Reward Engineering for Spatial Epidemic Simulations

A Reinforcement Learning Platform for Individual Behavioral Learning

Radman Rakhshandehroo
DEPARTMENT OF COMPUTER SCIENCE · UBC

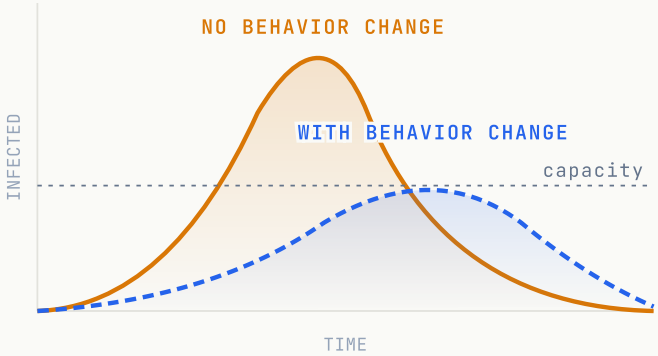
Daniel Coombs
MATHEMATICS & INST. OF APPLIED MATH · UBC

TRANSACTIONS ON MACHINE LEARNING RESEARCH · TMLR 2026

Epidemics are driven by individual behavior.

Who moves, who masks, who distances. These choices shifted the curve in 2020 more than any single policy intervention.

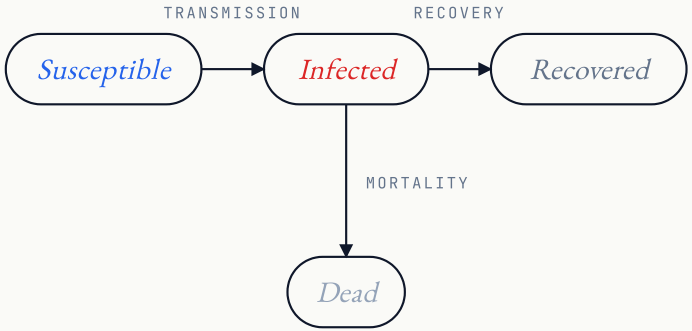
But the tools we use to model that behavior are bad at it.



Same disease, two epidemic curves. The difference is how people behaved.

COMPARTMENTAL (ODE) MODELS

Well-mixed by assumption.

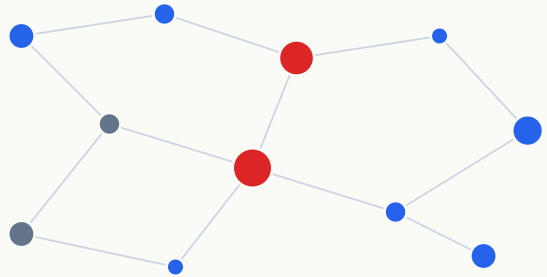


POPULATIONS FLOW AS RATES · NO INDIVIDUALS

Populations flow between compartments by ODEs. No individuals, no space.

AGENT-BASED MODELS

Heterogeneity by hand-coding.



`if neighbor.infected → infect(β)`

HAND-CODED RULES

Individuals differ — but their behavior is prescribed by the modeller.

Reward design steers the policy. For epidemic models, it's been barely studied.

RL replaces hand-coded rules with a learned policy. Every step, the reward grades the action and the policy updates. The reward function decides what the agent ends up doing.

That's the gap this work fills.

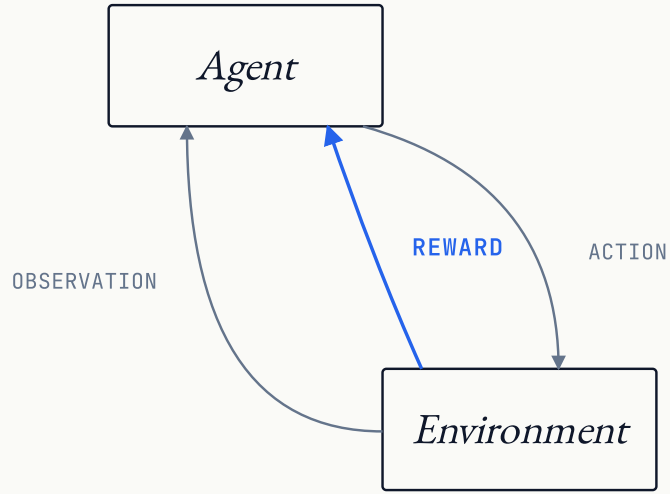


Figure 1 — The reinforcement-learning loop.

