Concurrent Individual and Social Learning

Justin Girard
justingirard@paxculturastudios.com

CTO at FleetOps
Intelligent Freight Logistics

Founder at Pax Cultura Studios
Intellectual-Capital Management and Data Product Innovation

B.A.Sc - Software Engineering
M.A.Sc - Aerospace Engineering - Robotics Knowledge Representation


EXPLORE A
PROCEDURALLY
GENERATED
GALAXY.

SIX IS A SMART SIGNAL LIGHT.

TURN SIGNALLING

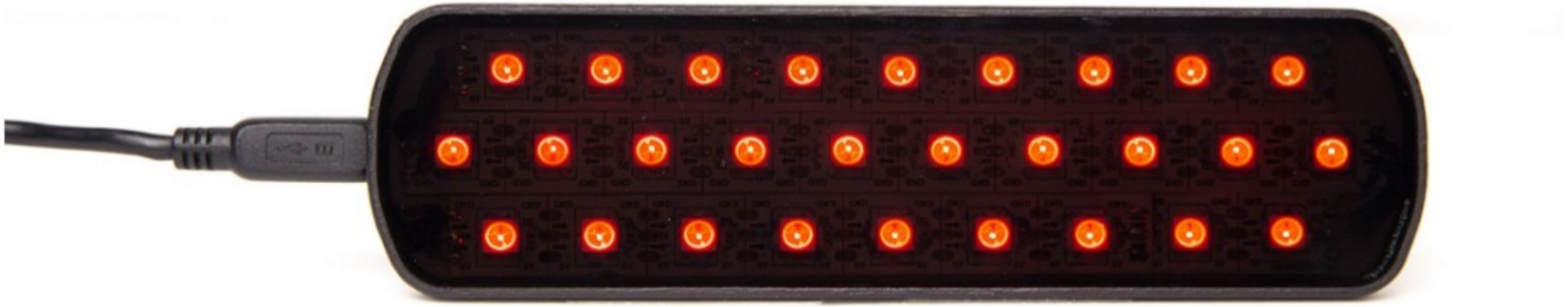
Moving your arm triggers a turn signal on SIX, indicating which direction you intend on turning with bright LEDs.

