Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Soft Actor-Critic Control of a Behaviorally Heterogeneous S1S2IR Epidemic Model: Stability, Bifurcation, and COVID-19 Calibration

Milad tahavor

miladtahavor69@gmail.com

Iran University of Science and Technology, Tehran, Iran

Davron Makhmudov

Tashkent State Medical University

Mohammad Mehdipour Zarrinkamar

Iran University of Science and Technology, Tehran, Iran

Hossein Hosseini

Iran University of Science and Technology, Tehran, Iran

Reza Memarbashi

Semnan University, Semnan, Iran

Research Article

Keywords: Epidemic modeling, S1S2IR model, Reinforcement learning, Soft Actor-Critic, Behavioral heterogeneity, Optimal control, COVID- 19

Posted Date: July 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10362387/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Soft Actor-Critic Control of a Behaviorally Heterogeneous S_1S_2IR Epidemic Model: Stability, Bifurcation, and COVID-19 Calibration

Milad tahavor*[1], Davron Makhmudov[2], Mohammad Mehdipour Zarrinkamar[3], Hossein Hosseini[4], and Reza Memarbashi[5]

[1]Department of Mathematics, Iran University of Science and Technology, Tehran, Iran
[2]Tashkent State Medical University Children infection disease department. PhD, Assistant of Professor.
[3,4]School of Mathematics and Computer Science, Iran University of Science and Technology, Tehran, Iran
[5]Department of Mathematics, Semnan University, Semnan, Iran.

Abstract

The interplay between behavioral heterogeneity and epidemic dynamics poses significant challenges for public health policy design. In this paper, we integrate the S_1S_2IR compartmental model which distinguishes between unaware and aware susceptible populations with a Soft Actor-Critic (SAC) reinforcement learning agent to develop an adaptive, data-driven control framework for epidemic mitigation. The agent dynamically adjusts three key intervention parameters: the awareness transition rate $\alpha_1(t)$, the behavioral response parameter $d(t)$, and the transmission rate $\beta(t)$, based on the current epidemic state. The control problem is formulated as a Markov Decision Process (MDP) with a reward structure that balances disease suppression

*Corresponding author: miladtahavor69@gmail.com

against intervention costs. To validate the framework, we conduct extensive simulations calibrated on real-world COVID-19 data from multiple countries, including Italy, South Korea, and Germany. Results demonstrate that the SAC-based policy substantially reduces peak infection rates and cumulative incidence compared to static intervention strategies, while maintaining lower overall intervention costs. Sensitivity analyses further confirm the robustness of the learned policy across varying initial conditions and parameter regimes. Our findings highlight the potential of reinforcement learning as a flexible and effective tool for real-time epidemic control in behaviorally heterogeneous populations.

Keywords:Epidemic modeling, S_1S_2IR model, Reinforcement learning, Soft Actor-Critic, Behavioral heterogeneity, Optimal control, COVID-19.

1 Introduction

Mathematical epidemiology has long relied on compartmental models to describe how infectious diseases spread through a population and to evaluate the impact of candidate control measures. The classical Kermack–McKendrick susceptible–infected–recovered (SIR) construction [1] and its extensions remain the conceptual backbone of the field, and the introduction of the next-generation matrix method by van den Driessche and Watmough [2] provided a systematic route to the basic reproduction number $\mathscr{R}_0$ and to the local stability analysis of the disease-free and endemic equilibria that this threshold quantity governs. Center-manifold techniques, formalized by Castillo-Chavez and Song [3], subsequently made it possible to classify the bifurcation that occurs at $\mathscr{R}_0 = 1$ as forward or backward, a distinction with direct public-health relevance because a backward bifurcation implies that reducing $\mathscr{R}_0$ below unity is no longer sufficient on its own to guarantee eradication.