05 — WHAT WE BUILT

ContagionRL: a platform for testing how reward design shapes epidemic behaviour.

A Gymnasium-compatible RL environment built on a spatial SIRS+D simulation.

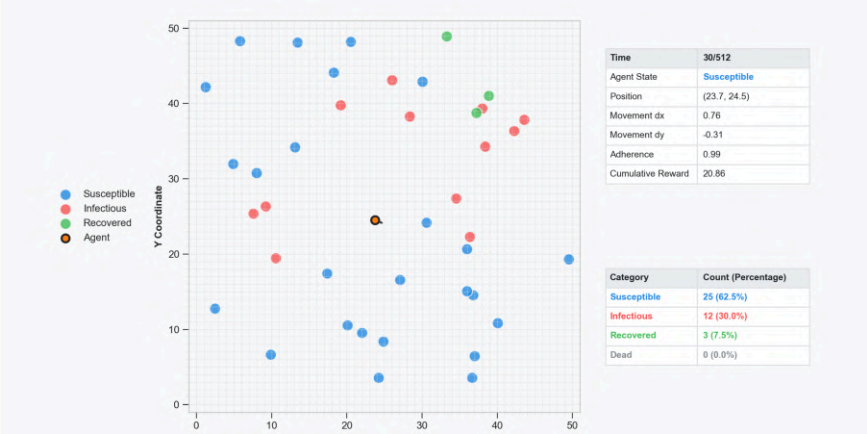


Figure 2 — ContagionRL environment render. Coloured discs are S/I/R/D agents on a toroidal grid.

ENVIRONMENT

Spatial SIRS+D on a toroidal grid.

AGENTS

One RL agent (PPO). Forty non-learning humans.

OBJECTIVE

Stay susceptible. Episode ends on first infection.

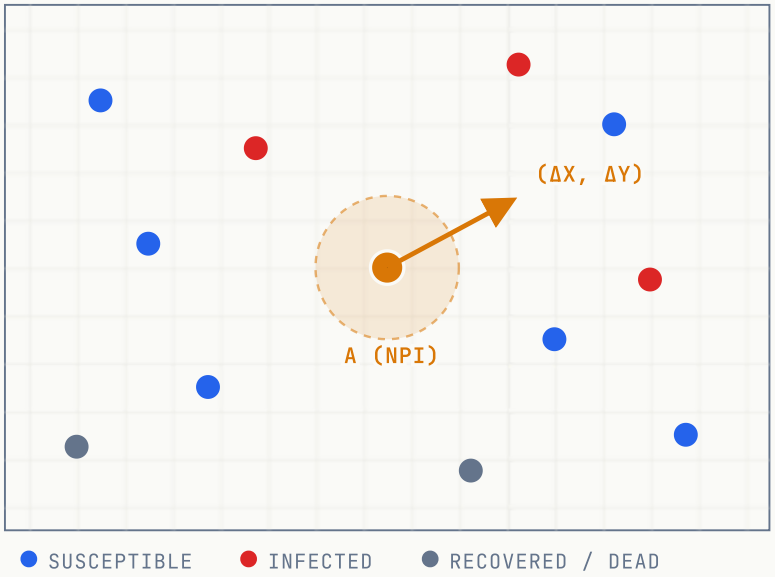


Figure 3 — Per-step action. The RL agent (orange) picks a move and an adherence level.

06 — THE AGENT'S CONTROLS

Two controls per step: a movement vector, and an NPI adherence level.

MOVEMENT

$$(\Delta x, \Delta y) \in [-1, 1]^2$$

Continuous step on the toroidal grid.

ADHERENCE A

$$\alpha \in [0, 1]$$

How strictly the agent follows **non-pharmaceutical interventions** like masking and distancing. Higher α cuts transmission risk, but never to zero.

$$\beta_{\text{eff}} = \beta \cdot (\varepsilon_\alpha + (1 - \varepsilon_\alpha)(1 - \alpha))$$

Higher α drops β_{eff}; residual risk ε_α always remains.

07 — THE QUESTION

Five rewards. One outcome target.

All five rewards aim at the same goal: stay susceptible. They differ only in how they shape the signal the agent learns from.

Constant

Flat +1 per step alive. Sparse, no spatial guidance.

Reduce P_{inf}

Maximize $1 - P_{inf}$. Local and myopic.

Combined

Reduce P_{inf} plus a survival floor for dense crowds.