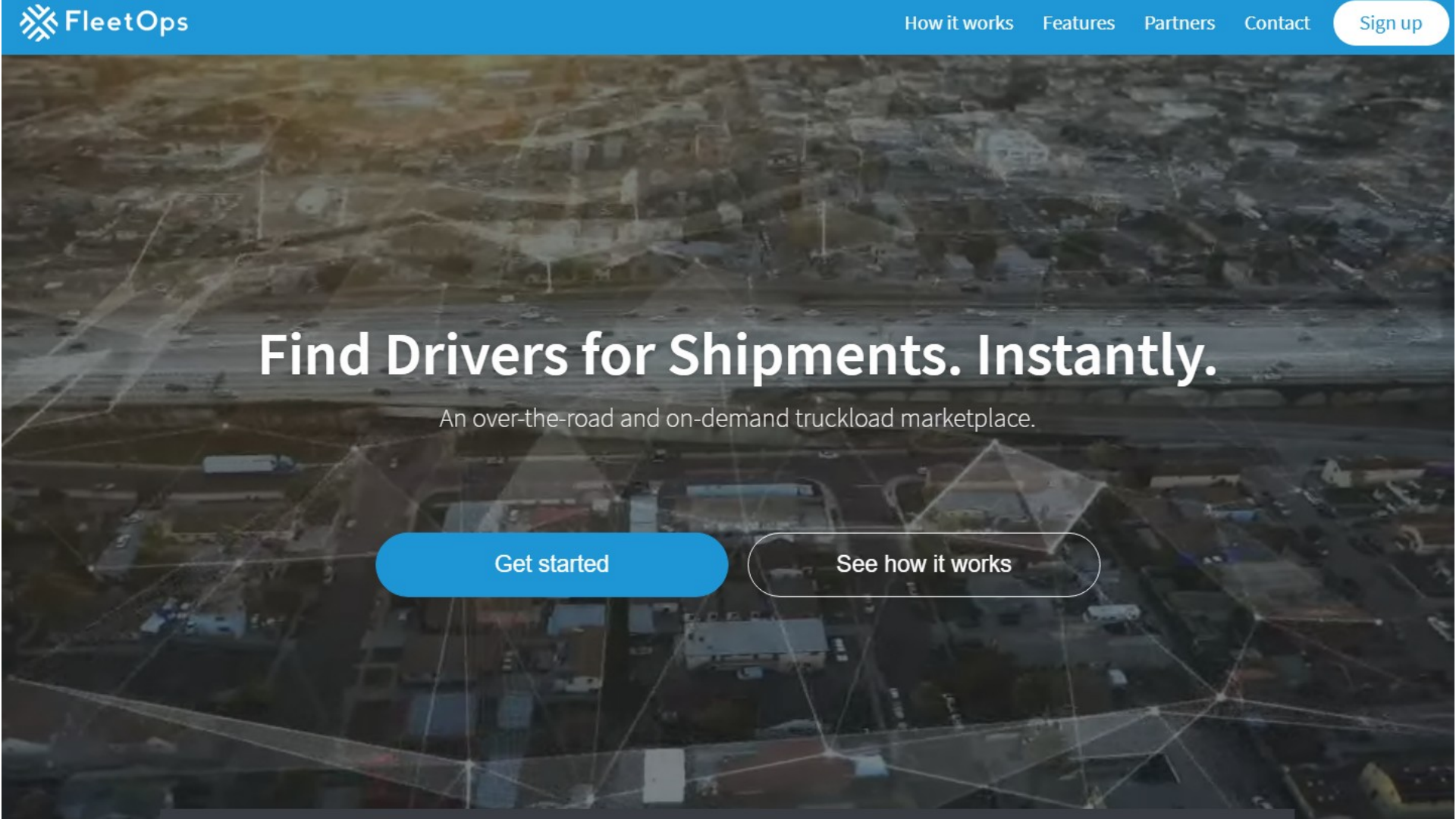
BRAKE SIGNALLING

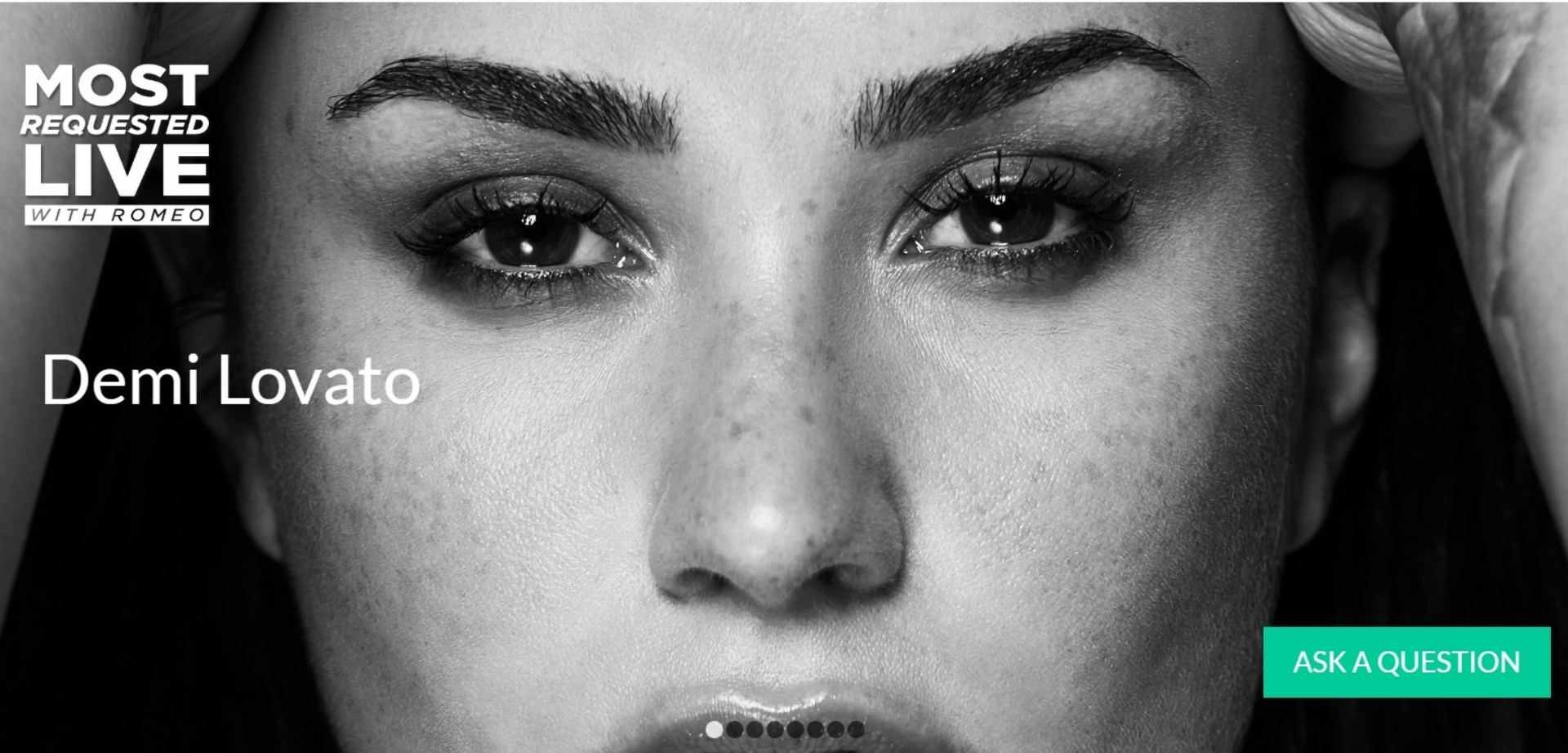
SIX automatically detects when you are braking, and signals to vehicles behind you that you are slowing down. It can be clipped to a backpack, a bike, or even the shoulder of your riding jacket.

A SMART DEVICE

SIX pairs with your phone and wearable, and alerts you of battery loss or disconnection. Never lose another bike light.




FleetOps
How it works
Features
Partners
Contact
Sign up
Find Drivers for Shipments. Instantly.
An over-the-road and on-demand truckload marketplace.
Get started
See how it works

