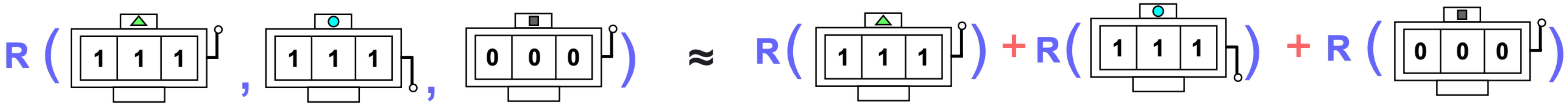
Main New Method: Linear- and Shapley-Whittle

Two Methods of decomposing global reward into Linear Sum

Main New Method: Linear- and Shapley-Whittle

Two Methods of decomposing global reward into Linear Sum

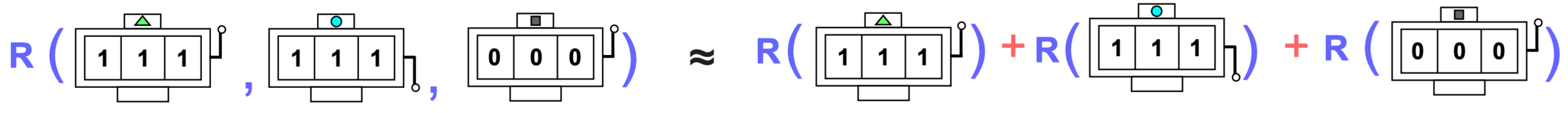
Linear-Whittle Index



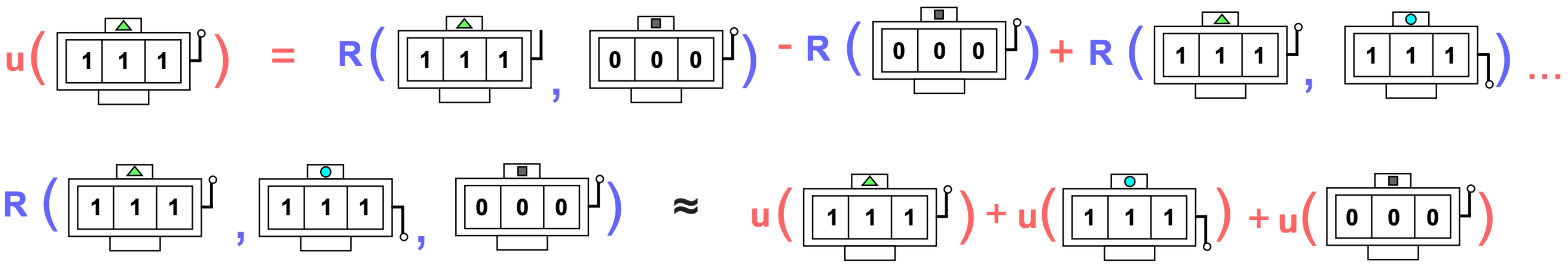
Main New Method: Linear- and Shapley-Whittle

Two Methods of decomposing global reward into Linear Sum

Linear-Whittle Index



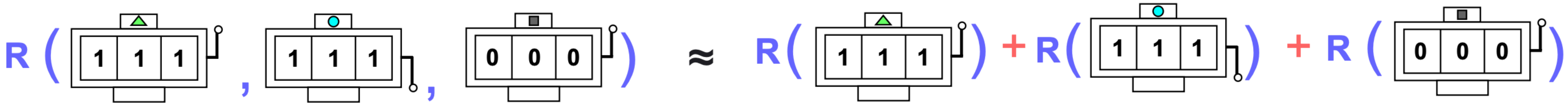
Shapley-Whittle Index



Main New Method: Linear- and Shapley-Whittle

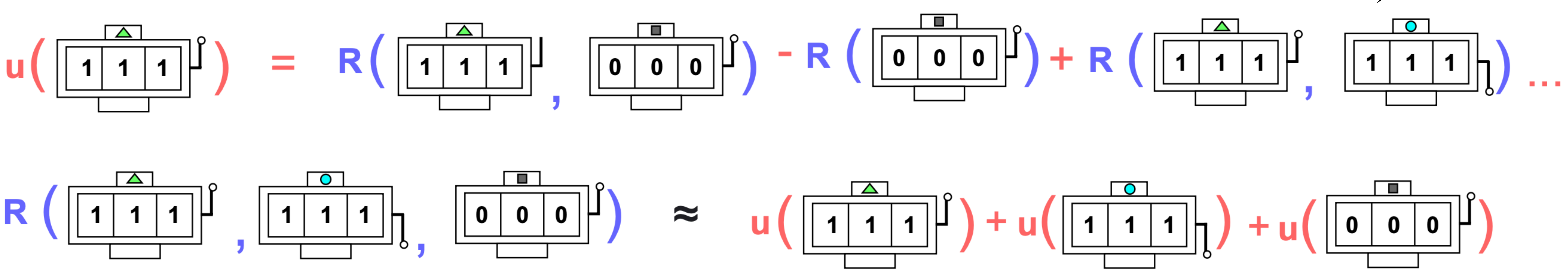
Two Methods of decomposing global reward into Linear Sum

Linear-Whittle Index



Approximate Shapley Value of one arm

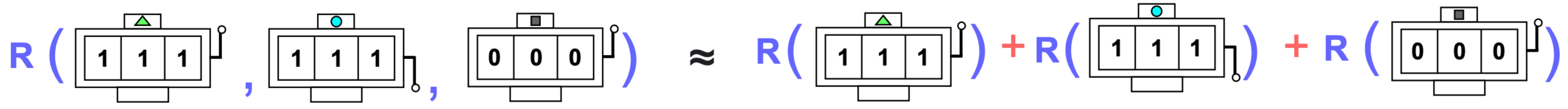
Shapley-Whittle Index



Main New Method: Linear- and Shapley-Whittle

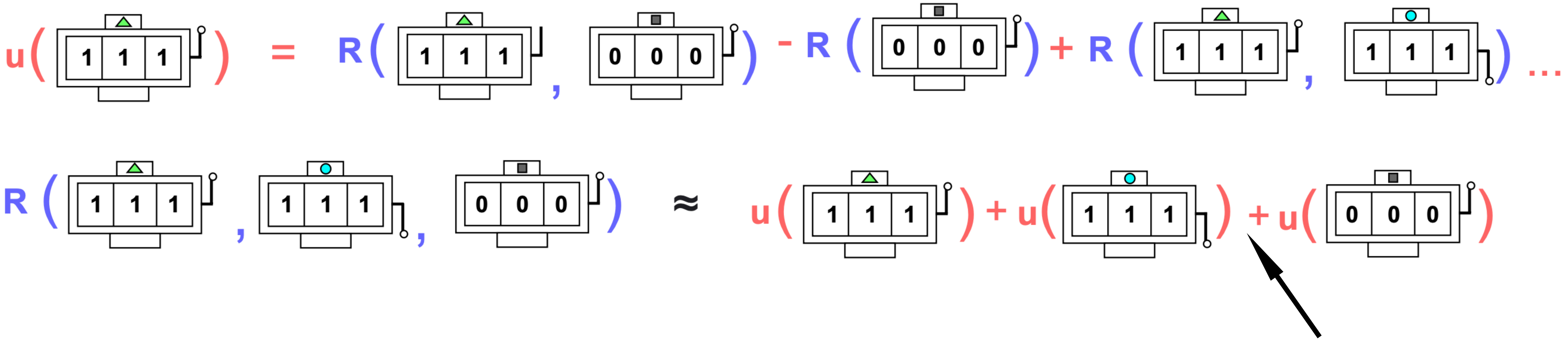
Two Methods of decomposing global reward into Linear Sum

Linear-Whittle Index



Approximate Shapley Value of one arm

Shapley-Whittle Index

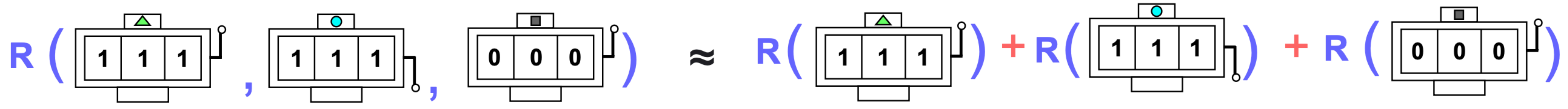


Decompose into sum of Shapley Values

Main New Method: Linear- and Shapley-Whittle

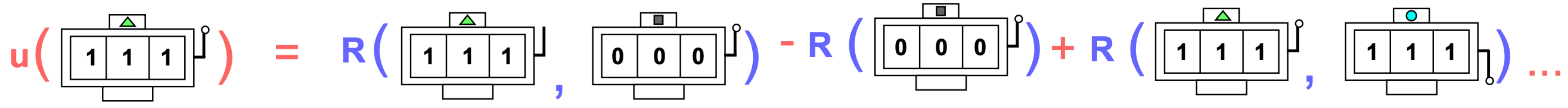
Two Methods of decomposing global reward into Linear Sum

