# Reinforcement learning with combinatorial actions for coupled restless bandits


**Lily Xu**

UNIVERSITY OF OXFORD

Bryan Wilder

*CMU*

Elias B. Khalil

*U. Toronto*

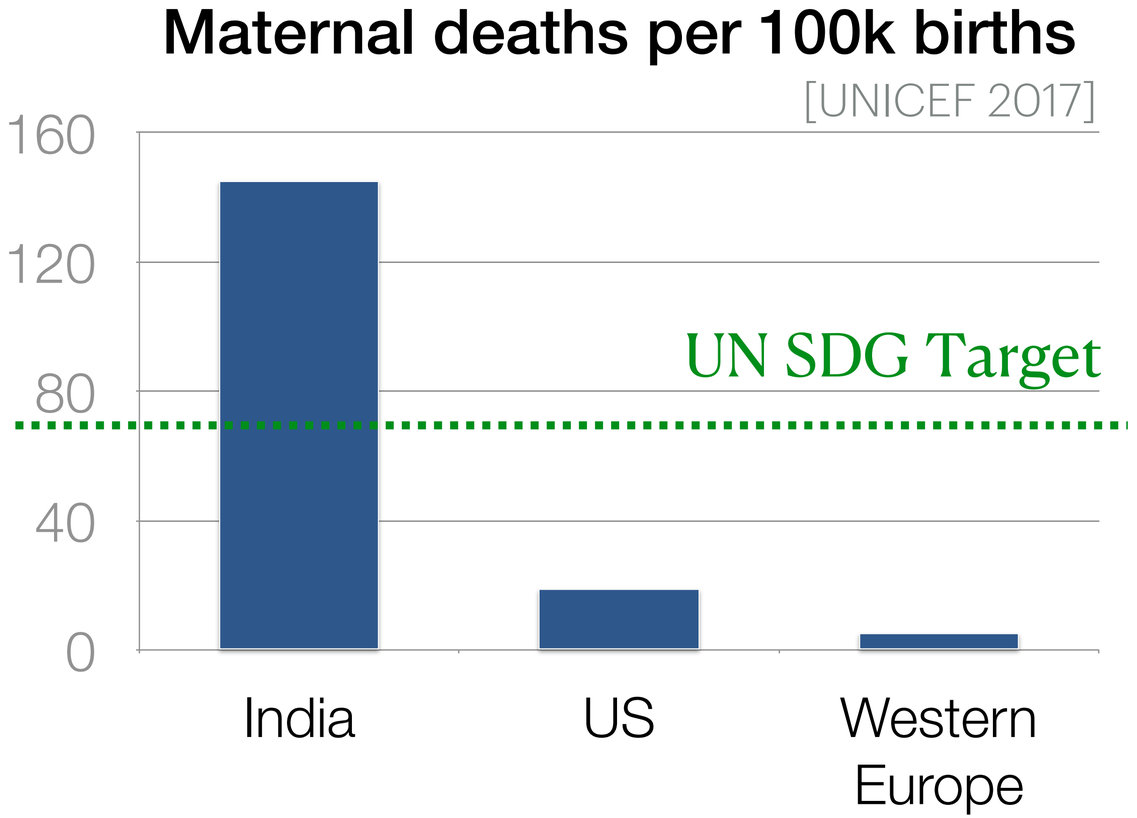
Milind Tambe

*Harvard*


*Fall 2025: Columbia IEOR*

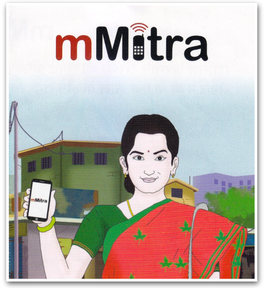
# Maternal mortality crisis

Photo by Save the Children / Flickr creative commons




Many deaths are preventable

# Intervention planning for maternal health




Weekly 2-minute automated voice messages

2.2 million beneficiaries

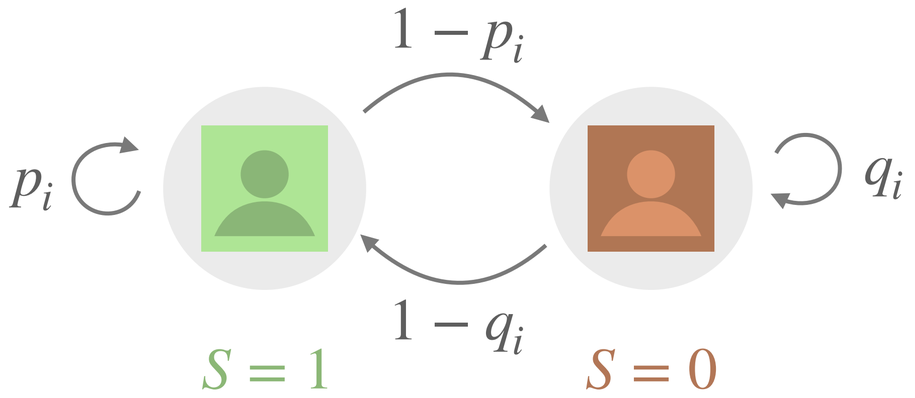