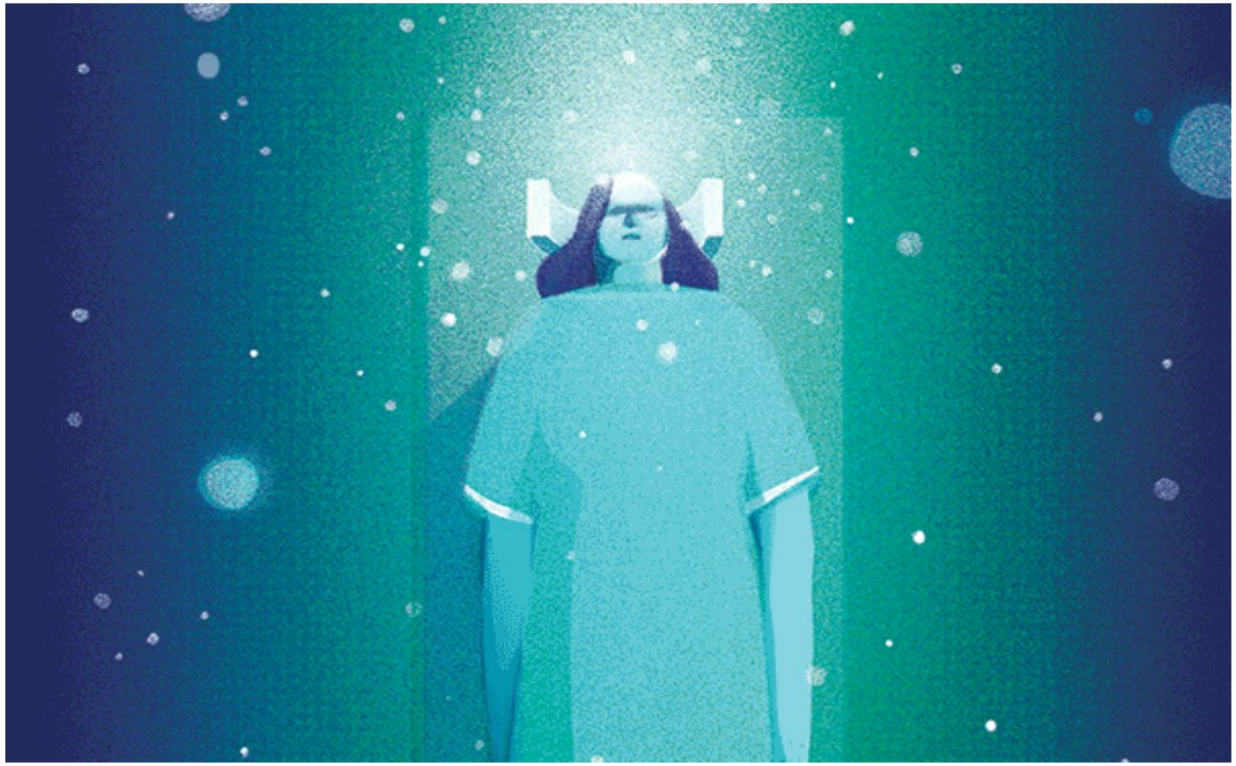

MOST
REQUESTED
LIVE
WITH ROMEO
Demi Lovato
ASK A QUESTION

MEMOIR

How I Found My Way After an MS Diagnosis at Twenty-Seven

Socrates and Plato helped me realize that charting the future with a chronic disease means embracing ambiguity

BY MEREDITH WHITE

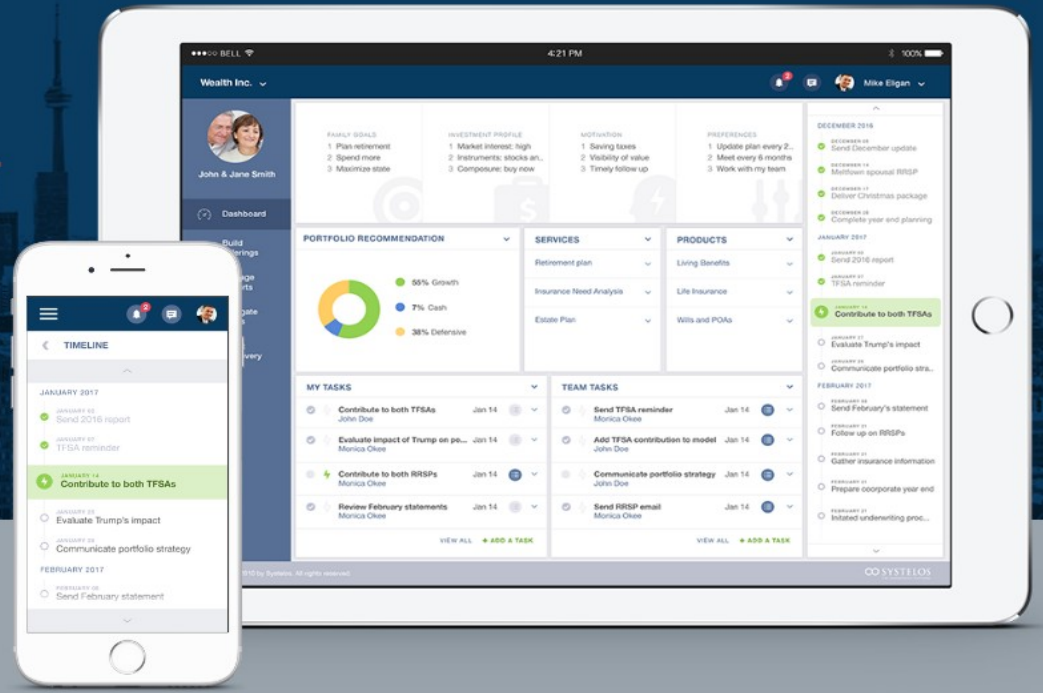


SYSTELOS
THE EMPOWERING PLATFORM

Take a Tour | We're Hiring | Login | Join waiting list →

The wealth management world is changing and this could mean more for everyone.

You will get better financial results with ease



We built a platform that automates and improves how wealth is

ASSETS

24

20 enroute

4 idle

2 alert

DRIVERS

43

3 drivers violating HoS

6 drivers nearing HoS violations

Best: Chris A • Joshua S • Casiino P
Worst: Karl F • Naguesh P • Aprup S

POINTS OF INTEREST

20

3 assets in client hubs

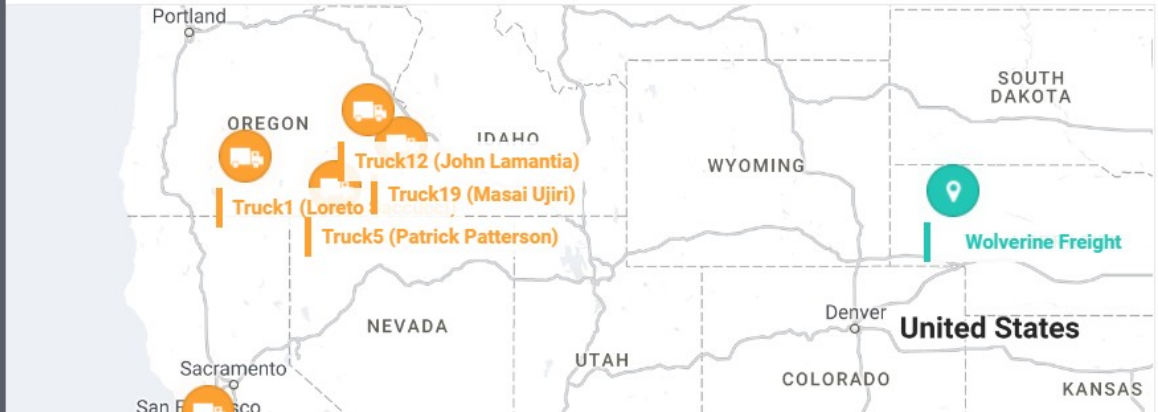
4 assets in fuel / repair stations

12 assets in other POIs

1024
miles covered today • avg 952

12
Idle hours today • avg 10

All (44) | Alerts (2) | Favorites (2) | Assets | Drivers | POIs | Search ...



Wolverine 02
90 mph at Menlo Park, CA

Magneto 07
55 mph at Santa Cruz, CA

Deadpool 05
Stationary at Houston, TX

DCCPER HQ, CA
Yard • No assets

Concurrent Individual and Social Learning

Justin Girard
justingirard@paxculturastudios.com

Idea

People learn from their world individually

Groups need to coordinate work assignments efficiently.