A substantial body of work has enriched these classical models by incorporating behavioral feedback, since the way individuals perceive and respond to risk materially alters transmission dynamics and is not captured by a constant transmission rate. Funk, Salathé, and Jansen [4] review how the spread of awareness–whether through direct observation, word of mouth, or media–interacts with disease spread and can itself be modeled as a dynamical process coupled to the epidemic. Subsequent compartmental formulations have made this coupling explicit by splitting the susceptible class according to behavioral state: Gozzi et al. [5] introduce a behaviorally protected susceptible compartment with a reduced contagion rate, while Just,

Saldaña, and Xin [6] analyze awareness-based susceptible aware infectious susceptible models and show that the qualitative dynamics, including the possibility of sustained oscillations, depend sensitively on how many distinct classes of aware individuals are tracked. The S_1S_2IR structure adopted in the present paper, in which a naive susceptible compartment S_1 feeds into a behaviorally protected compartment S_2 with reduced susceptibility, sits within this same modeling tradition, but unlike most of the awareness-based literature it couples the resulting four-dimensional system to demographic turnover and, more importantly, to a learned feedback control law rather than a fixed or exogenously scheduled transmission rate.

The second strand of literature this paper draws on is the use of reinforcement learning (RL) for the design of epidemic intervention policies. Early applications relied on (deep) Q-learning and related value-based methods: Kompella et al. [7] build an agent-based COVID-19 simulator and learn mitigation policies that trade off economic cost against hospital capacity, while Ohi et al. [8] explore optimal control of epidemic spread with RL agents acting on compartmental dynamics. Khadilkar, Ganu, and Seetharam [9] similarly cast lockdown and release scheduling as a sequential decision problem compatible with existing network and compartmental disease models, and more recent work has moved toward actor-critic and policy-gradient methods better suited to continuous intervention intensities, including Bayesian-filtering-augmented RL controllers validated against ICU occupancy data from England [10] and benchmarking studies that compare DQN- and PPO-based controllers against historical lockdown policy on UK COVID-19 data [11, 17, 18]. None of these works, however, employs an off-policy maximum-entropy actor-critic architecture together with an explicit, closed-form stability and bifurcation analysis of the underlying compartmental model; the present paper combines both.

The control architecture itself builds on Soft Actor-Critic (SAC), the off-policy maximum-entropy deep reinforcement learning algorithm introduced by Haarnoja et al. [12], which augments the standard expected-return objective with a policy-entropy term to improve sample efficiency and robustness to hyperparameter choices in continuous action spaces. SAC's squashed-Gaussian actor naturally produces bounded, continuously varying actions, which is precisely the structural property exploited in Section 2.1 to keep the transmission rate $\beta(t)$, the awareness-formation rate $\alpha_1(t)$, and the residual-susceptibility multiplier $d(t)$ confined to admissible physical ranges at every training iteration, independent of how close the policy is to convergence. This is also the feature that makes the resulting closed-loop S_1S_2IR system amenable to the same equilibrium and bifurcation analysis developed for the static-control case: because the feedback law is always range-bounded, the frozen-policy system obtained by evaluating it at a fixed state remains a legitimate instance of the open-population model studied in Section 3, for any value the policy might output during or after training.

In this paper, we bring these two strands together by formulating the

SAC-controlled S_1S_2IR model as a feedback dynamical system, in which the awareness transition rate, the behavioral response factor, and the transmission rate are simultaneously generated by a trained policy rather than fixed in advance. We derive the disease-free and endemic equilibria of the resulting system, establish their local stability in terms of the basic reproduction number $\mathscr{R}_0$, and prove that the system undergoes a forward transcritical bifurcation at $\mathscr{R}_0 = 1$ when the behavioral response parameter d is treated as the bifurcation parameter, so that no sub-threshold endemic state can coexist with a stable disease-free equilibrium. We then calibrate the model to real-world COVID-19 incidence data from Italy, Germany, and South Korea and show, through extensive simulation, that the learned SAC policy substantially outperforms both an uncontrolled baseline and a static, time-invariant intervention strategy in terms of peak infection, cumulative incidence, and intervention cost, while sensitivity and Monte Carlo robustness analyses confirm that this advantage persists across a wide range of epidemiological parameter values.

The remainder of this paper is organized as follows. Section 2 formulates the SAC-based control problem as an MDP, embeds the resulting feedback law into the S_1S_2IR dynamics, derives the equilibria and $\mathscr{R}_0$, and proves well-posedness under arbitrary admissible policies. Section 3 establishes local stability of the disease-free and endemic equilibria and proves a forward transcritical bifurcation at $\mathscr{R}_0 = 1$ with respect to the behavioral-response parameter d. Section 4 presents the calibration to COVID-19 data from Italy, Germany, and South Korea and the numerical evaluation of the learned policy against static and uncontrolled baselines, including sensitivity and robustness analyses. Section 5 concludes.