Linear-Whittle Index



Approximate Shapley Value of one arm

Shapley-Whittle Index

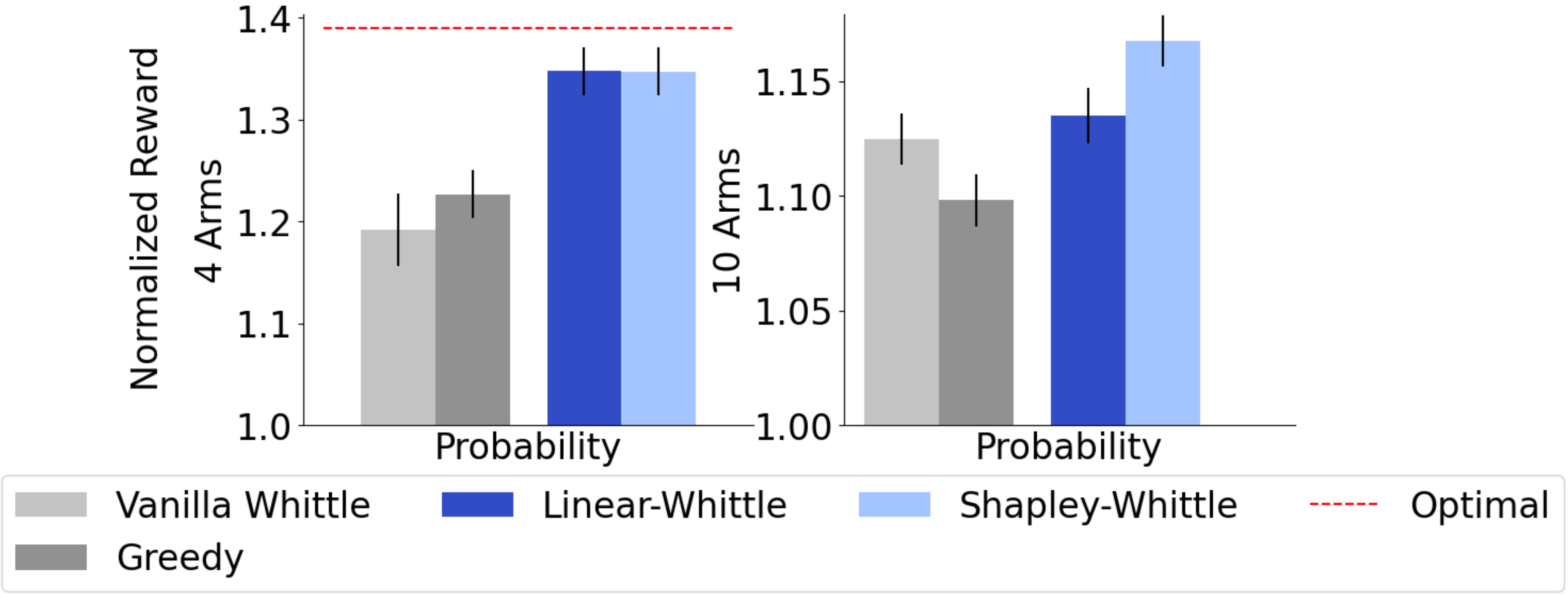


Decompositions allow us to use Whittle Indices

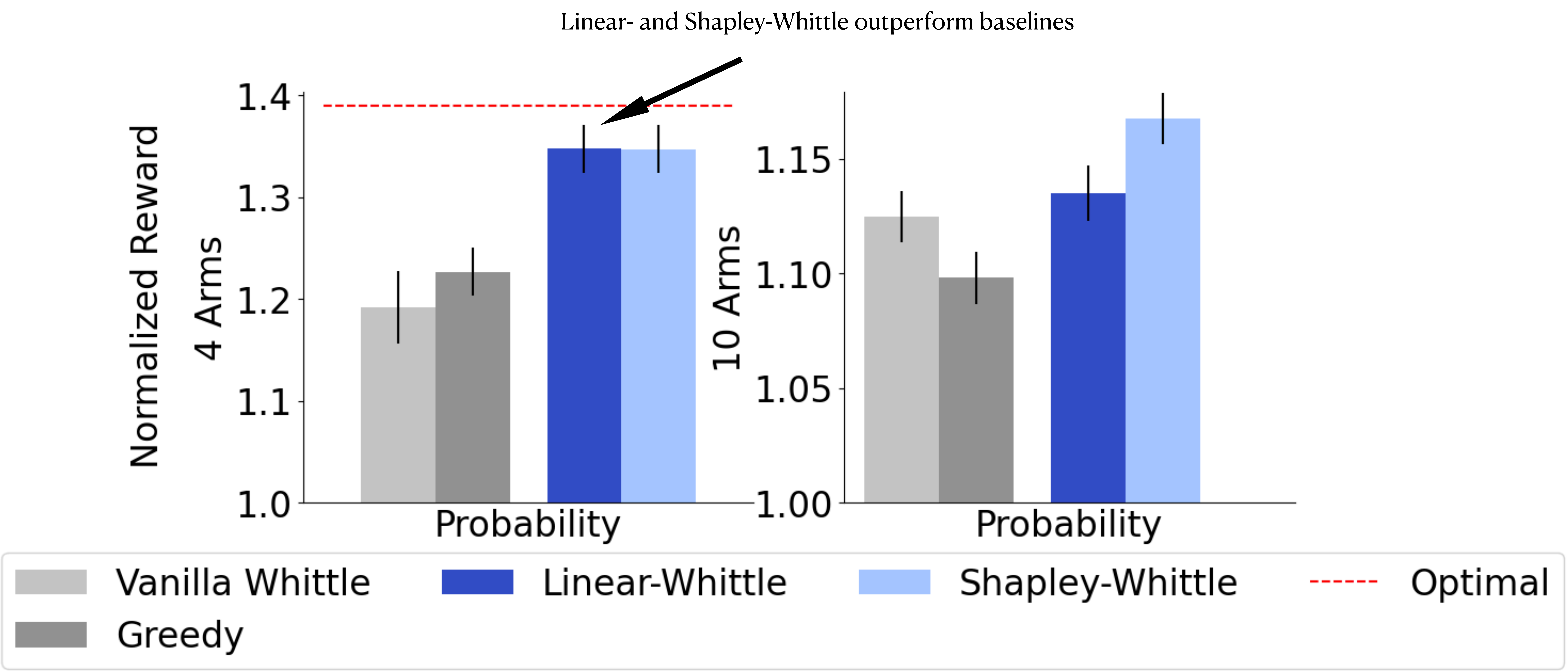
Decompose into sum of Shapley Values

Synthetic Empirical Verification

Synthetic Empirical Verification



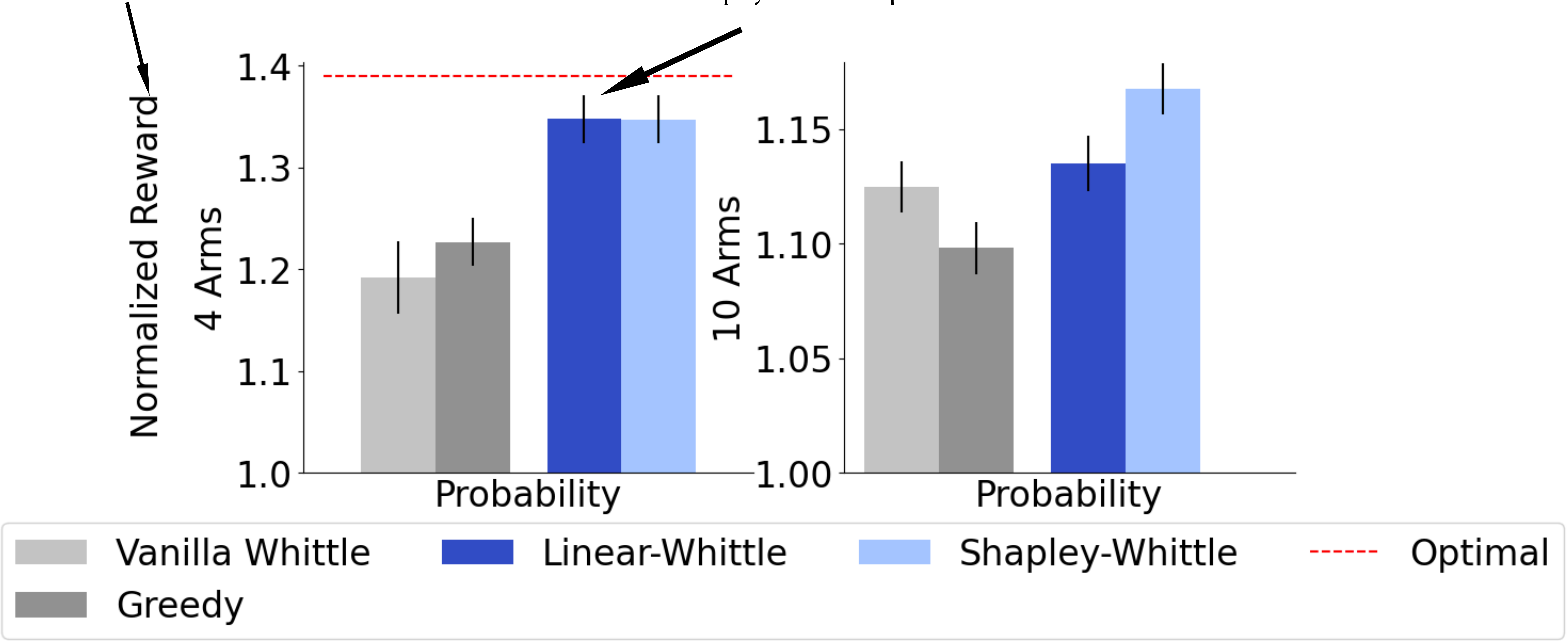
Synthetic Empirical Verification



Synthetic Empirical Verification

Normalized with respect to random baseline

Linear- and Shapley-Whittle outperform baselines



Constructing a Food Rescue Simulation

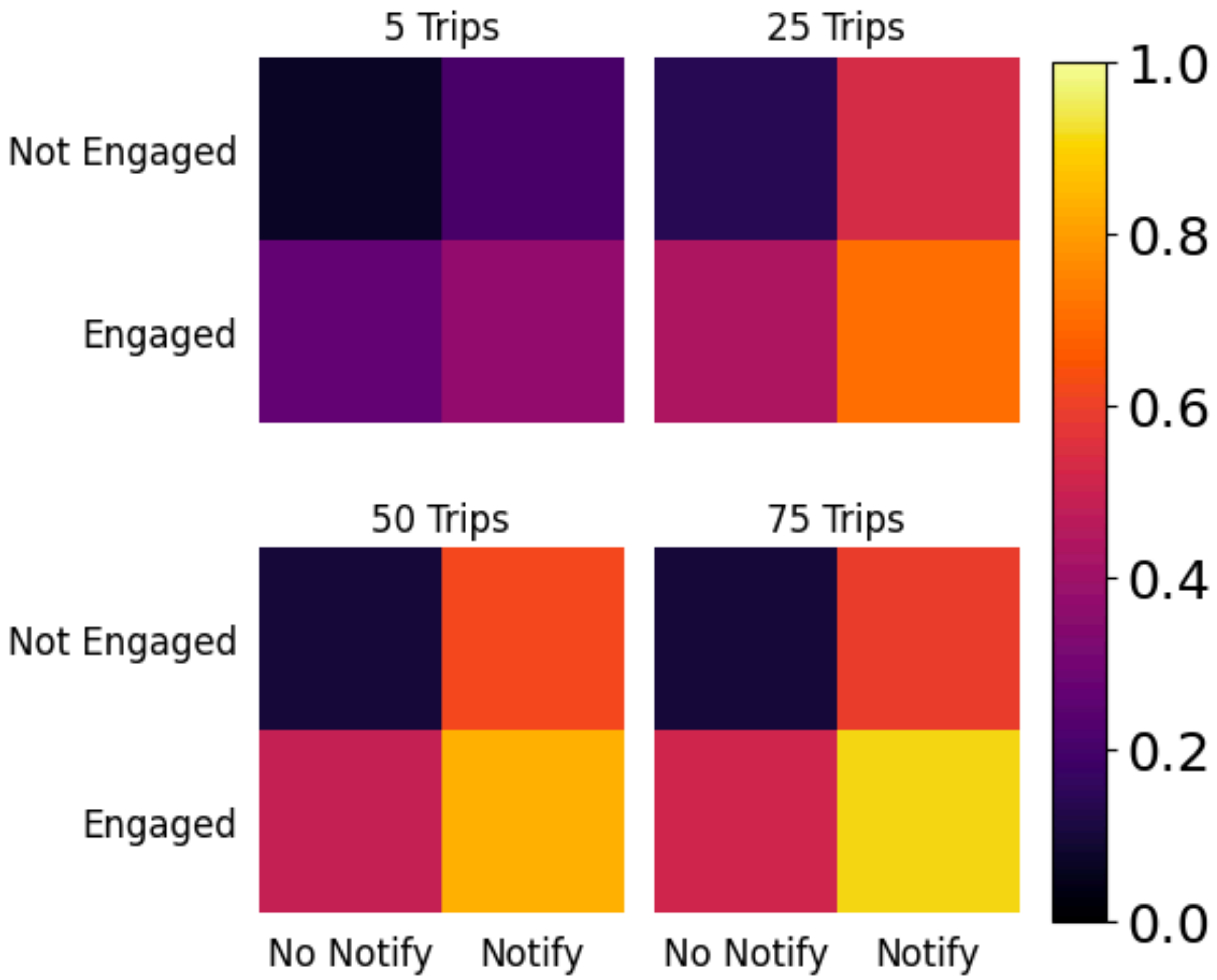
Constructing a Food Rescue Simulation

Engaged: Completed a trip in past 2 weeks
Not Engaged: No Trip Completion over past 2 weeks

States

Constructing a Food Rescue Simulation

