TRANSACTIONS ON MACHINE LEARNING RESEARCH (TMLR)

Reward Engineering for Spatial Epidemic Simulations

A Reinforcement Learning Platform for Individual Behavioral Learning

Radman Rakhshandehroo
Department of Computer Science
University of British Columbia

Daniel Coombs
Department of Mathematics & Institute of Applied Mathematics
University of British Columbia

01 — BACKGROUND: EPIDEMIC MODELS

Epidemic Models

SIR

Susceptible → Infected → Recovered

The simplest compartmental model. Individuals move through three states governed by ordinary differential equations (ODEs).

SEIR

+ Exposed compartment

Adds a latent *Exposed* state between Susceptible and Infected, capturing incubation periods.

SIRS+D

+ Waning immunity & Death

Recovered individuals can become Susceptible again. A *Dead* state captures mortality. This is the model used in ContagionRL.

- All ODE-based models assume homogeneous, well-mixed populations: no individual behavior, no spatial structure.

02 — MOTIVATION

Individual behavior shapes epidemics, but how do we model it?

ODE (ORDINARY DIFFERENTIAL EQUATION) MODELS

SIR/SEIR assume homogeneous populations. No individual behavior or spatial interactions.

AGENT-BASED MODELS

Capture heterogeneity but rely on prescribed, rule-based behaviors rather than learned policies.

THE GAP

No platform for systematic study of how reward design affects learned epidemic behavior.

03 — BACKGROUND: REINFORCEMENT LEARNING

Reinforcement Learning

An agent learns by **trial and error**, taking actions in an environment and receiving rewards. The goal is to learn a **policy** that maximizes cumulative reward.

MDP — MARKOV DECISION PROCESS

Defined by states, actions, rewards, and transitions. The agent observes the *full* environment state at each step.

POMDP — PARTIALLY OBSERVABLE MDP

The agent can only see a *partial* observation of the true state. More realistic, but harder to solve.

NPI — NON-PHARMACEUTICAL INTERVENTION

Behavioral measures like masking and social distancing that reduce transmission without medication or vaccines.

ALGORITHMS WE EVALUATE

PPO (Proximal Policy Optimization), **SAC** (Soft Actor-Critic), and **A2C** (Advantage Actor-Critic). Standard policy-gradient RL algorithms for continuous control.

04A — PLATFORM

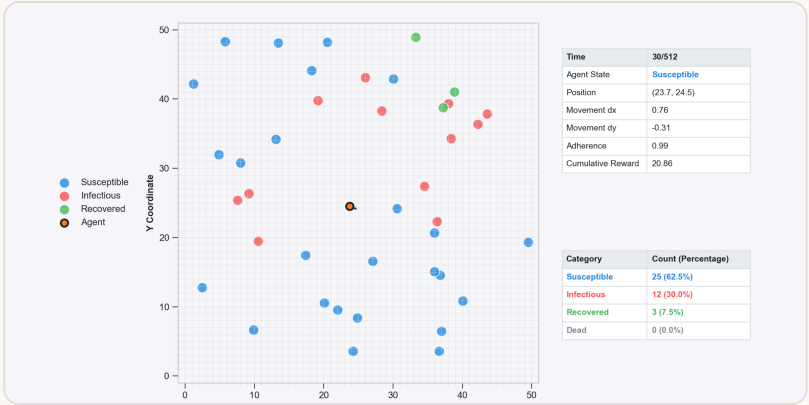
ContagionRL

SIRS+D Compartmental model with Susceptible, Infected, Recovered, and Dead states on a toroidal grid

AGENT Single RL agent among 40 non-learning humans. Continuous action: movement vector + NPI adherence level

OBS Adherence level, relative positions, normalized distances, infection status of each human

GOAL Remain susceptible as long as possible. Episode ends upon first infection.