This causes conflict

Concurrent Individual and Social Learning

Individual: Individual processes learn from their environment

Social: Individuals learn how to allocate effort with each other

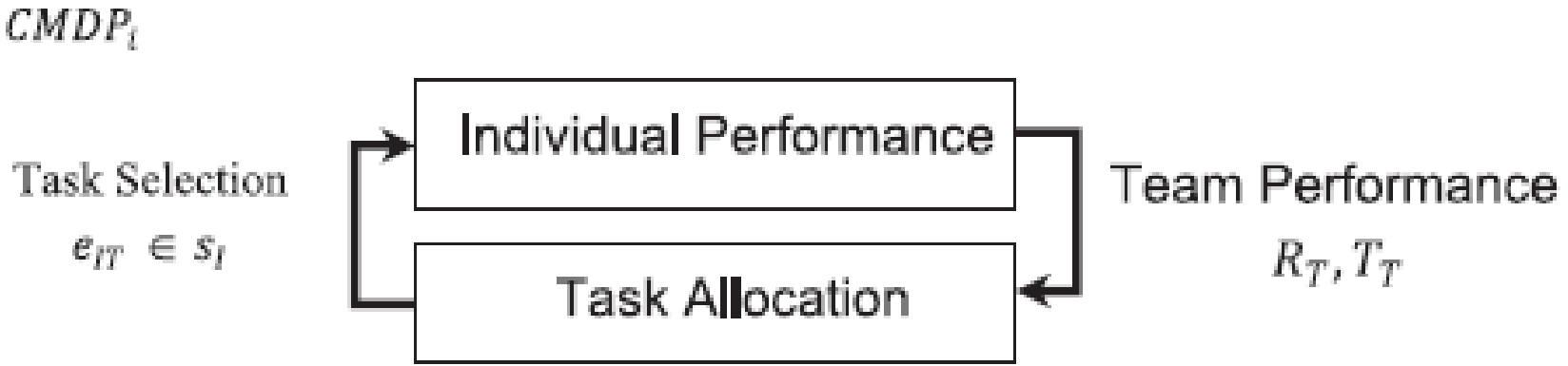
Social: Individuals learn from each other

CISL

ABSTRACT

Multi-agent learning, in a decision theoretic sense, may run into deficiencies if a single Markov decision process (MDP) is used to model agent behaviour. This paper discusses an approach to overcoming such deficiencies by considering a multi-agent learning problem as a concurrence between individual learning and task allocation MDPs. This approach, called *Concurrent MDP* (CMDP), is contrasted with other MDP models, including decentralized MDP. The individual MDP problem is solved by a Q-Learning algorithm, guaranteed to settle on a locally optimal reward maximization policy. For the task allocation MDP, several different concurrent individual and social learning solutions are considered. Through a heterogeneous team foraging case study, it is shown that the CMDP-based learning mechanisms reduce both simulation time and total agent learning effort.

One Individual Agent



Individual Markov Decision Process

3.1.1. Individual performance MDP

Individual agent progress toward a sub-task can be defined as a $\langle S_I, A_I, T_I, R_I \rangle$ tuple,

- S_I denotes a discrete set of states, whose intrinsic value is partially specified by the task allocation process through a set of evidence E_{IT}.
- A_I denotes a discrete set of actions.
- $T_I(s_I, a_I, s'_T)$ denotes a stable transition model, i.e., the probability of executing action a_I starting from state s_I and ending up in state s'_I.
- $R_I(s_I, a_I, s'_I)$ denotes a positive real number as a reward received for transitioning from state s_I into state s'_I using action a_I.