Engaged: Completed a trip in past 2 weeks
Not Engaged: No Trip Completion over past 2 weeks

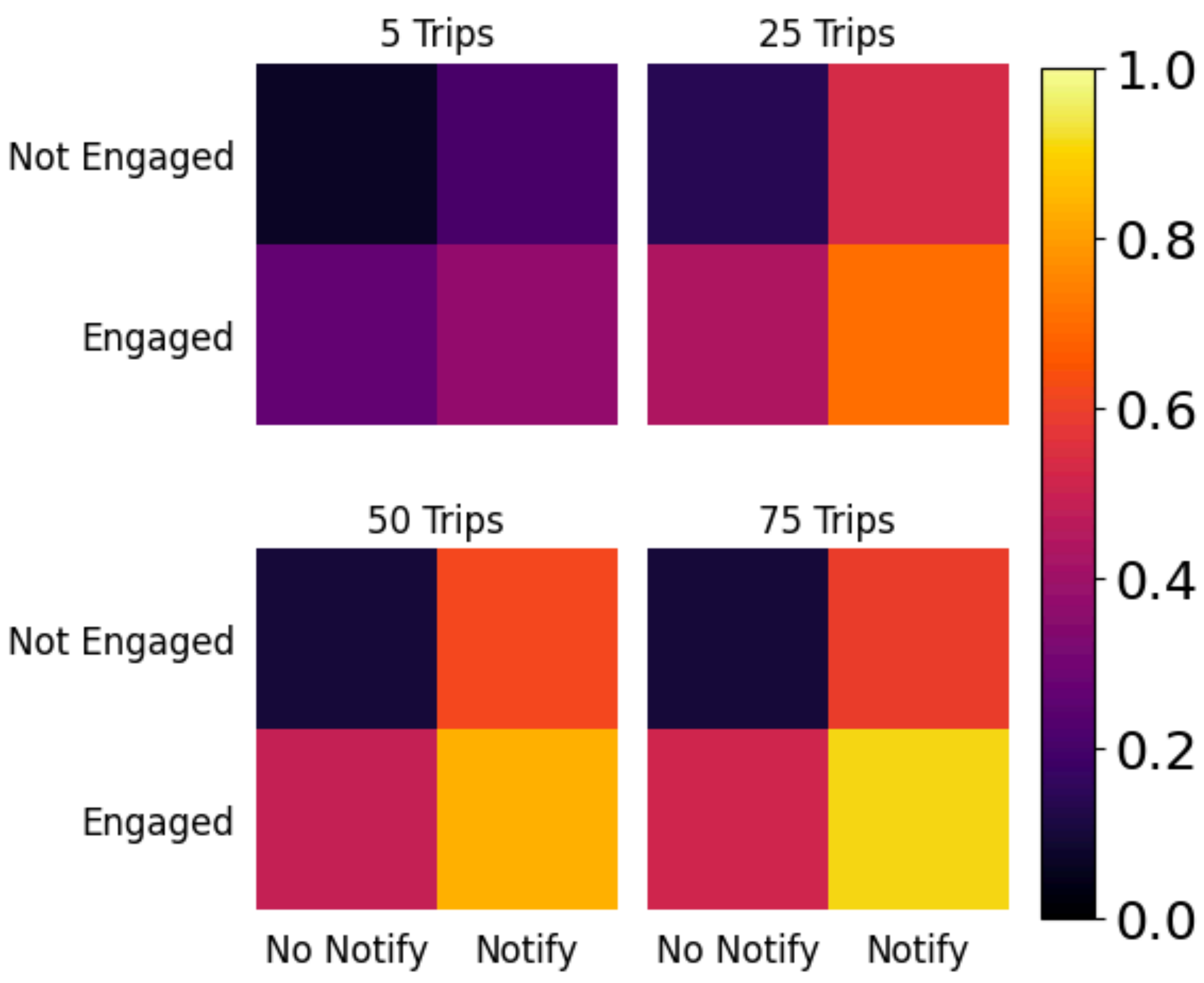


States

Transitions

Constructing a Food Rescue Simulation

Engaged: Completed a trip in past 2 weeks
Not Engaged: No Trip Completion over past 2 weeks



States

Transitions

Learn real transition matrices, states from volunteer data from 412 Food Rescue

Two Food Rescue Settings

Two Food Rescue Settings

Notifications: Volunteers are notified en-masse about rescue
Large Budget (K) and number of volunteers (N), but low match probability

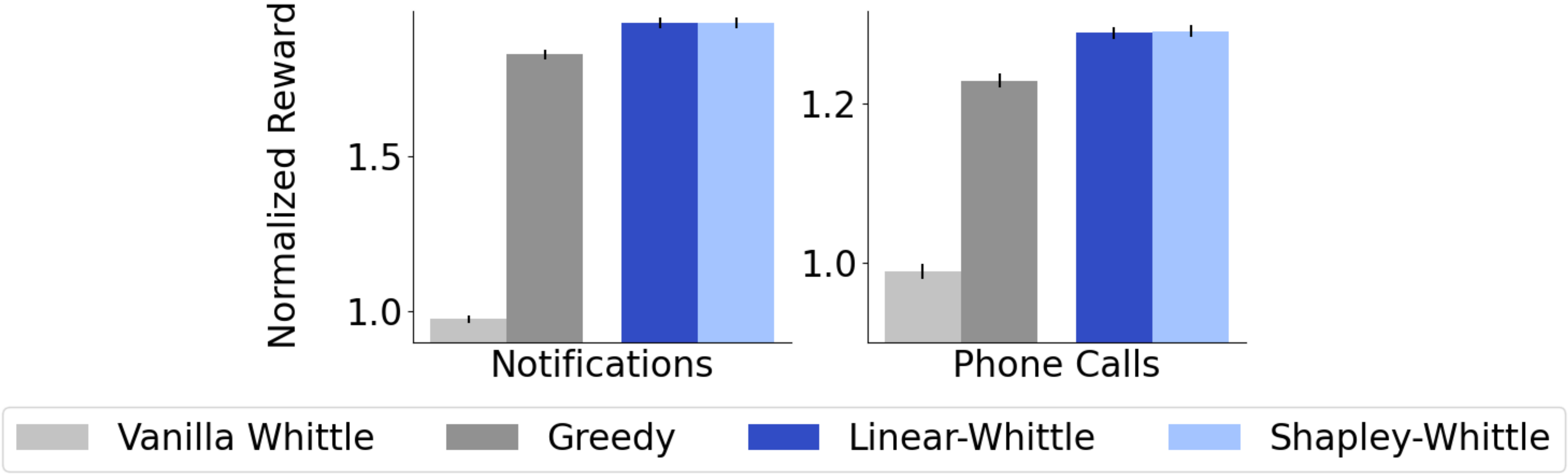
Two Food Rescue Settings

Notifications: Volunteers are notified en-masse about rescue
Large Budget (K) and number of volunteers (N), but low match probability