Max Distance

Reward distance from nearest human; capped past contagion threshold.

THE COMPOSITE

Potential Field.

+ Health bonus

+ Adherence reward

+ Directional movement field

Only reward with a direction, not just a score.

08 — HEADLINE RESULT

One reward dominates, and the win isn't algorithm-specific.

POPULATION SPILLOVER

The Potential Field agent survives longer. It also cuts new infections in the surrounding population by 21% versus the random baseline.

Selfish agents make better neighbours, apparently.

FIGURE 4A · POPULATION INFECTION RATE

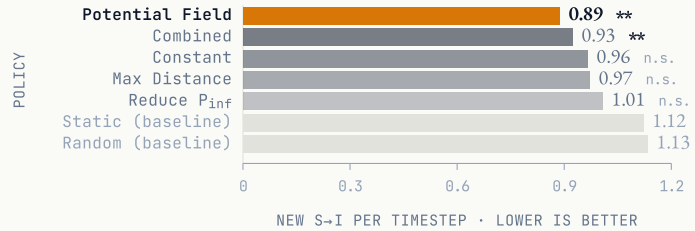
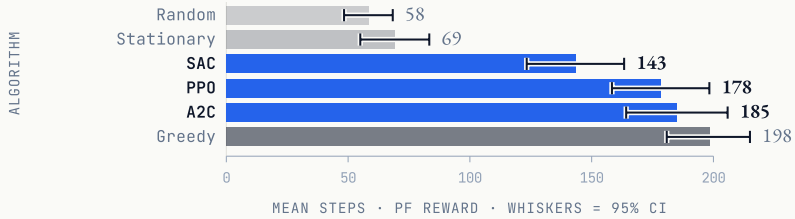


FIGURE 4B · ACROSS ALGORITHMS



ALGORITHM-AGNOSTIC

PPO, SAC, and A2C all learn effective policies and beat the random and stationary baselines by 2–3×.

The result isn't tied to one algorithm. The platform makes swapping them easy.

09 — REWARD COMPARISON

The shape of the reward decides the policy.

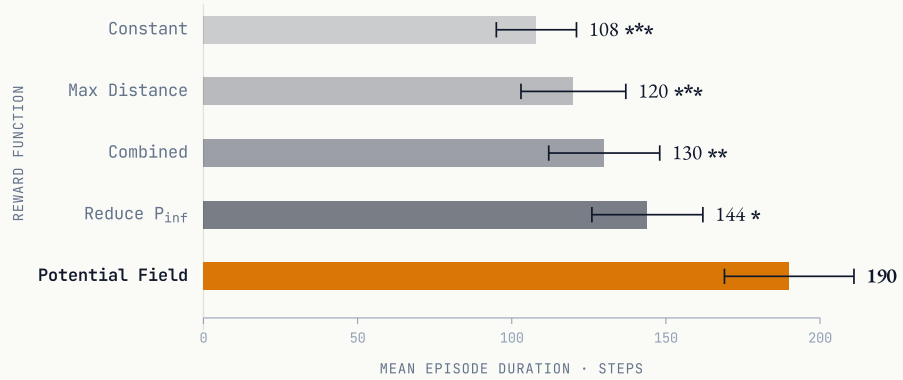
Mean episode duration across 300 evaluation runs per reward (3 seeds × 100 episodes).

Sparse rewards give the policy no spatial signal.

Local-dense rewards plateau in short-sighted optima.

Potential Field is the only directional signal.

FIGURE 5 · SURVIVAL BY REWARD (MEAN STEPS, PPO)



*Bars: mean episode duration. Whiskers: 95 % bootstrap CI. * p<0.05, ** p<0.01, *** p<0.001 vs. Potential Field (Mann–Whitney U, Bonferroni-corrected).*

10 — ABLATION

Two pieces do the work.

We strip one Potential-Field component at a time and retrain. Some removals collapse the policy; most leave it intact.

Direction: alignment with the repulsion field.