Gymnasium-compatible · Stable-Baselines3 · Open source

04B — PLATFORM

A highly configurable simulation

EPIDEMIOLOGICAL

Infection rate (β), recovery rate (ρ), immunity loss (λ), lethality (δ), distance decay (k_d), adherence effectiveness (ε_α)

SPATIAL

Grid size, population count, initial infected count, safe distance, max infection distance, toroidal boundary wrapping

MOVEMENT

Stochastic (random walk) or deterministic (workplace/home cycles). Configurable movement scale and patterns

OBSERVABILITY

Full visibility (MDP) or configurable visibility radius (POMDP). Tested at r = 10, 15, 20, and ∞

REWARD ENGINEERING

Five reward functions with ablation variants. Pluggable design for custom reward functions

EPISODE CONTROL

Max timesteps, reinfection mechanism, initial agent distance, configurable termination conditions

20+ configurable parameters · Supports PPO, SAC, A2C via Stable-Baselines3

05A — CORE MECHANISM

Distance-based infection

$$P_{\text{inf}}(h_s) = 1 - \exp\left(- \underbrace{\beta}_{\substack{\text{base infection} \\ \text{rate}}} \underbrace{\sum_{h_i}}_{\substack{\text{sum over} \\ \text{infected}}} \underbrace{e^{-k_d \cdot d(h_s, h_i)}}_{\substack{\text{exponential decay} \\ \text{with distance}}} \right)$$

Each infected individual contributes to infection risk based on proximity. The total exposure is summed over all infected neighbors, then converted to a probability.

Parameters

$P_{inf}(h_s)$	infection probability for susceptible h_s
β	base infection rate
h_i	each infected individual
k_d	distance decay factor
$d(h_s, h_i)$	distance between susceptible and infected

05B — CORE MECHANISM

Adherence modulates risk

$$\underbrace{\beta_{\text{eff}}}_{\text{effective rate}} = \underbrace{\beta}_{\text{base rate}} \cdot \Big(\underbrace{\epsilon_\alpha}_{\text{residual risk (never zero)}} + (1-\epsilon_\alpha)\ \underbrace{(1-\alpha)}_{\text{non-adherence fraction}} \Big)$$

Higher adherence α reduces the effective infection rate, but never eliminates risk entirely. A residual ε_α always remains. The agent controls both its movement and adherence level at each step.

Parameters

β_{eff}	effective infection rate after adherence
β	base infection rate
α	NPI adherence level $\in [0, 1]$
ε_α	residual risk at full adherence (never zero)
$(\Delta x, \Delta y, \alpha)$	full action space: movement + adherence

05C — CORE MECHANISM

What the agent sees and does

Observation: per visible human h_j

Relative position	$(\Delta x_j, \Delta y_j) \in [-0.5, 0.5]$
Normalized distance	d_j / d_{max}
Infection status	$I_j \in \{0, 1\}$
Agent adherence	$\alpha_t \in [0, 1]$

Action space: continuous, 3D

Movement vector	$(\Delta x, \Delta y) \in [-1, 1]^2$
NPI adherence	$\alpha \in [0, 1]$
Position update	$x_{t+1} = (x_t + \Delta x) \bmod G$

Observability formulations

MDP Agent observes all humans on the grid. Full state information at every step.

POMDP Only humans within a visibility radius r are observed. Beyond r, individuals are invisible. Tested at r = 10, 15, 20.

Key design choice

The reward signal uses global state (privileged critic), while the policy only sees the local observation. This intentional asymmetry lets the agent learn from perfect information but act under realistic constraints.

06A — REWARD DESIGN SPACE

Five reward functions, sparse to dense

01

Constant

Fixed positive reward while alive. Sparse and uninformative.

02

Reduce Inf. Prob.

Directly incentivizes minimizing calculated infection likelihood.

03

Combined

Constant bonus plus infection probability reduction together.

04

Max Distance

Rewards maintaining distance from nearest individuals.

05

Potential Field

Force-based directional guidance + health + adherence incentives.