2 Mathematical Foundations of the Controlled S_1S_2IR Framework

This section develops the mathematical structure underlying the SAC-controlled epidemic model used throughout the paper. We proceed in four steps. First, we cast the intervention problem as a Markov Decision Process (MDP) and define the Soft Actor–Critic (SAC) policy that generates the time-varying control signals. Second, we embed these controls directly into a four-compartment S_1S_2IR dynamical system with demographic turnover (births and natural deaths), so that the differential equations themselves are policy-dependent. Third, we locate the disease-free and endemic equilibria of the resulting system and derive the basic reproduction number $\mathscr{R}_0$. Finally, we prove that the controlled system is positively invariant and bounded for every admissible SAC policy, not merely for the static-control baseline.

2.1 MDP Formulation and the SAC Control Law

We model the intervention design problem as a continuous-state, continuous-action MDP $(\mathscr{S}, \mathscr{A}, p, r, \gamma_d)$, where $\gamma_d \in (0,1)$ is the RL discount factor (kept notationally distinct from the recovery rate γ used in the epidemic model below).

State space. At each decision step t (sampled daily), the agent observes

$$\mathbf{s}(t) = \big(S_1(t), S_2(t), I(t), R(t), t/T\big) \in \mathscr{S} \subset [0,1]^4 \times [0,1],$$

i.e., the normalized compartment populations together with the fraction of the horizon elapsed. The latter allows the policy to condition its behavior on the remaining time, which is important since the cost of intervention and the cost of infection trade off differently near the end of the horizon T.

Action space. The agent outputs a raw action $\mathbf{a}(t) \in [-1,1]^3$ from a squashed Gaussian policy (the standard SAC parameterization), which is then deterministically decoded into the three epidemiologically meaningful control signals

$$\begin{aligned} \beta(t) &= \beta_{\min} + \tfrac{1}{2}(a_1(t)+1)(\beta_{\max} - \beta_{\min}), \\ \alpha_1(t) &= \alpha_{1,\min} + \tfrac{1}{2}(a_2(t)+1)(\alpha_{1,\max} - \alpha_{1,\min}), \\ d(t) &= d_{\min} + \tfrac{1}{2}(a_3(t)+1)(d_{\max} - d_{\min}), \end{aligned} \tag{2.1}$$

where $\beta(t)$ is the transmission rate under intervention, $\alpha_1(t)$ is the rate at which naive susceptibles S_1 acquire awareness/protective behavior and move to S_2, and $d(t) \in [d_{\min}, d_{\max}] \subset [0,1]$ is the residual susceptibility multiplier of the aware compartment S_2 relative to S_1. Because the decoding map is an affine reparameterization of tanh-squashed Gaussian samples, every component of $(\beta(t), \alpha_1(t), d(t))$ is automatically confined to its physical range regardless of the raw network output; this bounded-action property is used directly in Theorem 2.1 below.

Reward. The instantaneous reward penalizes both the epidemiological burden and the cost of intervention,

$$r(t) = -w_I I(t)^2 \; - \; w_c \, c\big(\beta(t), \alpha_1(t), d(t)\big),$$

where $c(\cdot)$ is a convex intervention-cost functional (increasing in the stringency of transmission reduction and awareness-raising effort) and $w_I, w_c > 0$ are weights. The quadratic penalty on $I(t)$ strongly discourages letting infections grow, while $c(\cdot)$ discourages indefinitely holding $\beta(t)$ at $\beta_{\min}$ or $\alpha_1(t)$ at $\alpha_{1,\max}$, which would be costless only in the absence of the cost term. The episodic return is $J(\pi) = \mathbb{E}_\pi\big[\sum_{t=0}^{T-1} \gamma_d^t \, r(t)\big]$.