Phone Calls: Operators manually call top volunteers
Small Budget (K) and number of volunteers (N), but high match probability

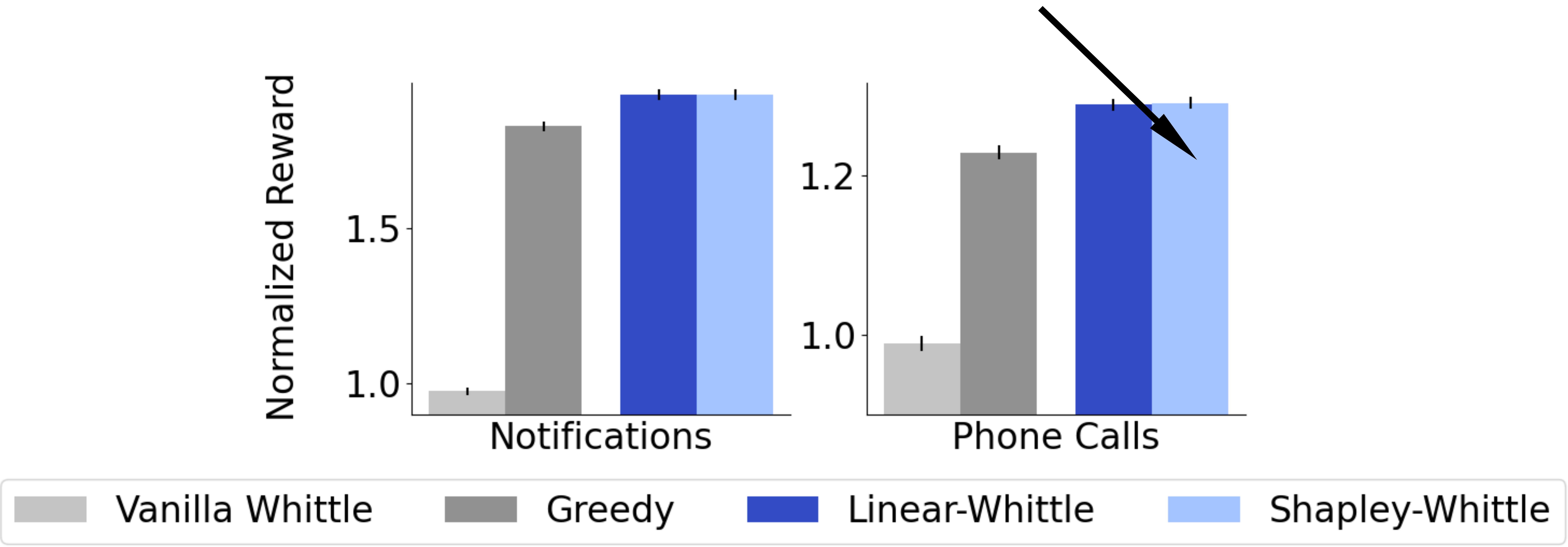
Food Rescue Empirical Verification

Food Rescue Empirical Verification

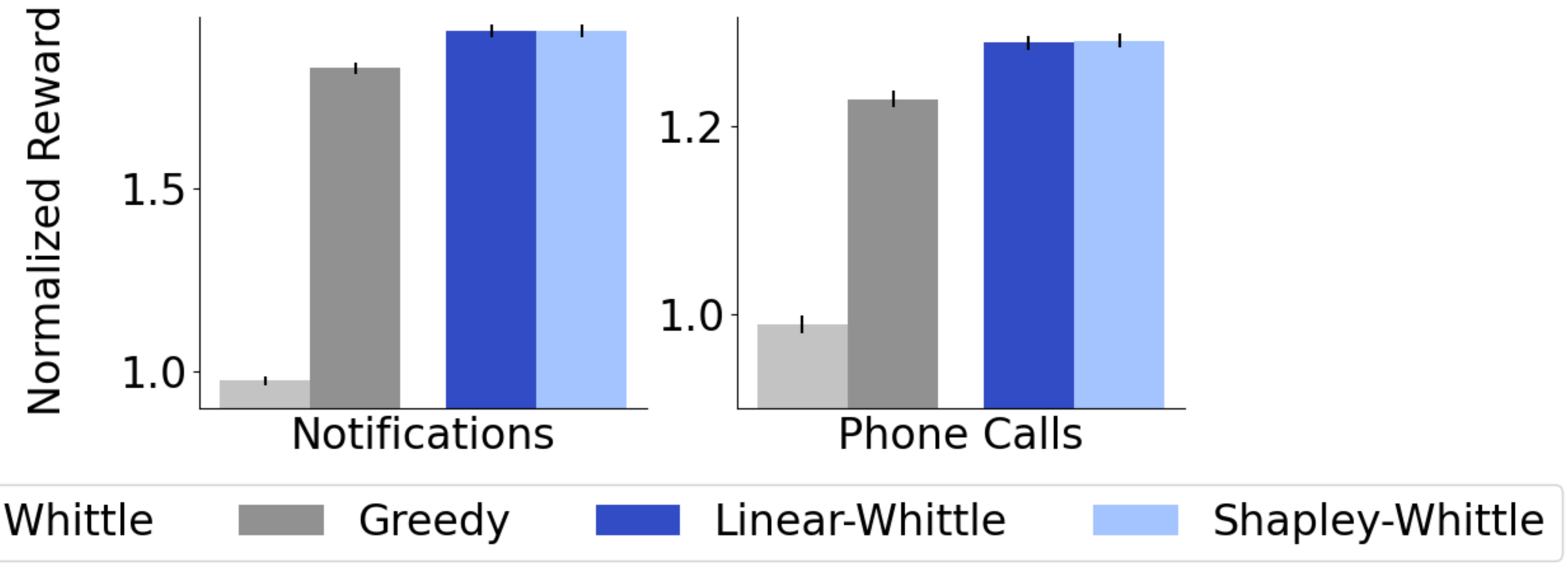


Food Rescue Empirical Verification

Due to reward linearity, Linear- and Shapley-Whittle are similar



Food Rescue Empirical Verification



Adapting to Reward Linearity

Why do we need Adaptivity?

Why do we need Adaptivity?

Using Linear- or Shapley-Whittle indices can lead to poor performance in some scenarios

Why do we need Adaptivity?

Using Linear- or Shapley-Whittle indices can lead to poor performance in some scenarios

Example:
Consider K=N Arms; all arms start in state s=1
Pulling an arms forces it to state s=0, not pulling an arm leaves it state as is
Reward is:

$$R(\mathbf{s}, \mathbf{a}) = \max_i s_i a_i$$

So arms should be pulled separately
However, Linear- and Shapley-Whittle will play all arms simultaneously, leading to $\frac{1}{K}$ of the optimal reward

New Contribution: Two Forms of Adaptivity

New Contribution: Two Forms of Adaptivity

Two new forms of adaptivity that combine with Linear and Shapley-Whittle Indices

New Contribution: Two Forms of Adaptivity

Two new forms of adaptivity that combine with Linear and Shapley-Whittle Indices

Iterative Linear-Whittle: Select arms one-by-one by re-computing Whittle index, based on the arms already pulled

Previously: Marginal Reward for pulling arm 2 is $R(\mathbf{s}, \{0,1,\ldots,0\})$

Now: Reward for pulling arm 2, given arm 1 is pulled, is $R(\mathbf{s}, \{1,1,\ldots,0\}) - R(\mathbf{s}, \{1,0,\ldots,0\})$

New Contribution: Two Forms of Adaptivity

Two new forms of adaptivity that combine with Linear and Shapley-Whittle Indices

Iterative Linear-Whittle: Select arms one-by-one by re-computing Whittle index, based on the arms already pulled

Previously: Marginal Reward for pulling arm 2 is $R(\mathbf{s}, \{0,1,\ldots,0\})$

Now: Reward for pulling arm 2, given arm 1 is pulled, is $R(\mathbf{s}, \{1,1,\ldots,0\}) - R(\mathbf{s}, \{1,0,\ldots,0\})$

MCTS Linear-Whittle: Use Monte-Carlo Tree Search to search for best combination of arms
Compute $R(\mathbf{s}, \mathbf{a})$ for this combination of arms, then estimate future value via Linear-Whittle index

New Contribution: Two Forms of Adaptivity

Two new forms of adaptivity that combine with Linear and Shapley-Whittle Indices

Iterative Linear-Whittle: Select arms one-by-one by re-computing Whittle index, based on the arms already pulled