38% of mothers stop listening to messages

*Intervention:* Individualized service calls

Photo from ARMMAN

# Restless bandits for maternal intervention

## Model engagement as an MDP




for $i \in [N]$

**How should we intervene?**

Planning problem alone PSPACE hard

[Papadimitriou and Tsitsiklis 1994]

# Thinking critically about non-engagement

## How should we intervene?

Standard restless bandit: Call $K$ of $N$ mothers each week

Photo from ARMMAN

## Some reasons for non-engagement:

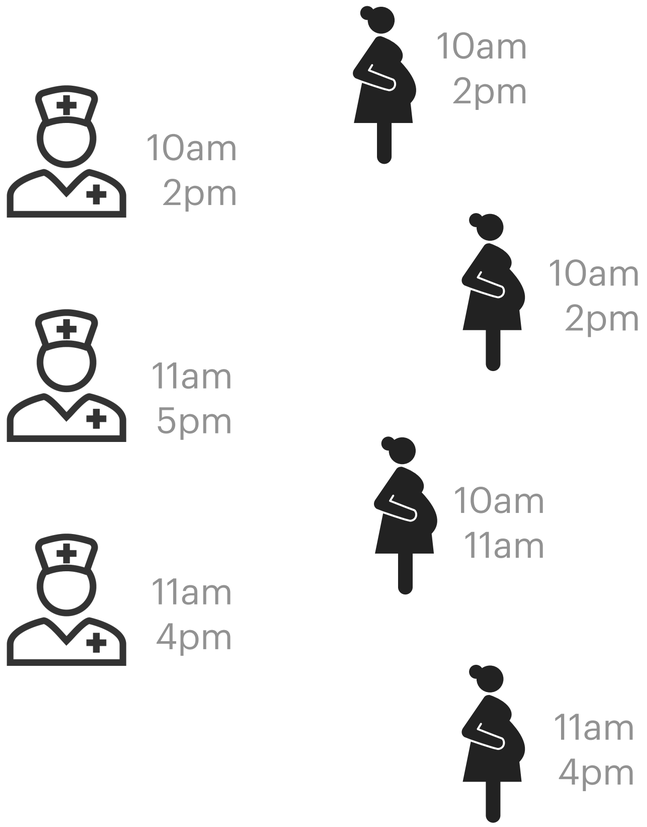
"Time didn't work for me"

"I wanted to talk to someone I was familiar with"