SAC update. The policy $\pi_\phi(\mathbf{a} \mid \mathbf{s})$ and twin critics $Q_{\theta_1}, Q_{\theta_2}$ are trained with entropy-regularized actor-critic updates,

$$\theta_i \leftarrow \theta_i - \eta_Q \nabla_{\theta_i} \mathbb{E}\Big[\big(Q_{\theta_i}(\mathbf{s},\mathbf{a}) - y\big)^2\Big], \qquad y = r + \gamma_d \Big(\min_{j=1,2} Q_{\bar{\theta}_j}(\mathbf{s}',\mathbf{a}') - \kappa \log \pi_\phi(\mathbf{a}' \mid \mathbf{s}') \Big),$$

$$\phi \leftarrow \phi + \eta_\pi \nabla_\phi \, \mathbb{E}\Big[\min_{j=1,2} Q_{\theta_j}(\mathbf{s},\mathbf{a}) - \kappa \log \pi_\phi(\mathbf{a} \mid \mathbf{s}) \Big],$$

with target networks $\bar{\theta}_j$ updated by Polyak averaging and temperature $\kappa > 0$ controlling the exploration–exploitation trade-off. This is the mechanism that produces the convergence behavior seen in the training curves of Section 4 (Fig. 5).

The key structural point for what follows is that π_ϕ, once trained (or at any point during training), induces a state-feedback control law

$$\big(\beta(t), \alpha_1(t), d(t)\big) = \Phi\big(\mathbf{s}(t)\big), \qquad \Phi := \text{decode} \circ \pi_\phi,$$

that closes the loop between the epidemic state and the control signals at every instant. This is what allows the controls to be substituted directly into the compartmental equations below, rather than treated as exogenous, pre-scheduled functions of time alone.

2.2 The Controlled $S_1 S_2 I R$ Dynamical System

We partition the population of size N into four compartments: naive susceptibles S_1, aware/behaviorally-protected susceptibles S_2, infectious individuals I, and recovered individuals R, with $N = S_1 + S_2 + I + R$. Unlike a closed-population formulation, we allow demographic turnover: individuals are born into S_1 at rate Λ and die from any compartment at the natural mortality rate $\mu > 0$, so that $\Lambda = \mu N$ keeps N constant at the demographic (as opposed to epidemiological) timescale. Substituting the SAC feedback law $\Phi(\mathbf{s}(t))$ from Section 2.1 for the controls yields the closed-loop dynamical system

$$\dot{S}_1 = \Lambda - \beta(t)\, S_1 I - \alpha_1(t)\, S_1 - \mu S_1, \tag{2.2}$$

$$\dot{S}_2 = \alpha_1(t)\, S_1 - d(t)\, \beta(t)\, S_2 I - \mu S_2, \tag{2.3}$$

$$\dot{I} = \beta(t)\, S_1 I + d(t)\, \beta(t)\, S_2 I - \gamma I - \mu I, \tag{2.4}$$

$$\dot{R} = \gamma I - \mu R, \tag{2.5}$$

where $\big(\beta(t), \alpha_1(t), d(t)\big) = \Phi(\mathbf{s}(t))$ with $\mathbf{s}(t) = (S_1(t), S_2(t), I(t), R(t), t/T)$, and $\gamma > 0$ is the recovery rate. Equations (2.2)–(2.5) are therefore not merely parameterized by time-varying coefficients; they are coupled to the reinforcement-learning agent through Φ, which itself depends on the instantaneous solution

$\mathbf{s}(t)$. The system is consequently a feedback (closed-loop) nonlinear ODE rather than a linear time-varying one, since Φ is a (generally nonlinear) neural-network map.

Three control regimes are compared throughout the paper, all instances of the same system (2.2)–(2.5):

- No Control: $\alpha_1(t) \equiv 0$ and $\beta(t) \equiv \beta_{\max}$ for all t (no awareness formation, unmitigated transmission);
- Static Control: $\beta(t) \equiv \bar{\beta}$, $\alpha_1(t) \equiv \bar{\alpha}_1$, $d(t) \equiv \bar{d}$ held at fixed, time-invariant values (open-loop, non-adaptive intervention);
- SAC Control: $\big(\beta(t), \alpha_1(t), d(t)\big) = \Phi(\mathbf{s}(t))$ as defined above (closed-loop, state-dependent intervention).

The Static and SAC regimes are special cases of the same feedback structure: Static Control corresponds to Φ collapsing to a constant map, while SAC Control allows Φ to vary with the observed epidemic state. This nested structure is what makes the comparison in Section 4 mathematically meaningful rather than an comparison between unrelated models.