- The potential field reward uses global state to compute forces during training (privileged critic), while the agent's policy observes only what its visibility setting allows: full grid (MDP) or a limited radius (POMDP). The reward always sees everything; the agent may not.

The potential field reward

$$R_{\text{PF}} = \underbrace{w_h}_{0.1} \cdot \underbrace{r_{\text{health}}}_{\substack{\text{alive}\\\text{bonus}}} + \underbrace{w_\alpha}_{0.2} \cdot \underbrace{r_{\text{adherence}}}_{\substack{\text{NPI}\\\text{level}}} + \underbrace{w_m}_{0.7} \cdot \underbrace{r_{\text{move}}}_{\substack{\text{directional}\\\text{guidance}}}$$

r_{health} = 1 if susceptible, 0 otherwise · $r_{adherence}$ = α · r_{move} dominates at 70% (detailed below)

Repulsive force field

$$\mathbf{F} = \sum_j \underbrace{\frac{w_j}{d_j^2}}_{\substack{\text{inverse-square}\\\text{weighting}}} \cdot \underbrace{\Delta\mathbf{p}_j}_{\substack{\text{displacement}\\\text{from human } j}}$$

Each human generates a repulsive force. Infected push harder (W_I = 1.0) than susceptible (W_S = 0.5). Forces decay with inverse-square distance.

Movement reward components

$$r_{\text{dir}} = \text{clip}\left(\frac{\mathbf{a} \cdot \hat{\mathbf{F}}}{\|\mathbf{a}\|},\ -1,\ 1\right)$$

$$r_{\text{mag}} = \text{clip}\big(1 - \big|\|\mathbf{a}\| - \min(\|\mathbf{F}\|,\ 1)\big|,\ -1,\ 1\big)$$

$$r_{\text{move}} = \underbrace{0.75}_{1-\beta_m} \cdot\ r_{\text{dir}} \ +\ \underbrace{0.25}_{\beta_m} \cdot\ r_{\text{mag}}$$

`a` agent's movement vector

$\hat{F}$ unit vector of resultant force

Δp_j toroidal displacement to human j

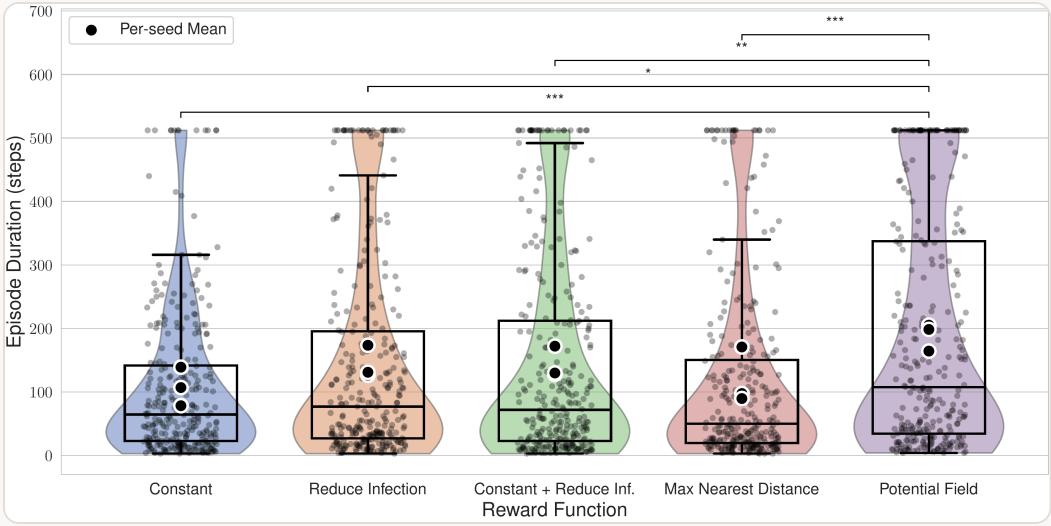
r_{dir} is the agent moving the right way?

r_{mag} is the agent moving the right speed?

w_j W_I = 1.0 (infected), W_S = 0.5 (susceptible)

07 — RESULT: REWARD COMPARISON

Reward choice dramatically impacts survival



Episode duration across five reward functions (higher is better)

Potential field wins

Significantly longer episode durations than all other reward designs ($p < 0.001$).