Previously: Marginal Reward for pulling arm 2 is $R(\mathbf{s}, \{0,1,\ldots,0\})$

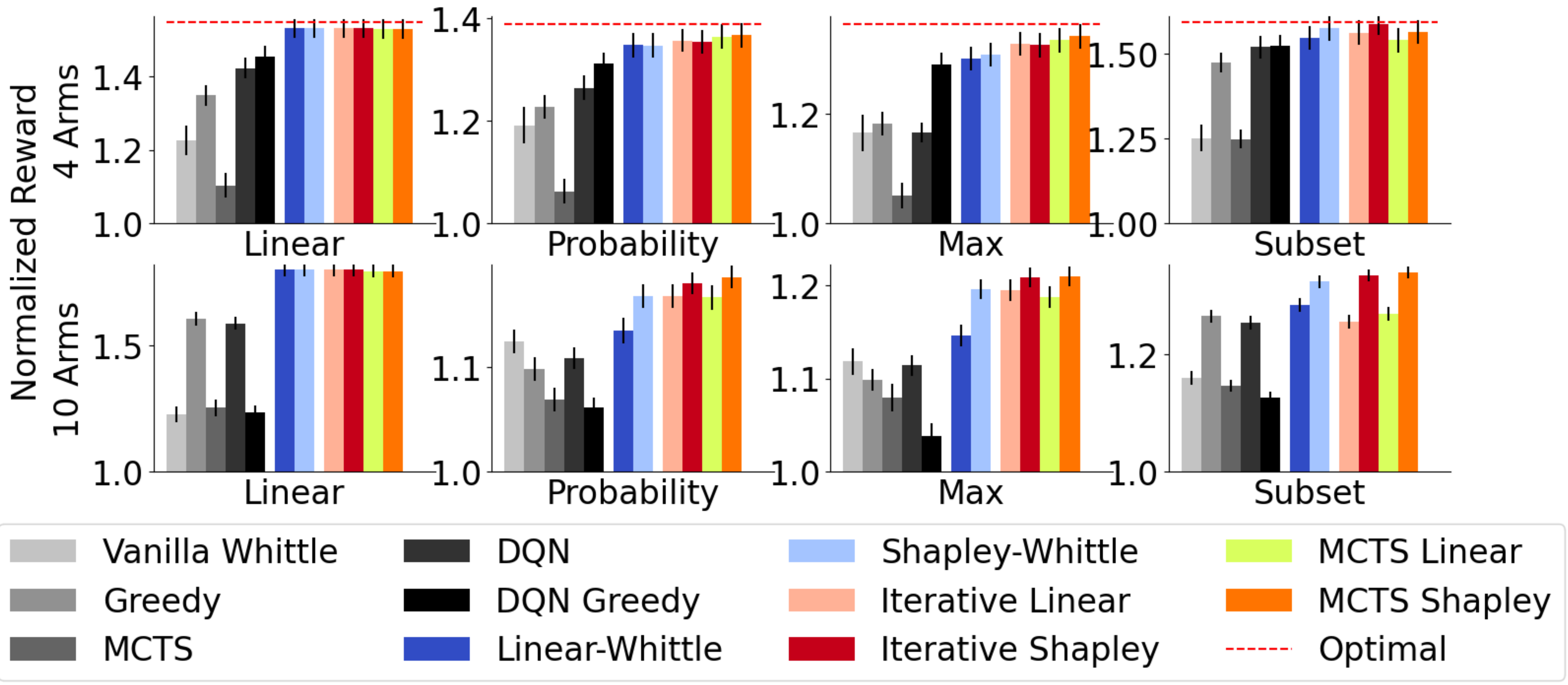
Now: Reward for pulling arm 2, given arm 1 is pulled, is $R(\mathbf{s}, \{1,1,\ldots,0\}) - R(\mathbf{s}, \{1,0,\ldots,0\})$

MCTS Linear-Whittle: Use Monte-Carlo Tree Search to search for best combination of arms
Compute $R(\mathbf{s}, \mathbf{a})$ for this combination of arms, then estimate future value via Linear-Whittle index

Analogous definitions for Shapley-Whittle as well!