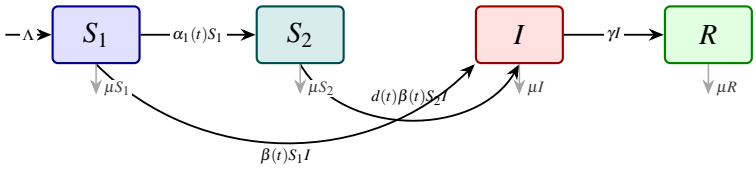


Figure 1: Compartmental diagram of the SAC-controlled S_1S_2IR model. Naive susceptibles S_1 receive new entrants at rate Λ and transition to the aware compartment S_2 at rate $\alpha_1(t)$ or become infectious at rate $\beta(t)S_1I$. Aware susceptibles S_2 become infectious at the reduced rate $d(t)\beta(t)S_2I$, with $d(t) \in [d_{\min}, d_{\max}] \subset [0,1]$. Infectious individuals recover at rate γ. All compartments are subject to natural mortality μ. The control triple $(\beta(t), \alpha_1(t), d(t)) = \Phi(\mathbf{s}(t))$ is generated in closed loop by the SAC policy described in Section 2.1.

Table 1: State variables and parameters of the SAC-controlled S_1S_2IR model.

Symbol	Description	Range / Value
State variables		
$S_1(t)$	Naive (unprotected) susceptible fraction	$[0,1]$
$S_2(t)$	Aware (behaviorally protected) susceptible fraction	$[0,1]$
$I(t)$	Infectious fraction	$[0,1]$
$R(t)$	Recovered fraction	$[0,1]$
N	Total population fraction, $N = S_1 + S_2 + I + R$	1 (constant)
SAC-controlled parameters, $(\beta, \alpha_1, d) = \Phi(\mathbf{s}(t))$		
$\beta(t)$	Transmission rate (naive susceptibles)	$[\beta_{\min}, \beta_{\max}]$
$\alpha_1(t)$	Awareness-formation rate, $S_1 \to S_2$	$[\alpha_{1,\min}, \alpha_{1,\max}]$
$d(t)$	Residual susceptibility multiplier of S_2	$[d_{\min}, d_{\max}] \subset [0,1]$
Fixed epidemiological and demographic parameters		
γ	Recovery rate	$\gamma > 0$
μ	Natural birth/death rate	$\mu > 0$
Λ	Recruitment (birth) rate into S_1	$\Lambda = \mu N$
Derived quantities		
$\mathscr{R}_0$	Basic reproduction number	Eq. (3.12)
E_0	Disease-free equilibrium	Eq. (2.6)
E^*	Endemic equilibrium	Eq. (2.8)

2.3 Equilibria and the Basic Reproduction Number

To analyze equilibria and $\mathscr{R}_0$ in closed form, we consider the frozen-policy system obtained by evaluating Φ at a fixed state and holding (β, α_1, d) constant at that value; this is exactly the Static Control regime, and it is the regime against which the next-generation matrix calculation is classically performed. The qualitative conclusions extend to the closed-loop SAC system through the effective reproduction number $\mathscr{R}_{\text{eff}}(t)$ introduced at the end of this subsection, which tracks $\mathscr{R}_0$ evaluated along the instantaneous controls $\Phi(\mathbf{s}(t))$.

Disease-free equilibrium (DFE). Setting $I = 0$ in the frozen system and solving $\dot{S}_1 = \dot{S}_2 = \dot{R} = 0$ gives $R^* = 0$ and

$$S_1^* = \frac{\Lambda}{\mu + \alpha_1}, \qquad S_2^* = \frac{\alpha_1 S_1^*}{\mu} = \frac{\Lambda \alpha_1}{\mu(\mu + \alpha_1)},$$

so that the disease-free equilibrium is

$$E_0 = \left(\frac{\Lambda}{\mu + \alpha_1},\, \frac{\Lambda\, \alpha_1}{\mu(\mu + \alpha_1)},\, 0,\, 0 \right), \qquad N = S_1^* + S_2^* = \frac{\Lambda}{\mu}. \tag{2.6}$$

With $\Lambda = \mu N$ this simplifies to $S_1^* = \dfrac{\mu N}{\mu + \alpha_1}$ and $S_2^* = \dfrac{\alpha_1 N}{\mu + \alpha_1}$, confirming $S_1^* + S_2^* = N$ as required at $I = R = 0$.