Sparse rewards fail

Constant reward provides too little signal for nuanced spatial decision-making.

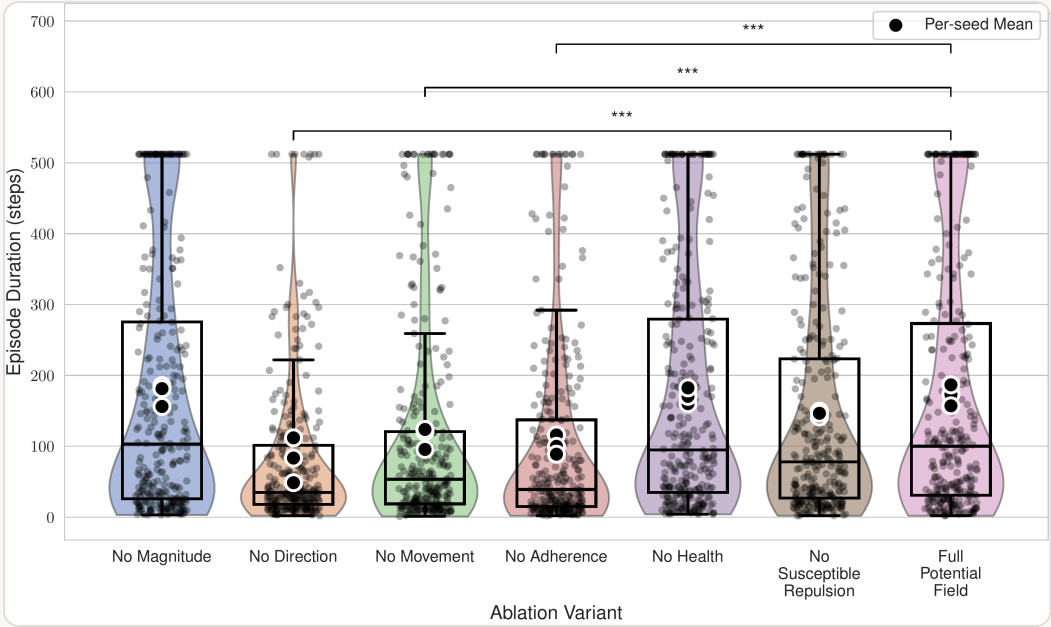
Local optima trap

Max Distance and Reduce Infection Prob. rewards lead to myopic optimization.

Universal adherence

All reward functions learn maximal NPI adherence. The challenge is spatial navigation.

Direction and adherence are the critical ingredients



Ablation of potential field reward components: which pieces matter most?

No Direction

Significant degradation ($p < 0.001$). Agent cannot learn effective spatial strategies without directional cues.

No Adherence

Significant degradation. With complex rewards, adherence must be explicitly incentivized.

No Magnitude

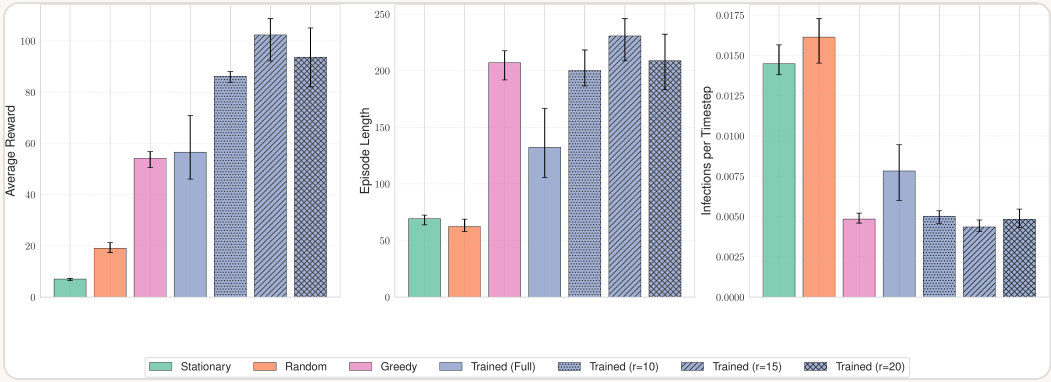
No significant change. The direction of movement matters more than its scaling.