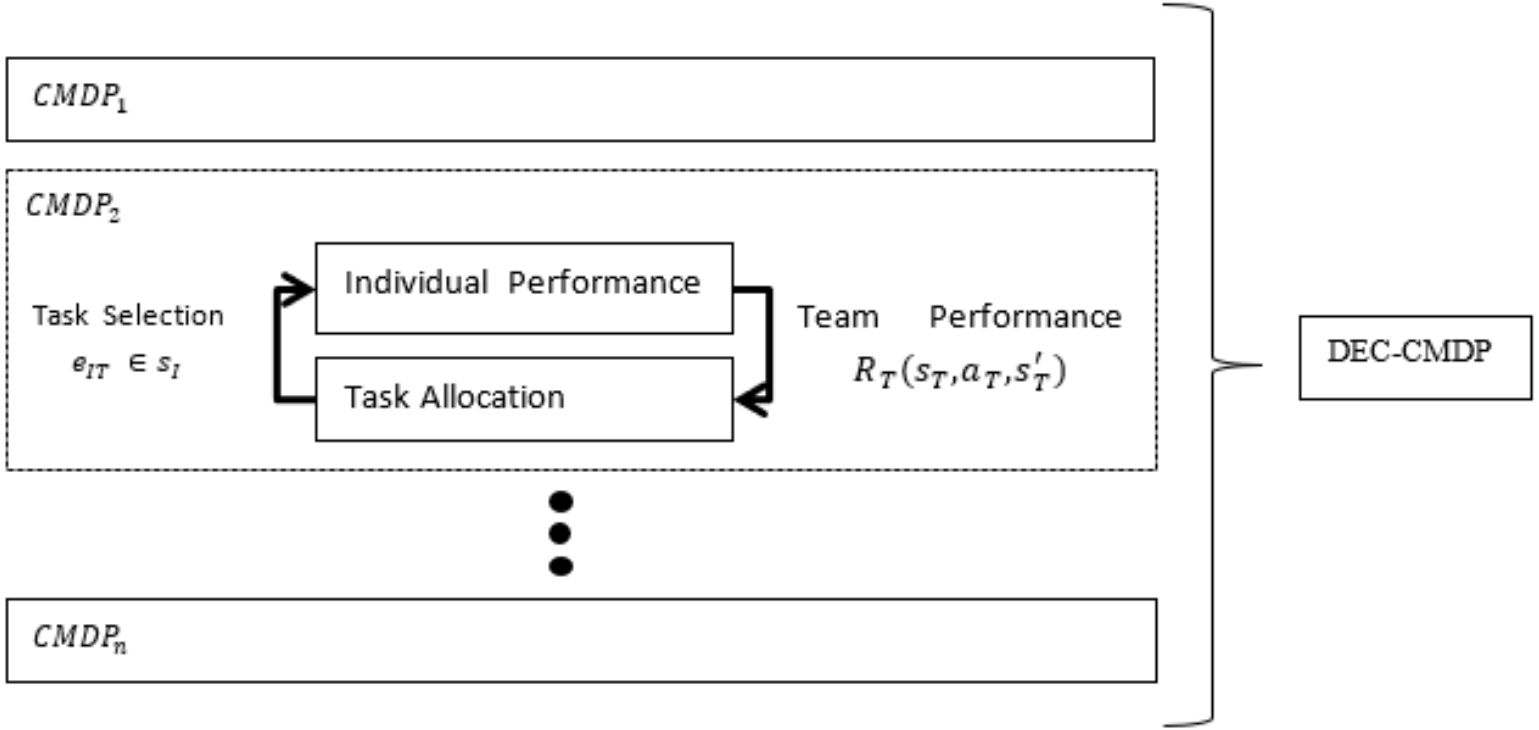
Group Markov Decision Process

3.1.2. Task allocation MDP

Team progress toward a goal can be defined as a $\langle S_T, A_T, T_T, R_T \rangle$ tuple,

- S_T denotes a discrete set of states, which captures the individual performance characteristics between all agents and all tasks.
- A_T denotes a discrete set of actions, i.e., a function that assigns all available tasks to available agents.
- $T_T(s_T, a_T, s'_T)$ denotes a stable transition model, i.e., the probability of executing action a_T starting from state s_T and ending up in state s'_T. The transition function is completely specified by the individual performance process through a set of evidence E_{TI}.
- $R_T(s_T, a_T, s'_T)$ denotes a positive real number as a reward that is received after taking a certain action a_T within state s_T when arriving in state s'_T. The reward function is completely specified by the individual performance process through a set of evidence E_{TI}.

A number of agents all study the space

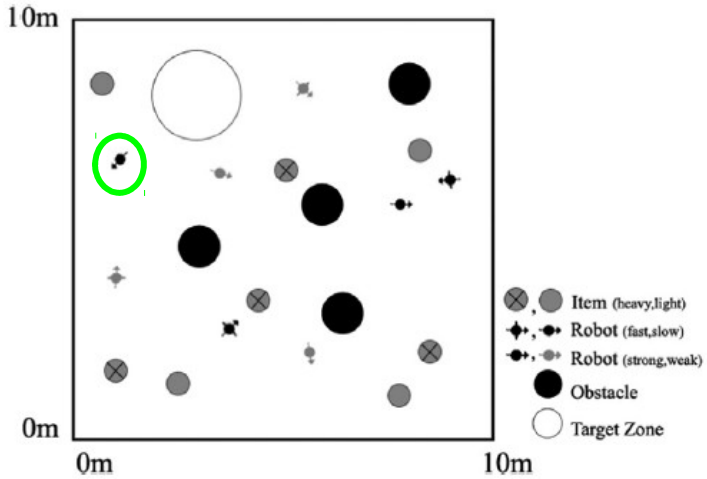


Individual Performance MDP

4.1.1. Individual performance MDP

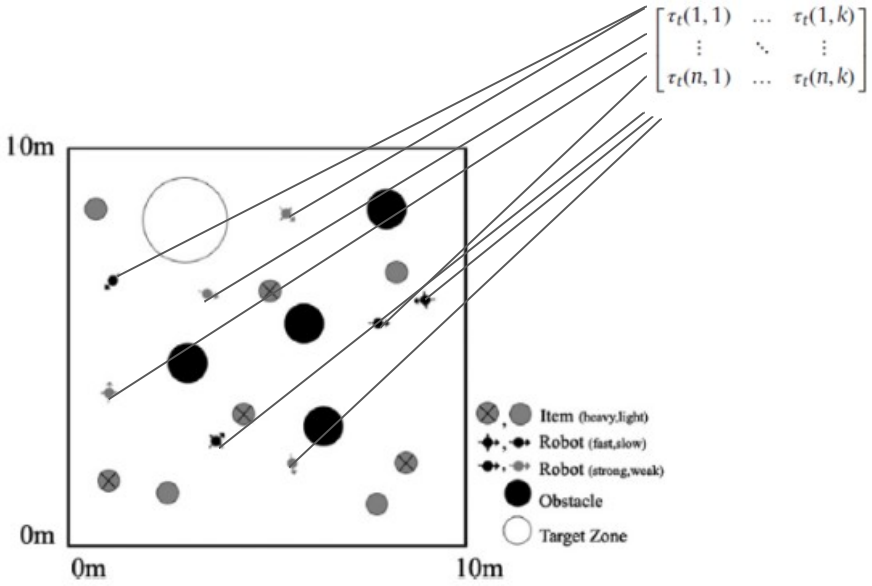
The individual MDP used in this case study consists of a $\langle S_I, A_I, T_I, R_I \rangle$ tuple defined as follows:

State $s_I \in S_I$: is defined by $s_I = \{r_x, r_y, r_\theta, o_x, o_y, l_x, l_y, l_p, g_x, g_y, b_x, b_y\}$, which includes the information about the robot r,



Action $a_I \in A_I$: is *move_forward*, *move_backward*, *turn_right*, *turn_left*, or *interact*. Each robot moves at either 0.20 m (slow robot) or 0.40 m (fast robot) per *move* action. All turns are fixed at $\pi/4$ radians per *turn*. *Interact* attaches the robot to an assigned item when the robot is currently not gripping another item.

Task Allocation MDP



4.1.2. Task allocation MDP

The task allocation learning problem is to determine the *optimal* assignment of items to the set of agents, where a task selection action at each iteration t is considered optimal when the robots minimize their average item return effort, $\overline{\tau}_t(i,i',j)$. First, the single-item-single-robot assignment will be discussed using the average trial duration, $\tau_t(i,j)$, and afterward the robots' pairwise *physical cooperation* will be discussed. The task allocation MDP is modeled as a $\langle S_T, A_T, T_T, R_T \rangle$ tuple:

(1) **State $s_T \in S_T$**: contains three pieces of information: each robot's average trial time toward each task, each items current *delivery status*, and each robot's *assignment status* toward each task. The state consists of three sub-states, $\left\{s^1_{T,t}, s^2_{T,t}, s^3_{T,t}\right\}$.