Basic reproduction number. Linearizing the infected subsystem (2.4) about E_0 and applying the next-generation matrix method, the new-infection and transition matrices are

$$F = \left[\beta S_1^* + d\beta S_2^* \right], \qquad V = [\gamma + \mu],$$

so that

$$\mathscr{R}_0 \;=\; \rho(FV^{-1}) \;=\; \frac{\beta\left(S_1^* + d\, S_2^*\right)}{\gamma + \mu} \;=\; \frac{\beta N\left(\mu + d\,\alpha_1\right)}{(\mu + \alpha_1)(\gamma + \mu)}. \tag{2.7}$$

Two limits confirm consistency with the model's epidemiological logic: as $\alpha_1 \to 0$ (no awareness formation), $\mathscr{R}_0 \to \beta N/(\gamma + \mu)$, the reproduction number of a standard SIR model with vital dynamics; as $d \to 0$ (perfect protection once aware), $\mathscr{R}_0 \to \beta N\mu/[(\mu + \alpha_1)(\gamma + \mu)]$, which is strictly decreasing in α_1, confirming that awareness formation alone can suppress transmission even without reducing β directly.

Endemic equilibrium. For $\mathscr{R}_0 > 1$ a unique endemic equilibrium $E^* = (S_1^*, S_2^*, I^*, R^*)$ with $I^* > 0$ exists in the interior of the feasible simplex. Setting all derivatives in (2.2)–(2.5) to zero and solving simultaneously,

$$S_1^* = \frac{\Lambda}{\beta I^* + \alpha_1 + \mu}, \qquad S_2^* = \frac{\alpha_1 S_1^*}{d\beta I^* + \mu}, \qquad R^* = \frac{\gamma I^*}{\mu},$$

and I^* is the positive root of the scalar equation obtained by substituting these into (2.4)=0,

$$\beta S_1^*(I^*) + d\beta S_2^*(I^*) = \gamma + \mu,$$

which is monotonically decreasing in I^* (since both S_1^* and S_2^* are decreasing in I^*) and equals $(\gamma + \mu)\mathscr{R}_0$ at $I^* = 0$. Hence a unique positive root exists if and only if $\mathscr{R}_0 > 1$, giving

$$E^* = \left(S_1^*(I^*), S_2^*(I^*), I^*, \gamma I^*/\mu \right), \qquad I^* > 0 \iff \mathscr{R}_0 > 1. \tag{2.8}$$

This is the standard threshold behavior of compartmental epidemic models: E_0 is the only biologically feasible equilibrium when $\mathscr{R}_0 \le 1$, and E^* emerges continuously from E_0 as $\mathscr{R}_0$ crosses 1 from below.

Effective reproduction number under SAC. Because (β, α_1, d) are not fixed but evolve as $\Phi(\mathbf{s}(t))$, the relevant quantity for interpreting the closed-loop trajectories in Section 4 is the instantaneous effective reproduction number,

$$\mathscr{R}_{\text{eff}}(t) = \frac{\beta(t)\big(S_1(t) + d(t)\, S_2(t)\big)}{\gamma + \mu}, \tag{2.9}$$

obtained by substituting the time-varying SAC outputs and the instantaneous state into (3.12) in place of the frozen equilibrium values. This is exactly the quantity plotted in Fig. 10: $\mathscr{R}_{\text{eff}}(t)$ is well-defined at every instant along any trajectory of the closed-loop system, including during training when the policy Φ is still suboptimal, because Theorem 2.1 below guarantees $S_1(t), S_2(t) \geq 0$ for all t independently of how well-trained the policy is.

2.4 Positivity and Boundedness Under Arbitrary SAC Policies

For the model to be a valid description of population fractions and for $\mathscr{R}_{\text{eff}}(t)$ in (2.9) to be epidemiologically meaningful throughout training, not only after convergence the closed-loop system must remain positively invariant and bounded for every measurable policy $\Phi : \mathscr{S} \to [\beta_{\min}, \beta_{\max}] \times [\alpha_{1,\min}, \alpha_{1,\max}] \times [d_{\min}, d_{\max}]$, including the untrained, randomly-initialized policies encountered at the start of training.