No Health / No Susc. Repulsion

No significant impact. Effects are captured by other components.

09 — RESULT: PARTIAL OBSERVABILITY

Less information, better performance



Full vs. limited visibility: average reward, episode length, and infection rate

Counterintuitive

Partially observable (POMDP) agents outperform the fully observable agent across all three metrics.

Information bottleneck

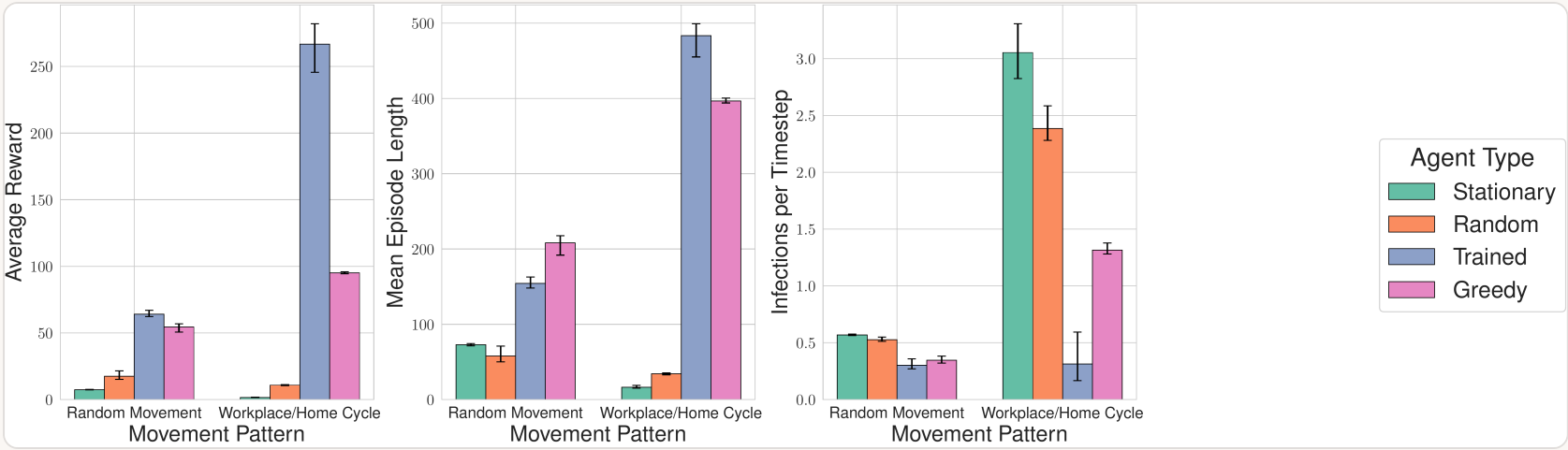
Limited visibility reduces observation noise, forcing more robust policy learning.

Plateau effect

No significant difference between r=10, 15, and 20. More info doesn't help beyond a threshold.