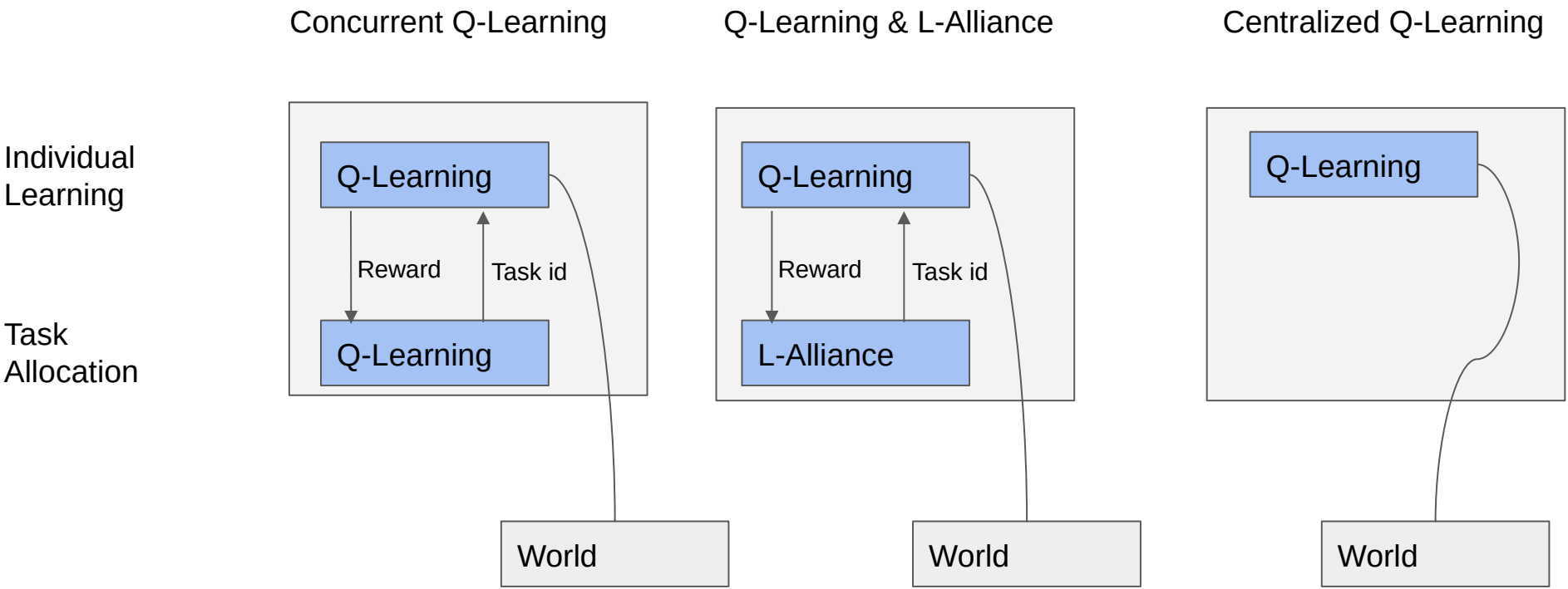
The robot performance sub-state is described by

$$s^1_{T,t} = \begin{bmatrix} \tau_t(1,1) & \dots & \tau_t(1,k) \\ \vdots & \ddots & \vdots \\ \tau_t(n,1) & \dots & \tau_t(n,k) \end{bmatrix}, \tag{21}$$

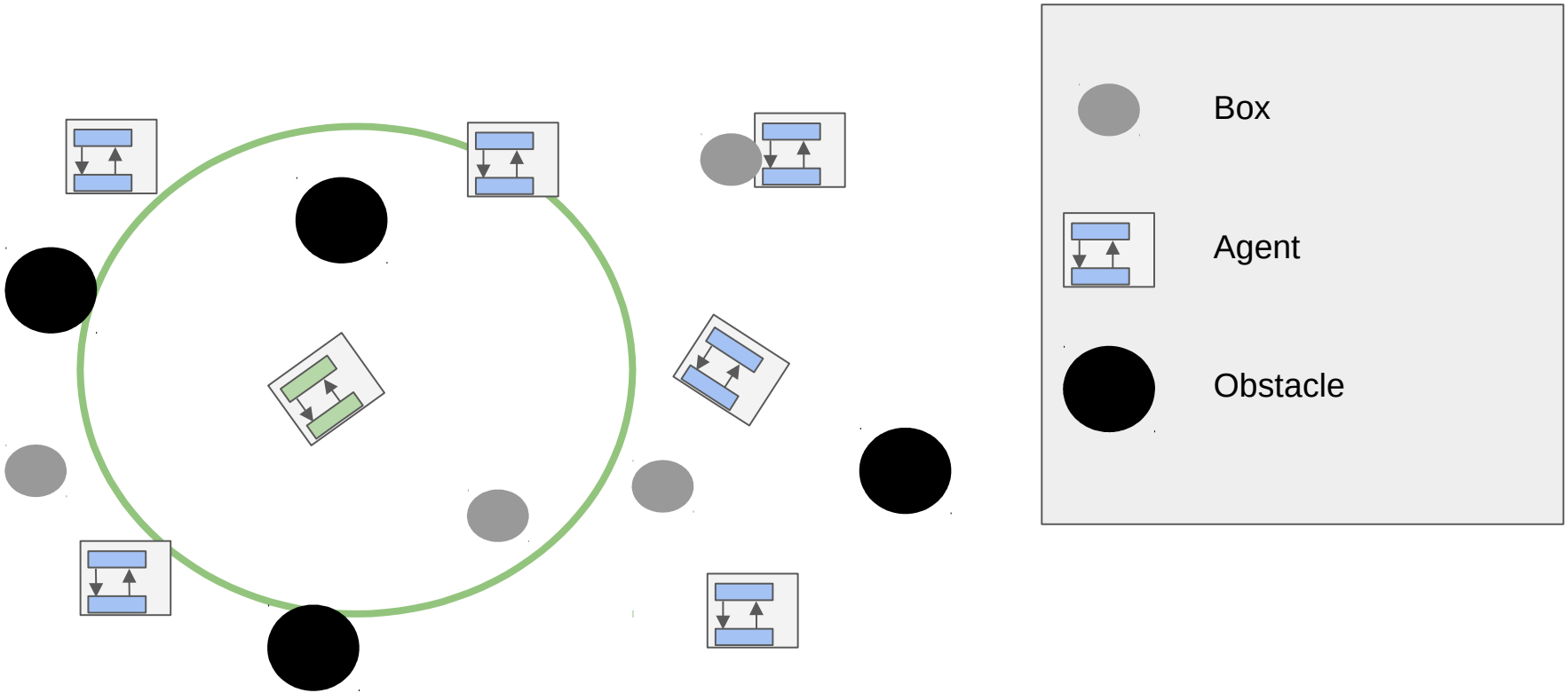
where n and k denote the number of robots and items, respectively. For the end of this section we note an inversed robot performance matrix:

$$\tilde{s}^1_{T,t} = \begin{bmatrix} 1/\tau_t(1,1) & \dots & 1/\tau_t(1,k) \\ \vdots & \ddots & \vdots \\ 1/\tau_t(n,1) & \dots & 1/\tau_t(n,k) \end{bmatrix} \tag{22}$$