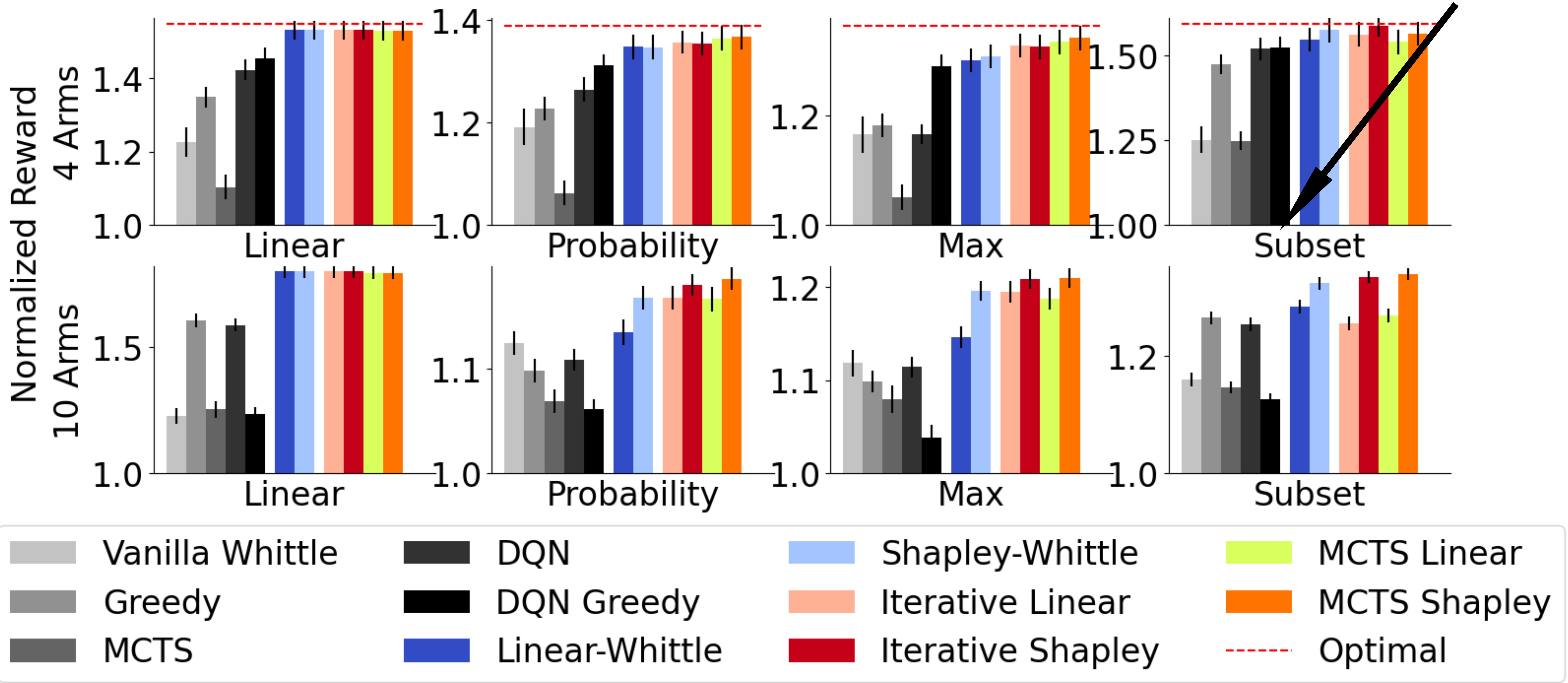
Comparison on Synthetic Data

Comparison on Synthetic Data

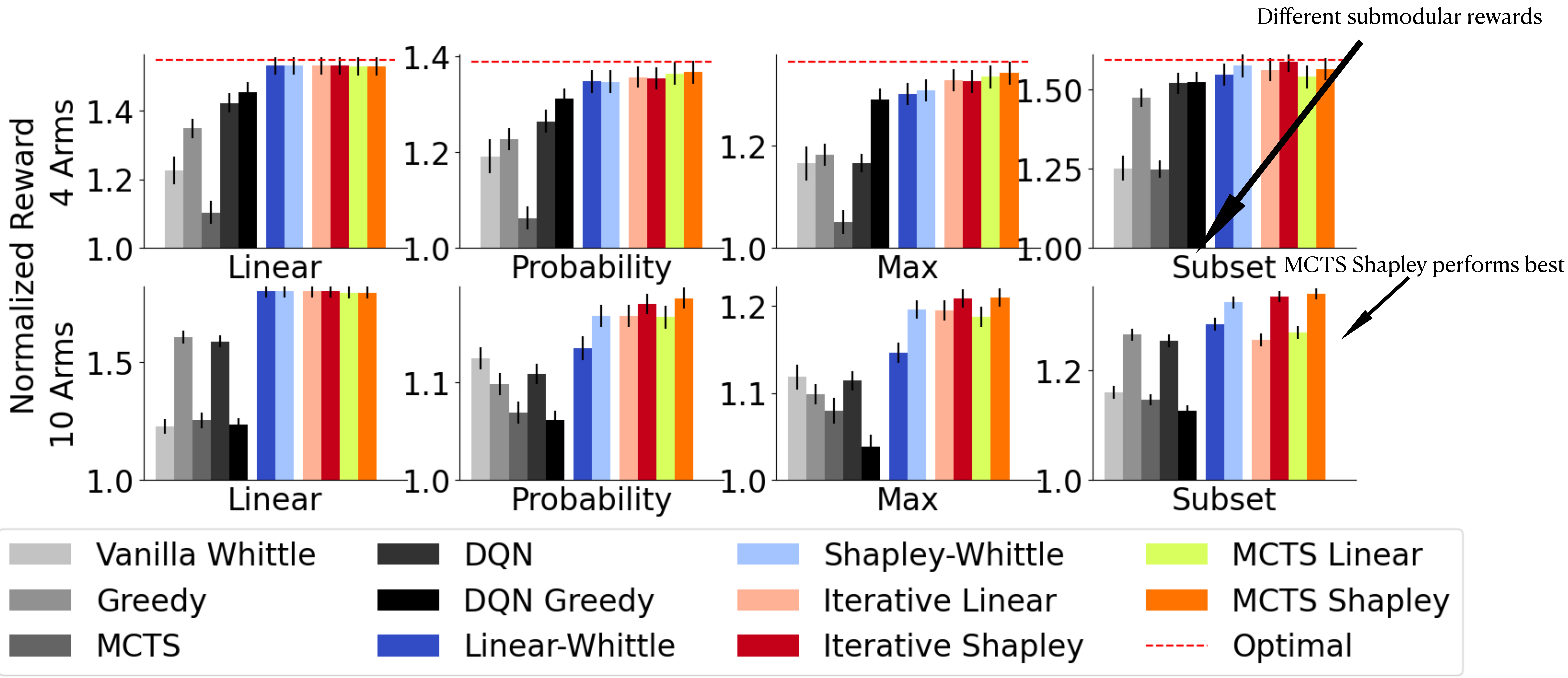


Comparison on Synthetic Data

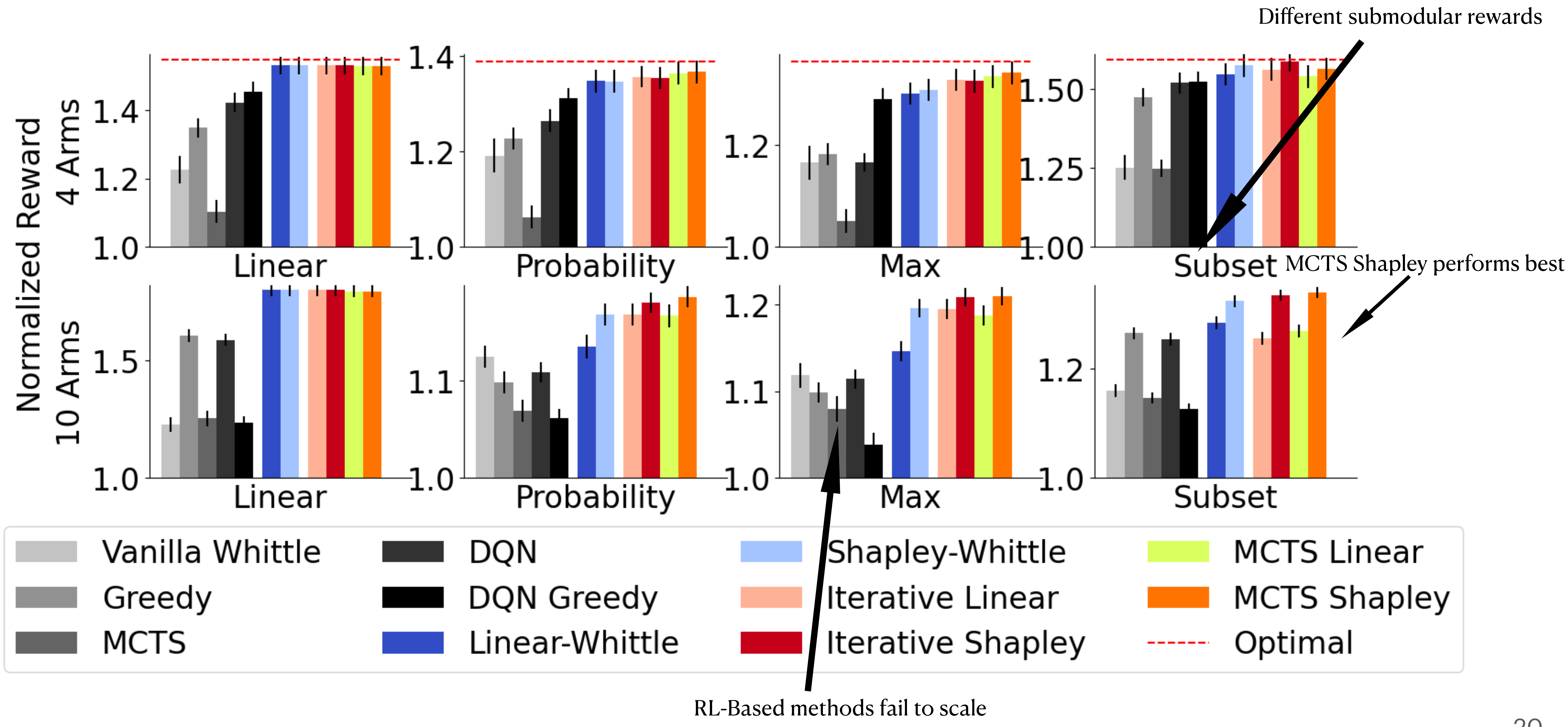
Different submodular rewards



Comparison on Synthetic Data

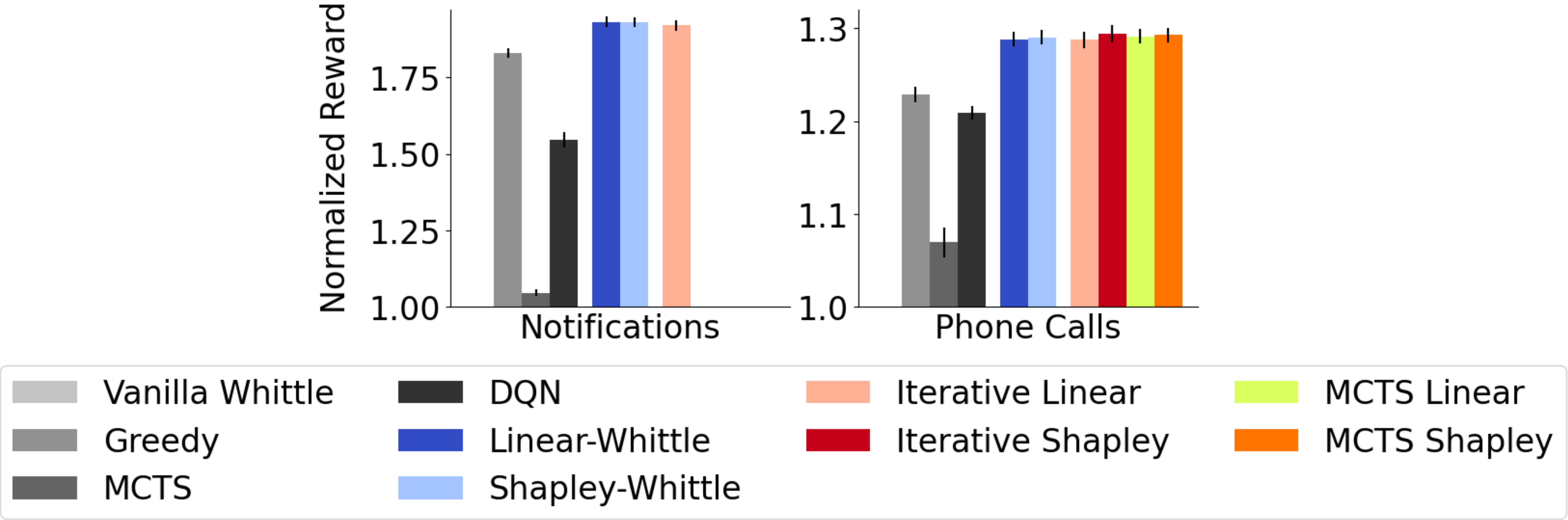


Comparison on Synthetic Data



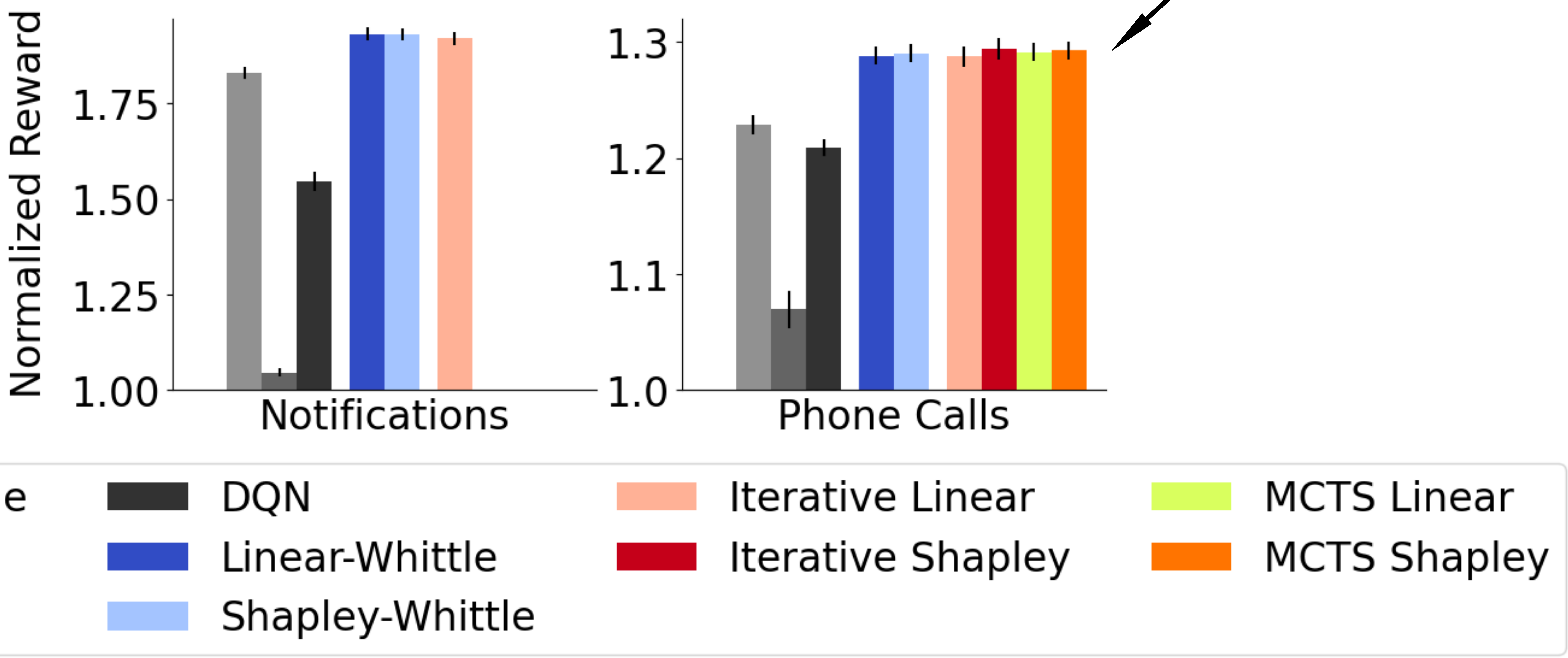
Comparison on Food Rescue

Comparison on Food Rescue