10A — MOBILITY PATTERNS

Structured mobility enables long-horizon planning



Structured vs. random human movement: reward, episode length, and infection rate

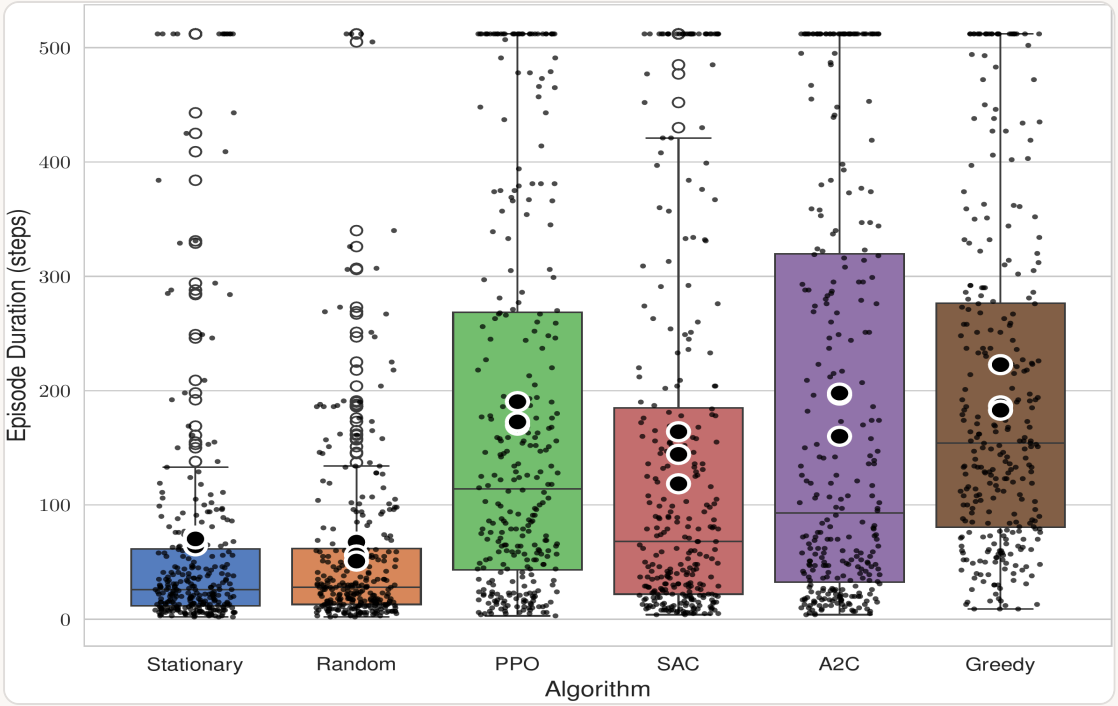
Spatio-temporal regularity

Trained agents exploit predictable workplace/home cycles for significantly longer survival ($p < 0.001$).

Population impact

A single trained agent reduces population-level infection rates by 10-21% relative to baselines ($p < 0.01$).

RL agents match strong heuristics



Episode duration: RL agents vs. baselines

Algorithm-agnostic

PPO, SAC and A2C all significantly outperform random and stationary baselines.

Competitive with heuristics

Mean RL performance is comparable to the handcrafted greedy heuristic, validating the platform.

11 — TAKEAWAYS

Key contributions

01 **Reward design dramatically shapes learned epidemic behavior.** Potential field reward with directional guidance achieves significantly superior survival.

02 **An open-source, highly configurable RL platform for epidemic research.** 20+ parameters, Gymnasium-compatible, works with any standard RL library. Accessible to epidemiologists and computer scientists alike.

03 **Directional guidance and adherence incentives are essential.** Ablation reveals these as the critical components; magnitude and health terms are secondary.

04 **Partial observability can improve robustness.** Information bottleneck forces more generalizable policies, outperforming full visibility.

05 **Structured mobility enables long-horizon planning.** Workplace/home cycles yield substantially better trained agent performance than random movement.

arXiv:2511.18000

TRANSACTIONS ON MACHINE LEARNING RESEARCH (TMLR)

Thank you

Radman Rakhshandehroo
Department of Computer Science
University of British Columbia

Daniel Coombs
Department of Mathematics & Institute of Applied Mathematics
University of British Columbia

rdmnr@protonmail.com github.com/redradman/ContagionRL arxiv.org/abs/2511.18000