Simulation (What kind of agents)



Simulation Setup



It works! Individual Performance

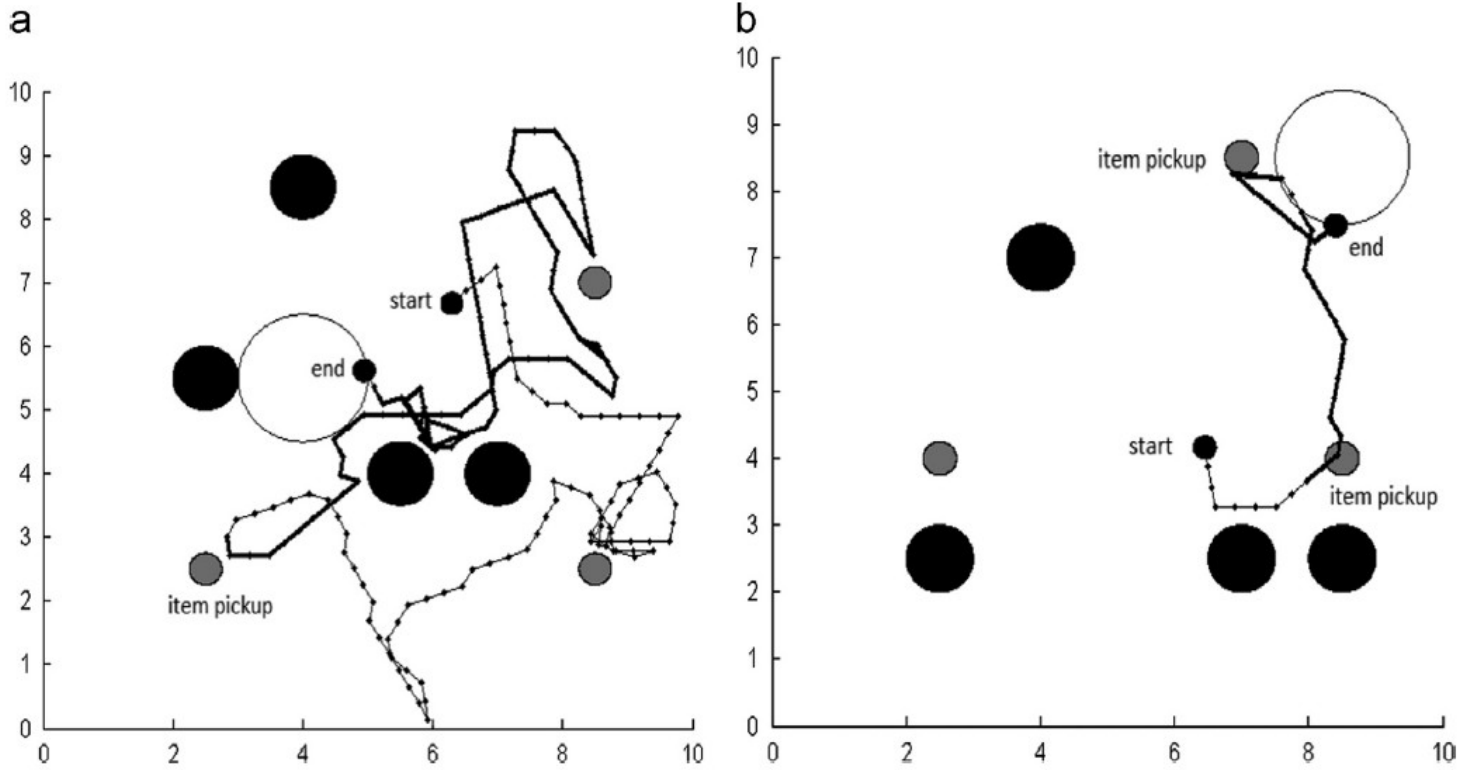
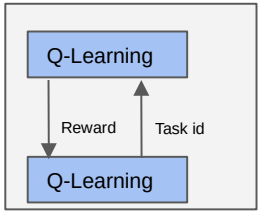
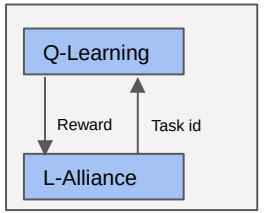
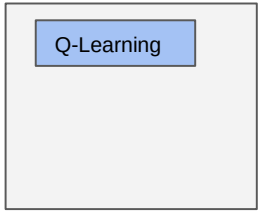


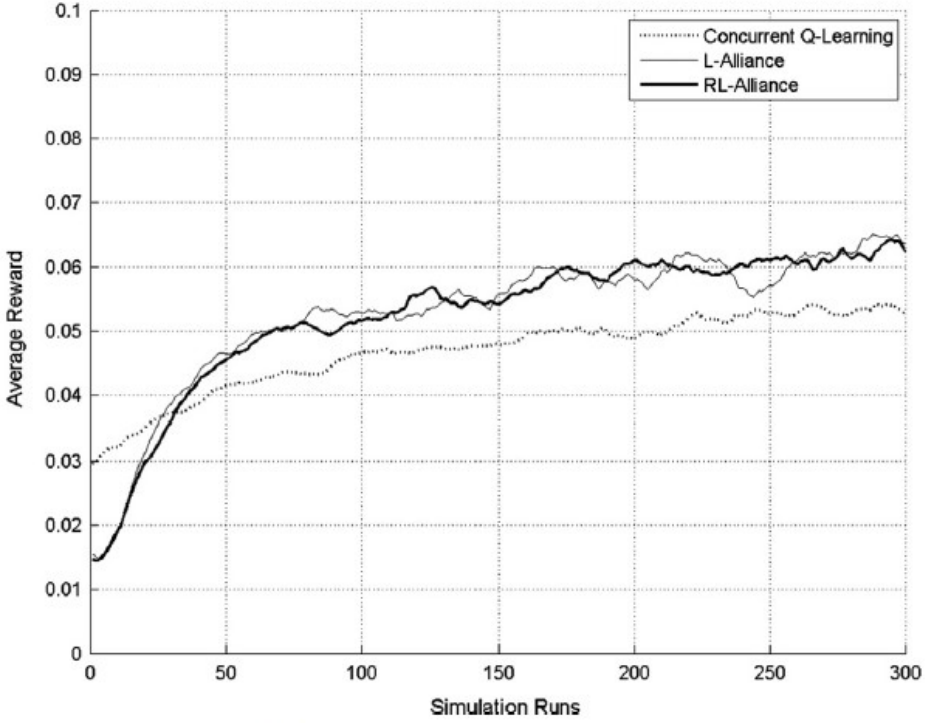
Fig. 3. Individual performance of an untrained robot (a) vs. the same robot when trained (b).