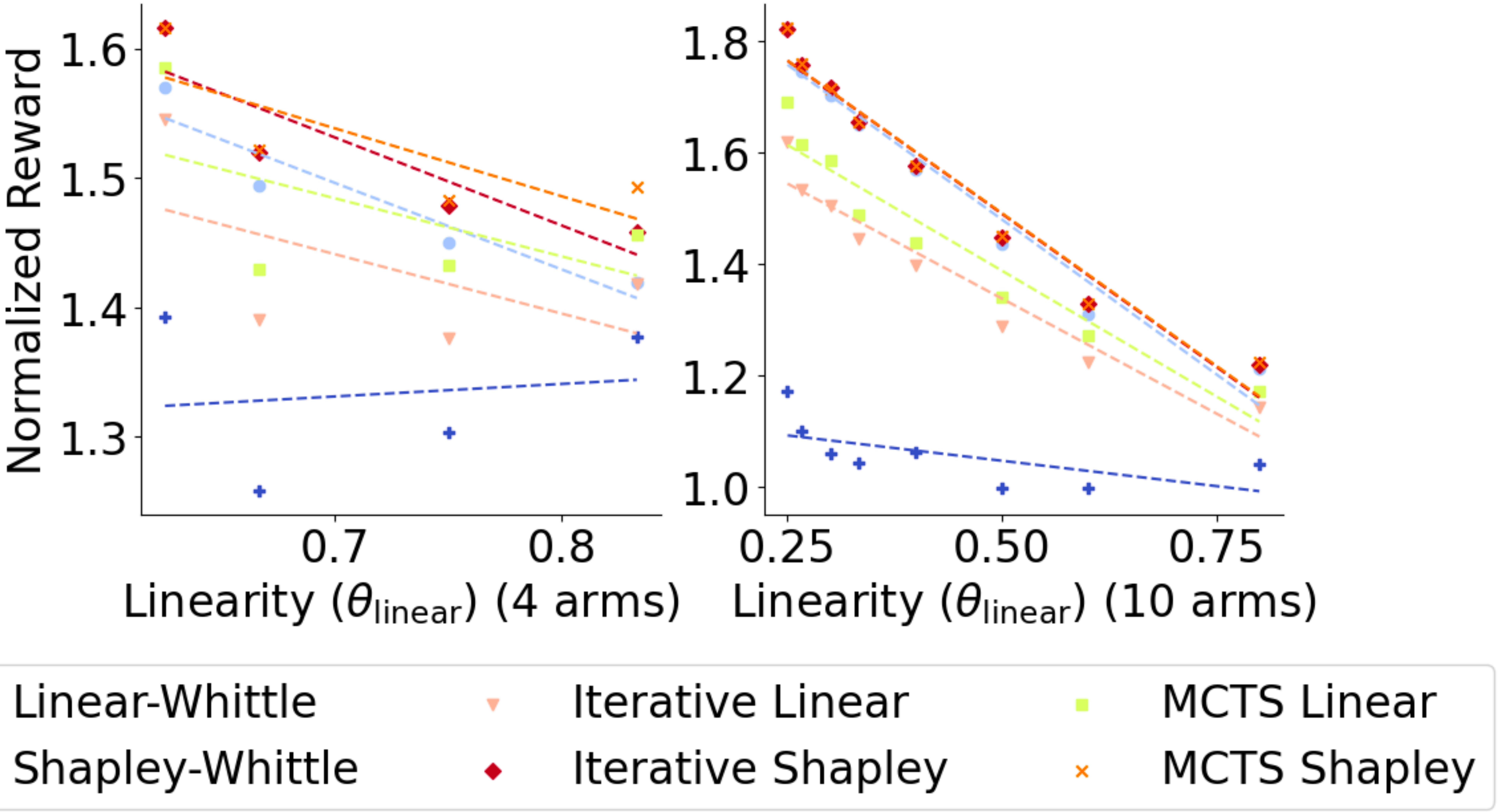
Comparison on Food Rescue

Adaptive Methods are slightly better

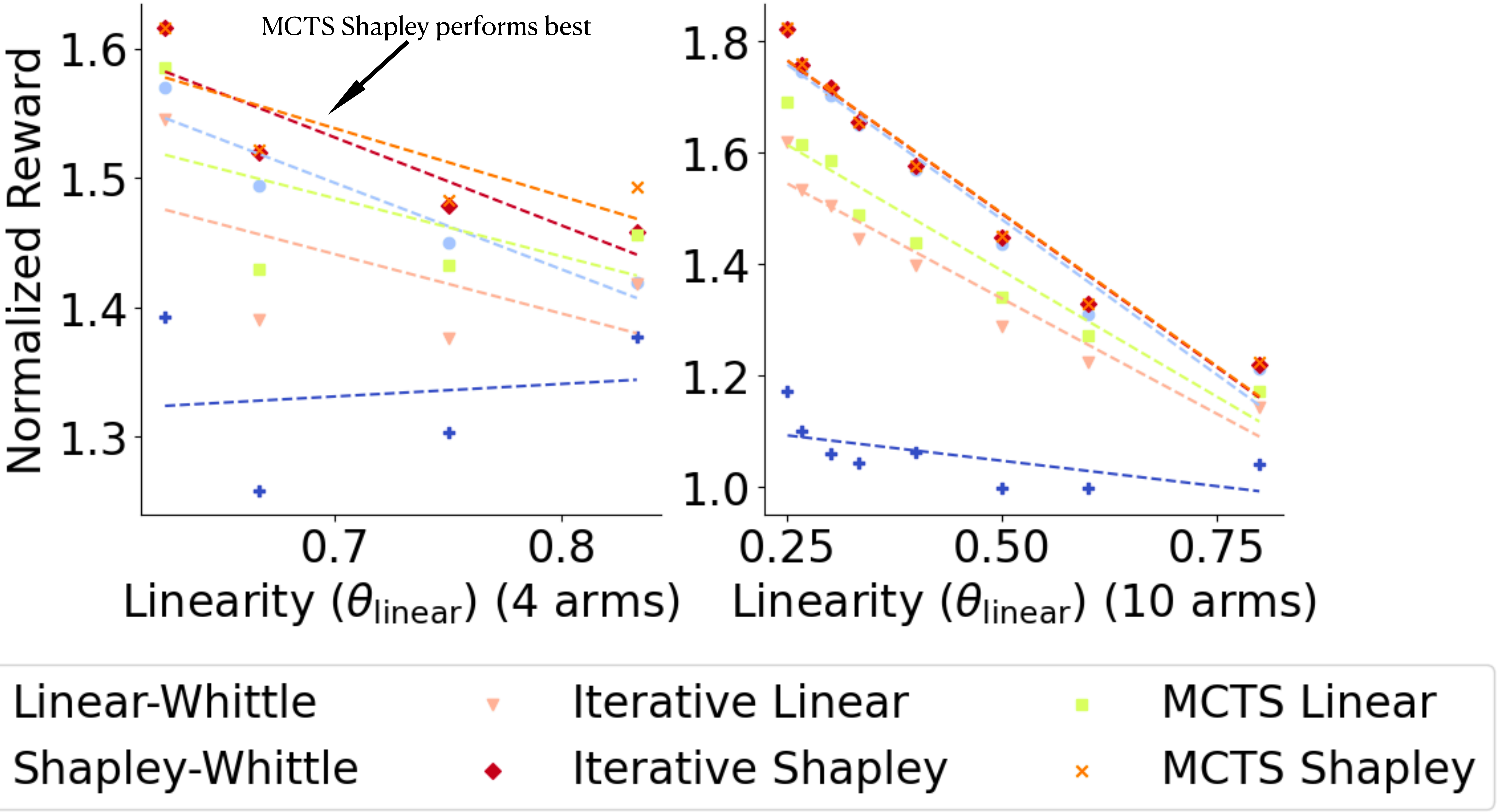


Comparison Across Reward Types

Comparison Across Reward Types



Comparison Across Reward Types



Performance Guarantees

Recall our Goal

Recall our Goal

Food Rescue Optimization

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1 - \prod_{i=1}^{N}(1 - p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]$$