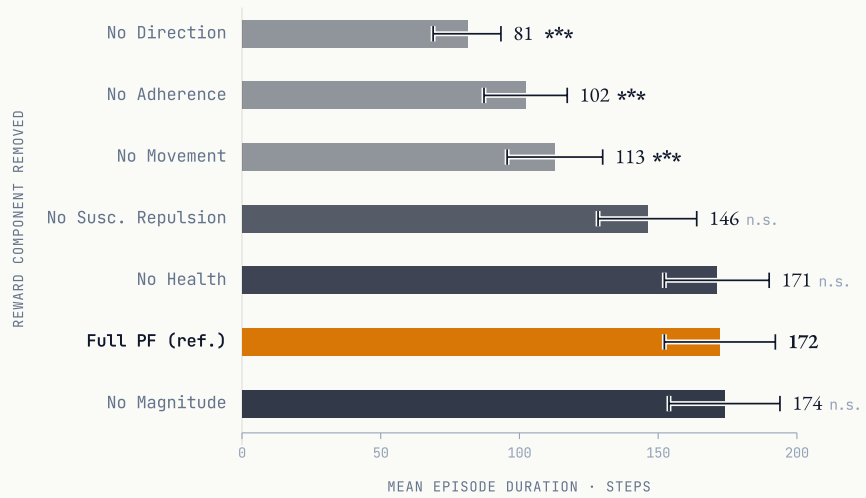
Adherence: explicit NPI incentive.

Magnitude: loss absorbed by the remaining field.

Health bonus: no measurable effect.

Susceptible repulsion: no measurable effect.

FIGURE 6 · ABLATION (MEAN STEPS)



*Whiskers: 95 % bootstrap CI. *** p<0.001 vs. Full PF; n.s. = not significant (Mann–Whitney U, Bonferroni-corrected).*

11 — COUNTERINTUITIVE

Less observation, better policy.

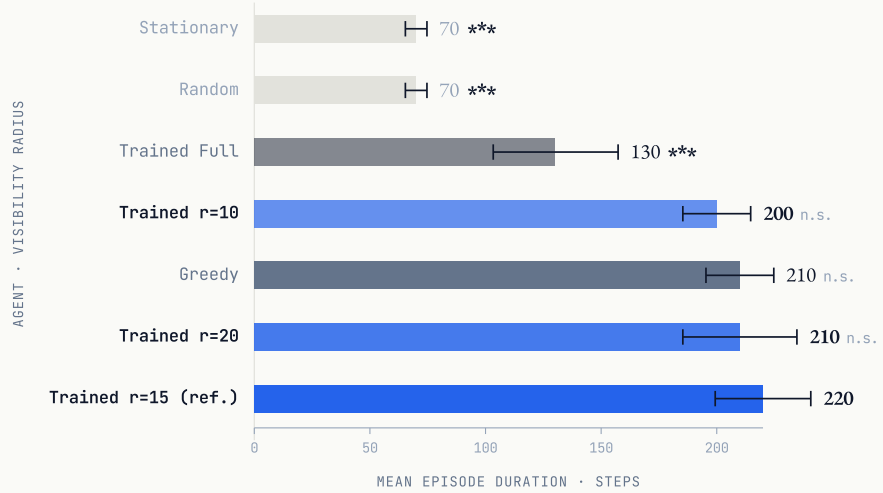
Visibility radius swept at r = 10, 15, 20, and full. We track average reward, episode length, and infections per timestep.

> **Limited visibility** beats full visibility on all three metrics.

> A narrow view forces focus on imminent threats.

> Reward still uses global state, so this is policy robustness rather than leakage.

FIGURE 7 · EPISODE LENGTH BY VISIBILITY (MEAN STEPS)



*Whiskers: 95 % bootstrap CI. *** p<0.001 vs. Trained r=15; n.s. = not significant (Mann–Whitney U, Bonferroni-corrected).*

12 — WHAT WE CONTRIBUTE

Three contributions to behavioural epidemic modelling.

01 · PLATFORM

A Gymnasium-compatible RL environment for spatial epidemics.

ContagionRL: open-source, spatial SIRS+D on a toroidal grid. Extensible to multi-agent setups and other epidemic dynamics.

02 · REWARD STUDY

First systematic comparison of reward functions for epidemic RL.

Five rewards × three algorithms. **Potential Field** dominates by 1.3×. Ablations isolate **direction** and **adherence** as the load-bearing components.

03 · ROBUSTNESS

An information-bottleneck finding under partial observability.

Limited-visibility agents (r = 10, 15, 20) outperform full-visibility ones. A narrower view forces focus on imminent threats.

FROM THE PAPER

Reward Engineering for Spatial Epidemic Simulations

QUESTIONS WELCOME

Thank you.

Radman Rakhshandehroo · Daniel Coombs

UBC · TRANSACTIONS ON MACHINE LEARNING RESEARCH (TMLR) 2026

PAPER
arxiv.org

CODE
github.com