Reward and Simulation Time









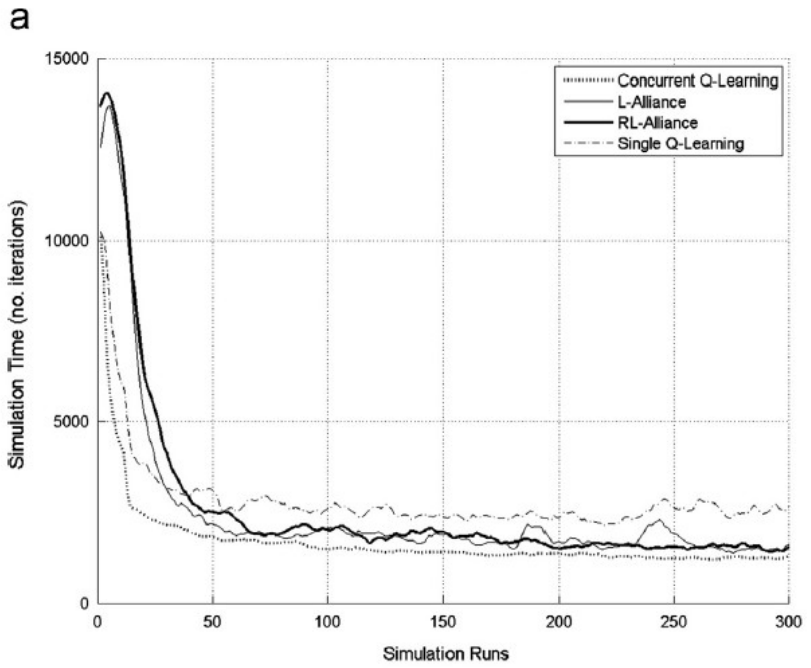
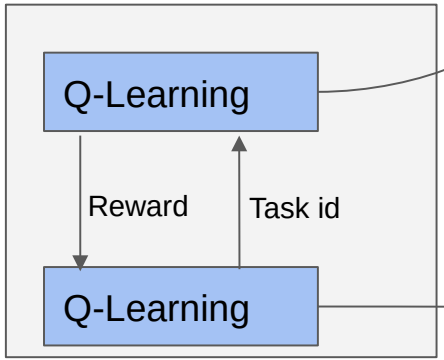


Fig. 4. Average individual reward (R_I) for a typical robot assigned to an item.

Wait. Is this optimal. Proof -- The main contribution

Concurrent Q-Learning



Individual Converges as “actions” approach infinity

Maximization of Individual Reward

Optimize Performance as “tasks-agent” approaches infinity

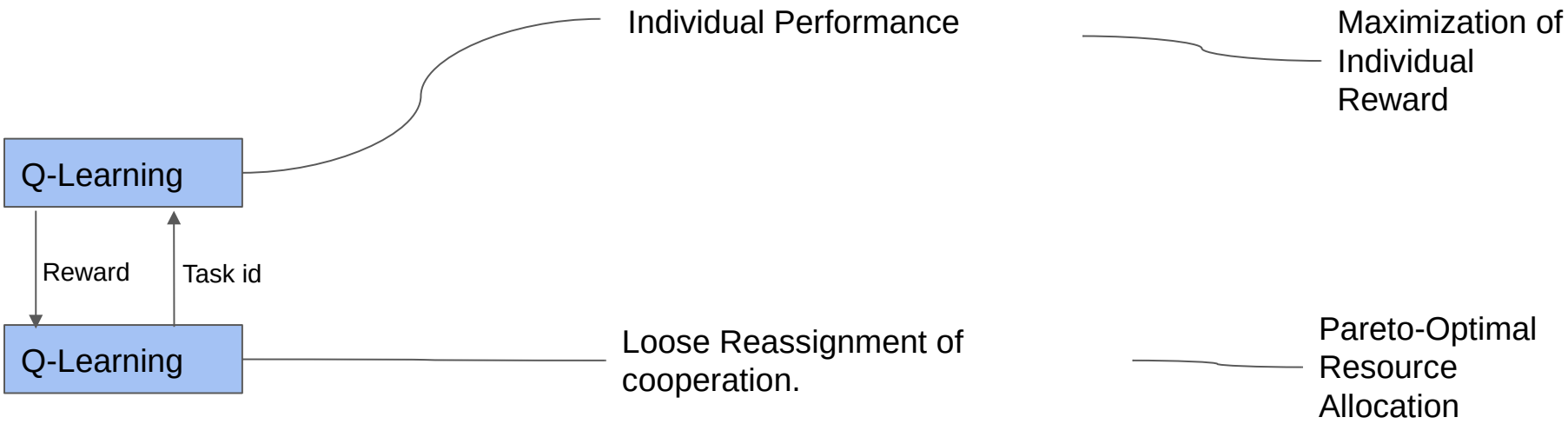
Pareto-Optimal Resource Allocation

Assumptions:

The Degenerate State Space perfectly Characterizes the problem (states change)

Reward is Stably Stochastic (individuals redefine their ‘happy’)

The General Paradigm: Separable Problems



Assumptions:

The Degenerate State Space perfectly Characterizes the problem (states change)

Reward is Stably Stochastic (individuals redefine their ‘happy’)