Reward Engineering for Spatial Epidemic Simulations: A Reinforcement Learning Platform for Individual Behavioral Learning

Radman Rakhshandehroo · Daniel Coombs
University of British Columbia · Canadian AI / CRV 2026 Nectar Spotlight · Transactions on Machine Learning Research (TMLR)

arXiv

Background

ODE models flatten heterogeneity; agent-based models hard-code the behaviour we'd actually want to study. RL can learn it instead — but the policy depends almost entirely on how the reward is written, and reward design has been barely studied for epidemics. **ContagionRL** closes that gap: an open, Gymnasium-compatible platform for systematic reward engineering on a spatial SIRS+D environment under continuous-control RL.

Method: Environment & Agent

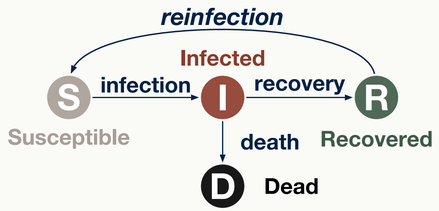
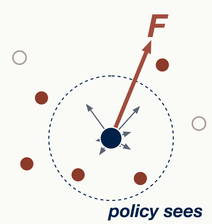


Figure 1. SIRS+D compartmental dynamics. R → S closes the loop after recovery; D is absorbing.

A single PPO agent shares a toroidal grid with $N = 40$ non-learning humans following SIRS+D dynamics. It sees each human's relative position and infection flag, then picks a movement vector $(\Delta x, \Delta y)$ and an adherence level $\alpha \in [0, 1]$ every step. Raising α drops the effective infection rate β_{eff}, but never to zero. The POMDP variant zeros out features for any human past a visibility radius r.

Method: Spatial Decision Geometry

Each human pushes the agent with an inverse-square force (infected push harder). The resultant $\boldsymbol{F}$ is the gradient the agent learns to follow.



Agent · Infected · Susceptible · Visibility radius · Net force $\boldsymbol{F}$

Figure 2. Each human exerts inverse-square repulsion on the agent; the reward aligns motion with the resultant $\boldsymbol{F}$. The reward sees every human; the policy sees only inside the dashed radius.

Method: Reward Functions

Five rewards, all aimed at the same outcome: keep the agent susceptible. **Constant** gives a flat $+1$ per step alive, with no spatial signal. **Reduce Inf. Prob.** uses $(1 - P_{\text{inf}})^2$, so the agent is rewarded for low instantaneous infection risk. **Combined** adds a 0.1 survival floor on top so reward never collapses in dense crowds. **Max Nearest Distance** rewards spacing from the closest human, capped past the contagion threshold D_β.

Potential Field, the winner, is the only composite reward. It sums three weighted pieces: **Health** ($w_h = 0.1$, a $+1$ while susceptible), **Adherence** ($w_\alpha = 0.2$, proportional to α), and **Movement** ($w_m = 0.7$, the dominant term). Movement itself splits into **Direction** (alignment with $\boldsymbol{F}$, 75% of its weight) and **Magnitude** (matching $|\boldsymbol{F}|$, the other 25%). $\boldsymbol{F}$ adds repulsions from infected ($W_I = 1.0$) and susceptible ($W_S = 0.5$) humans; turning W_S off is the **Susc. Repulsion** ablation. Each bold-italic piece above corresponds to one bar in the ablation chart.

Results: Reward Comparison

Potential Field beats the next-best reward by ≈1.3×, the worst by ≈1.8×. Sparse rewards (Constant) leave the agent without spatial guidance. Reduce-Inf-Prob and Max Distance optimise short-sighted heuristics and plateau in local optima. Only Potential Field's dense, directional signal teaches genuinely far-sighted avoidance.

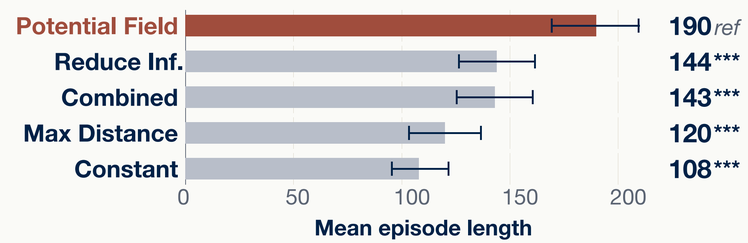


Figure 3. Episode length for the five reward designs; Potential Field is the reference, all others are *** vs PF.

Results: Ablation Study

Knock out direction, movement, or adherence and the policy collapses. Direction and Adherence are load-bearing; the rest are absorbed by the remaining field.

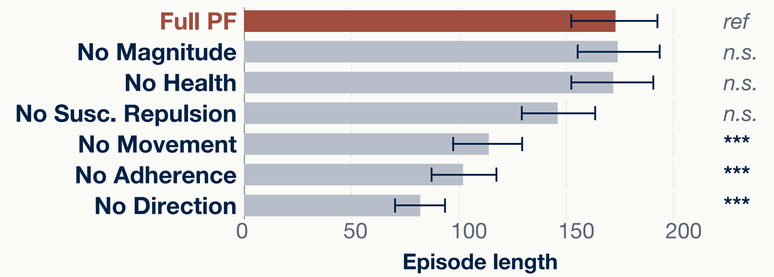


Figure 4. Mean episode length when one Potential Field component is removed at a time. Removing direction, movement, or adherence breaks the policy (***); the rest are n.s.

Results: Partial Observability

Less observation, better policy. A narrower view forces focus on imminent threats; the reward still uses the global state, so the gain is genuine robustness, not leakage.

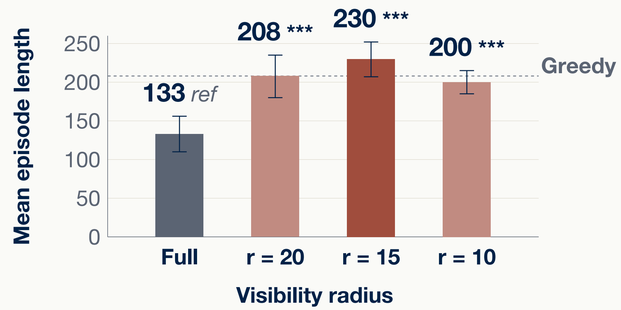


Figure 5. Limited-visibility agents ($r = 10, 15, 20$) outperform full visibility under a privileged-critic reward.

Key References

1. Schulman et al. **PPO**. arXiv:1707.06347, 2017.
2. Pinto et al. **Asymmetric Actor-Critic**. RSS 2018.
3. Ng & Russell. **Policy invariance under reward shaping**. ICML 1999.
4. Feng et al. **IDRLECA: Individual-Level Epidemic Control**. KDD 2023.
5. Manfredi & d'Onofrio (eds.) **Modeling Behaviour and Infectious-Disease Spread**. Springer 2013.
6. Amodei et al. **Concrete Problems in AI Safety**. arXiv:1606.06565, 2016.
7. Shihab et al. **Detecting and Mitigating Reward Hacking**. 2025.