Generalized Problem (RMAB-Global)

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t(R_{\text{glob}}(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) + \sum_{i=1}^{N} R_i(s_i^{(t)}, a_i^{(t)}))]$$

Recall our Goal

Food Rescue Optimization

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t((1 - \prod_{i=1}^{N}(1 - p_i a_i^{(t)} s_i^{(t)})) + \frac{1}{N}\sum_{i=1}^{N} s_i)]$$

Generalized Problem (RMAB-Global)

$$\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim(P,\pi)}[\sum_{t=0}^{\infty} \gamma^t(R_{\text{glob}}(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) + \sum_{i=1}^{N} R_i(s_i^{(t)}, a_i^{(t)}))]$$

How close are our proposed solutions to the optimal π

Lower Bounds and Intuition

Lower Bounds and Intuition

Theorem 1 (informal): Linear-Whittle is a β_{linear} approximation to the RMAB-G problem, where

$$\beta_{\text{linear}} = \min_{\mathbf{s}\in\mathcal{S}^N, \mathbf{a}\in[0,1]^N, \|\mathbf{a}\|_1\leq K} \frac{R(\mathbf{s},\mathbf{a})}{\sum_{i=1}^{N} (R_i(s_i, a_i) + p_i(s_i)a_i)} \geq \frac{1}{K}$$

Lower Bounds and Intuition

Theorem 1 (informal): Linear-Whittle is a β_{linear} approximation to the RMAB-G problem, where

$$\beta_{\text{linear}} = \min_{\mathbf{s}\in\mathcal{S}^N, \mathbf{a}\in[0,1]^N, \|\mathbf{a}\|_1 \leq K} \frac{R(\mathbf{s},\mathbf{a})}{\sum_{i=1}^{N} (R_i(s_i,a_i) + p_i(s_i)a_i)} \geq \frac{1}{K}$$

Linear Approximation of Global Reward

Lower Bounds and Intuition

Theorem 1 (informal): Linear-Whittle is a β_{linear} approximation to the RMAB-G problem, where

$$\beta_{\text{linear}} = \min_{\mathbf{s}\in\mathcal{S}^N, \mathbf{a}\in[0,1]^N, \|\mathbf{a}\|_1 \leq K} \frac{R(\mathbf{s},\mathbf{a})}{\sum_{i=1}^N \left(R_i(s_i,a_i) + p_i(s_i)a_i\right)} \geq \frac{1}{K}$$

Linear Approximation of Global Reward

Intuition: The Linear Approximation to a Submodular Function cannot be very far away from the original function, so perform at least β_{linear} as well as optimal

Upper Bounds and Intuition

Upper Bounds and Intuition

Theorem 2 (informal): For a given reward function, there exists transitions where Linear-Whittle achieves at most a θ_{linear} fraction of optimal reward for the RMAB-G problem, where

$$\theta_{\text{linear}} = \min_{\mathbf{s} \in \mathcal{S}^N} \frac{R(\mathbf{s}, \hat{\mathbf{a}}(\mathbf{s}))}{\max_{\mathbf{a} \in [0,1]^N, \|\mathbf{a}\|_1 \le K} R(\mathbf{s}, \mathbf{a})} \qquad \hat{\mathbf{a}}(\mathbf{s}) = \text{argmax}_{\mathbf{a} \in [0,1]^N, \|\mathbf{a}\|_1 \le K} \sum_{i=1}^{N} (R_i(s_i, a_i) + p_i(s_i)a_i)$$

Upper Bounds and Intuition

Theorem 2 (informal): For a given reward function, there exists transitions where Linear-Whittle achieves at most a θ_{linear} fraction of optimal reward for the RMAB-G problem, where

$$\theta_{\text{linear}} = \min_{\mathbf{s} \in \mathcal{S}^N} \frac{R(\mathbf{s}, \hat{\mathbf{a}}(\mathbf{s}))}{\max_{\mathbf{a} \in [0,1]^N, \|\mathbf{a}\|_1 \leq K} R(\mathbf{s}, \mathbf{a})} \qquad \hat{\mathbf{a}}(\mathbf{s}) = \text{argmax}_{\mathbf{a} \in [0,1]^N, \|\mathbf{a}\|_1 \leq K} \sum_{i=1}^{N} (R_i(s_i, a_i) + p_i(s_i) a_i)$$

Intuition: Even in the absence of stochasticity, submodular functions cannot be optimized perfectly, and so any policy is an imperfect approximation

Applications and Open Questions

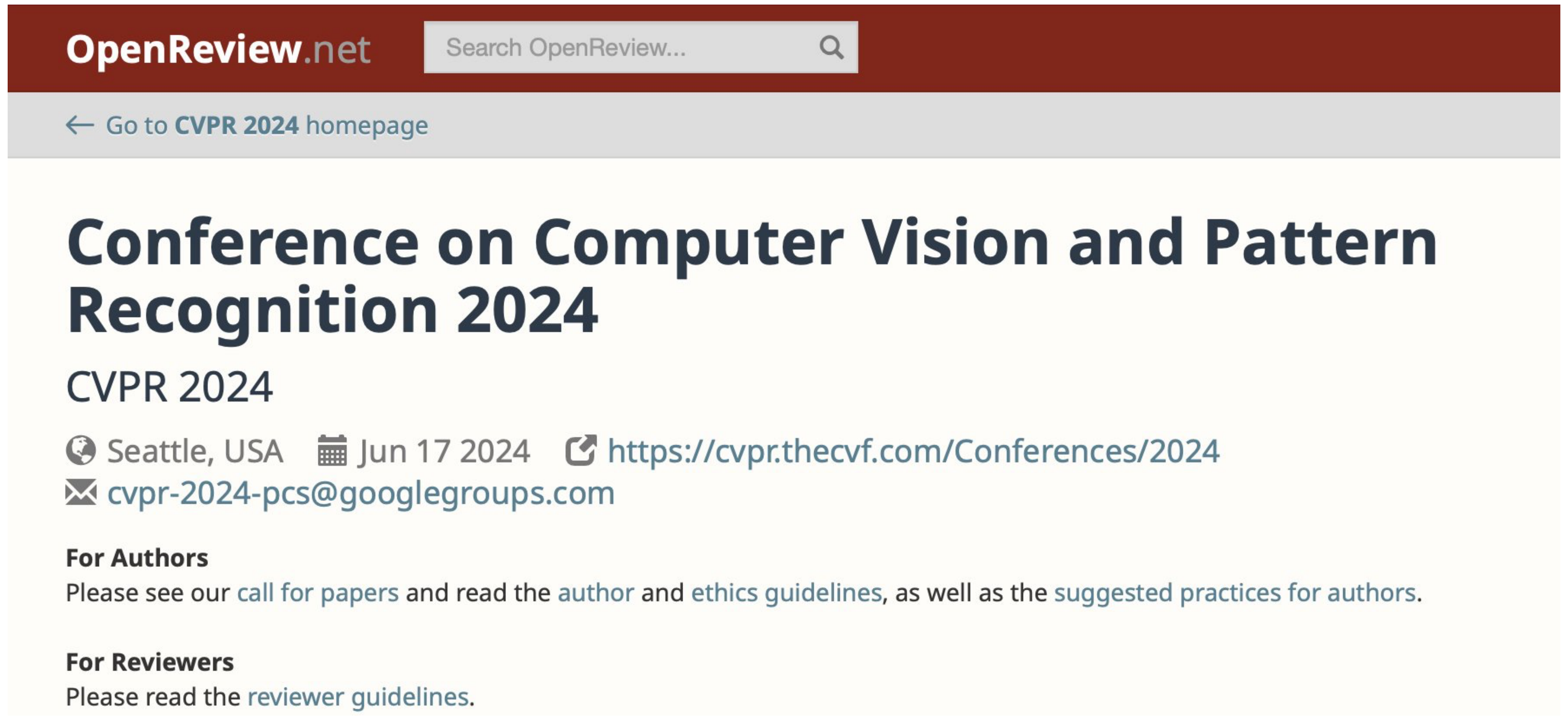
Other Applications

Other Applications

Volunteer Emergency Dispatch
Volunteers transition between availabilities + engagement, and emergency trips arrive online

Other Applications

Volunteer Emergency Dispatch
Volunteers transition between availabilities + engagement, and emergency trips arrive online



Peer Review
Reviewers transition in availability and new papers arrive online and need to be reviewed

Open + Future Questions

Open + Future Questions

What happens if volunteer match probabilities change over time or are **contextual** (e.g. dependent on trip location)?

Open + Future Questions

What happens if volunteer match probabilities change over time or are **contextual** (e.g. dependent on trip location)?

How can we model the **global** nature of matching; the fact that only one individual can actually match at any timestep?

Open + Future Questions

What happens if volunteer match probabilities change over time or are **contextual** (e.g. dependent on trip location)?

How can we model the **global** nature of matching; the fact that only one individual can actually match at any timestep?

What happens if reward parameters or functions are **unknown** and need to be learned?

Conclusion/Recap

Problem: How can we notify volunteers in food rescue with global rewards in a Restless Bandit scenario?

Solution 1: Linearize the global reward as a sum of local linear rewards using Shapley values

Solution 2: Improve on this by making linear approximations adaptive or iterative, essentially incorporating search techniques