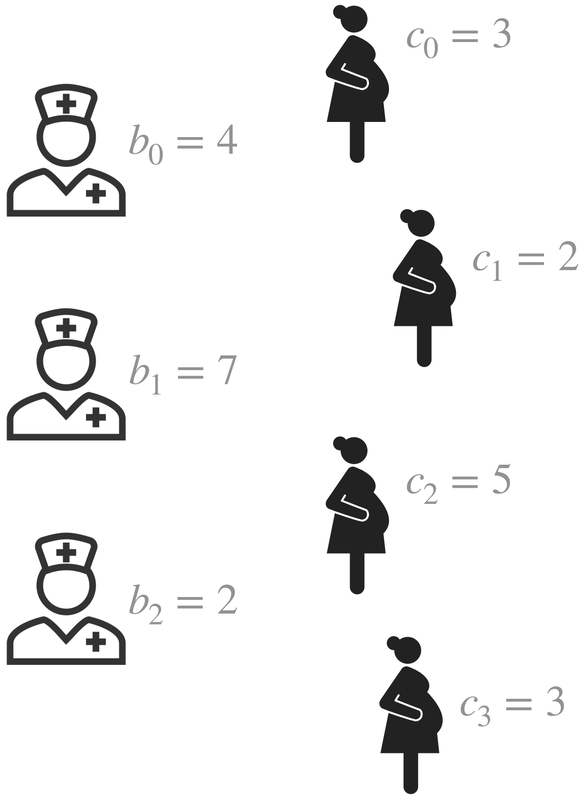
"I share the phone with my husband"

# Interventions beyond “top K”

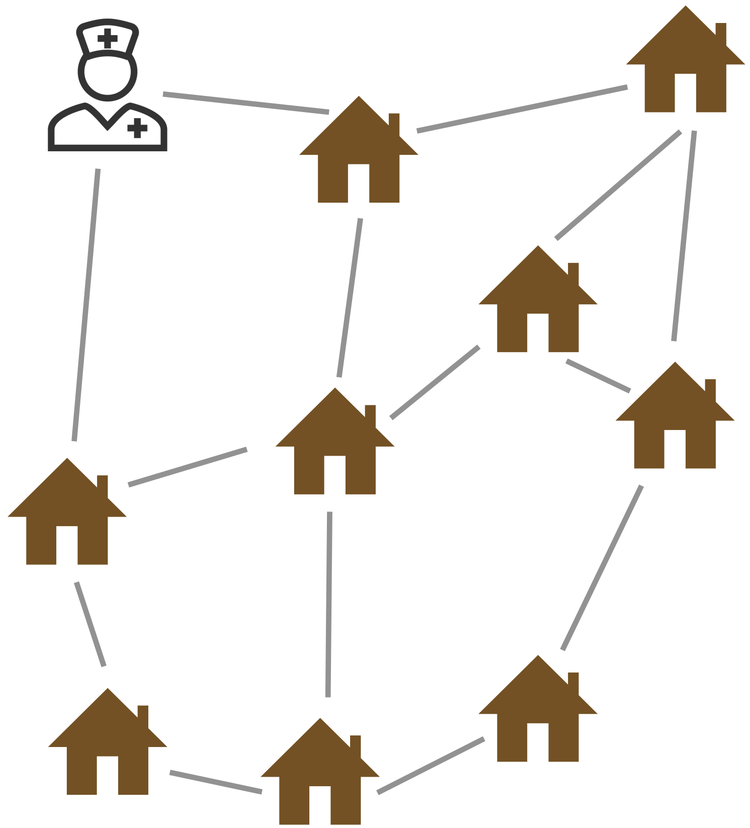
## Scheduling problem




## Capacity constrained

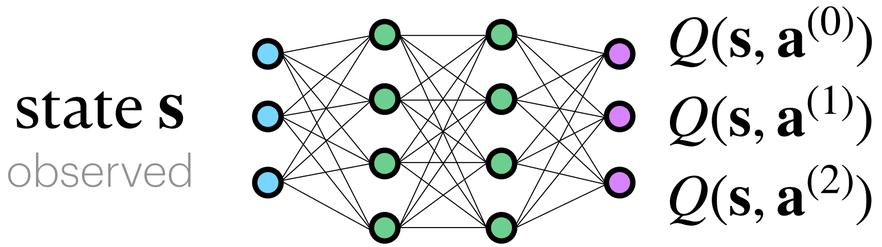



## Routing problem



# RL to solve sequential problems

DQN estimates the long-run value of action **a** from state **s**




But our actions are NP-hard discrete optimization problems

# Existing approaches in RL + combinatorial optimization

RL as heuristic solver for one-step combinatorial optimization

TSP [Dai et al. 2017]; max cut [Barrett et al. 2020]

RL as subset-selection heuristic solver for combinatorial optimization

capacitated VRP [Delarue et al. 2020]

RL with combinatorial action spaces

AlphaGo [Silver et al. 2016]; sampling approach [He et al. 2016]; tabular state space [Brantley et al. 2020]; linear approximation for MARL [Tkachuk et al. 2023]

# Integrating deep learning with MILPs

Neural networks are defined as a series of linear inequalities

$$x = \text{ReLU}(w^\top y + b)$$

Fischetti & Jo [2018] show that these neural networks can be expressed as a MILP:

$$w^\top y + b = x - s$$
$$x \geq 0,\ s \geq 0$$
$$z = 1 \implies x \leq 0,\ z = 0 \implies s \leq 0$$
$$z \in \{0,1\}$$

using a total of $O(DP)$ binary variables and linear constraints [Huchette et al. 2023]

# Solving RL with combinatorial action constraints

**SEQUOIA:** SEQUential cOmbinatorIal Actions

Combining deep RL with mixed-integer programming

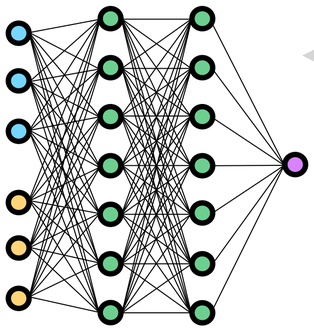
## 1. Training

Learn Q-function approximator

state **s**

action **a**

batched samples from memory



estimated Q-value $Q(\mathbf{s}, \mathbf{a})$

used to generate on-policy trajectories

embed trained deep Q-network into MILP

## 2. Evaluation

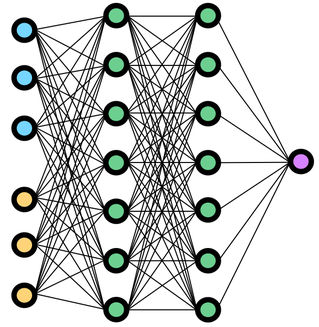
Solve MILP to find best action

state **s**

observed

action **a**

decision variables $a_1, \ldots, a_N$



estimated Q-value $Q(\mathbf{s}, \mathbf{a})$

decision variables from MILP are inputs into NN

use Q-value (from NN) as objective

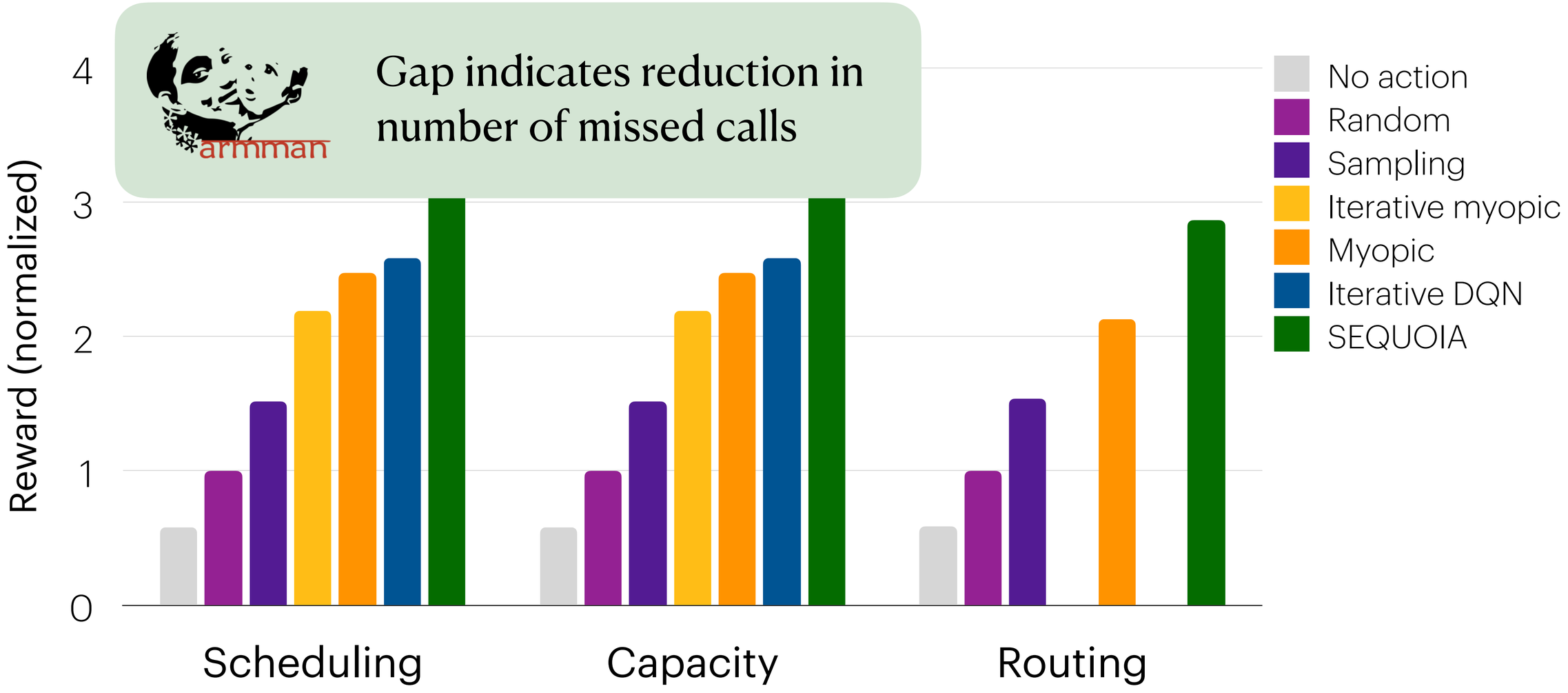
$\max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

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

combinatorial constraints on **a**

constraints and weights $\theta$ to embed NN

# Results

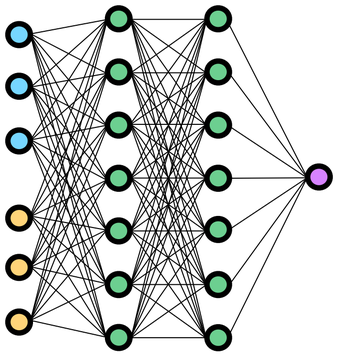

armman
Gap indicates reduction in number of missed calls
Reward (normalized)
0
1
2
3
4
Scheduling
Capacity
Routing
No action
Random
Sampling
Iterative myopic
Myopic
Iterative DQN
SEQUOIA

# Improving maternal health interventions with combinatorial restless bandits

*with Bryan Wilder, Elias B. Khalil, and Milind Tambe*

lily-x.github.io
@lilyxu0
@lilyxu

**Lily Xu**
Oxford / Columbia



Deep reinforcement learning for sequential planning

+

$$\max_{\mathbf{a}} \; Q(\mathbf{s}, \mathbf{a})$$
$$\text{s.t.} \quad \mathbf{a} \in \{0,1\}^N$$

combinatorial constraints on $\mathbf{a}$

constraints and weights $\theta$ to embed NN

Mixed-integer programming for combinatorial actions

**Combining deep RL with mixed-integer programming to solve RL with combinatorial actions**