Theorem 2.1 (Well-posedness of the closed-loop system). Let $\Phi : \mathscr{S} \to [\beta_{\min}, \beta_{\max}] \times [\alpha_{1,\min}, \alpha_{1,\max}] \times [d_{\min}, d_{\max}]$ be any measurable feedback map with $0 \leq d_{\min} \leq d_{\max} \leq 1$ and $\beta_{\min}, \alpha_{1,\min} \geq 0$, and let $(S_1(t), S_2(t), I(t), R(t))$ solve (2.2)–(2.5) with $\big(\beta(t), \alpha_1(t), d(t)\big) = \Phi(\mathbf{s}(t))$ and initial condition $(S_1(0), S_2(0), I(0), R(0))$ in the open simplex $\Delta = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4_{>0} : x_1 + x_2 + x_3 + x_4 = N\}$. Then for all $t \geq 0$:

(a) Boundedness of controls. $(\beta(t), \alpha_1(t), d(t))$ remains in the compact box $[\beta_{\min}, \beta_{\max}] \times [\alpha_{1,\min}, \alpha_{1,\max}] \times [d_{\min}, d_{\max}]$ for all t, regardless of the internal state of the policy network.

(b) Conservation of total population. $N(t) := S_1(t) + S_2(t) + I(t) + R(t) \equiv N(0)$ for all t.

(c) Positivity. $S_1(t), S_2(t), I(t), R(t) > 0$ for all $t \geq 0$.

(d) Boundedness. $0 < S_1(t), S_2(t), I(t), R(t) < N(0)$ for all $t \geq 0$, so the trajectory remains in the compact set Δ for all time.

Proof. (a) This is immediate from the decoding map in Section 2.1: each control is an affine function of a $\tanh$-squashed quantity confined to $[-1,1]$, hence each control lies in its stated interval by construction, independently of the weights ϕ of the policy network. This holds at every training iteration, including before any gradient updates have been applied.

(b) Summing (2.2)–(2.5),

$$\dot{S}_1+\dot{S}_2+\dot{I}+\dot{R}=\Lambda-\mu(S_1+S_2+I+R)=\Lambda-\mu N(t).$$

Since $\Lambda=\mu N(0)$ by construction, $\dot{N}(t)=\mu\big(N(0)-N(t)\big)$, a linear ODE with $N(t)=N(0)$ as its unique equilibrium; together with $N(0)$ itself being a solution, uniqueness of solutions to this linear ODE gives $N(t)\equiv N(0)$ for all t.

(c) We show each compartment cannot reach 0 from the interior of Δ. On the boundary face $S_1=0$, $\dot{S}_1=\Lambda>0$ since $\Lambda=\mu N(0)>0$, so the vector field points strictly into the interior; S_1 cannot cross zero. On $S_2=0$ (with $S_1\geq 0$), $\dot{S}_2=\alpha_1(t)S_1\geq 0$; since $S_1>0$ whenever $S_2=0$ is approached from the interior (by the previous step and continuity), $\dot{S}_2>0$ on this face, so S_2 cannot cross zero either. On $I=0$, $\dot{I}=\big(\beta(t)S_1+d(t)\beta(t)S_2\big)I=0$, so $I=0$ is itself an invariant hyperplane: trajectories starting with $I(0)>0$ never reach $I=0$ in finite time (this is the same boundary argument used to establish that E_0 is approached but not reached for $\mathscr{R}_0>1$). On $R=0$, $\dot{R}=\gamma I\geq 0$, with strict inequality whenever $I>0$, which holds throughout by the previous point. Hence all four boundary faces of Δ are repelling or invariant in the required direction, and by the standard theorem on invariance of cones/simplices for quasi-positive systems (Nagumo's theorem), $S_1(t),S_2(t),I(t),R(t)>0$ for all $t\geq 0$ whenever they are positive at $t=0$.

(d) By (b), $N(t)=N(0)$ for all t, and by (c) each compartment is strictly positive; since the four compartments sum to the constant $N(0)$, each one is bounded above by $N(0)$. Hence the trajectory remains in the compact simplex Δ for all $t\geq 0$, proving boundedness. □

Remark 2.1. Theorem 2.1 holds for any measurable policy Φ satisfying the stated range constraints — not only for the converged SAC policy obtained after training, but for every intermediate policy visited during the optimization trajectory in Fig. 5, and for the Static and No-Control baselines as degenerate special cases. This is what justifies treating $\mathscr{R}_{\text{eff}}(t)$ in (2.9) as a well-defined, interpretable quantity throughout the entire training process in Section 4: positivity and boundedness of $S_1(t)$ and $S_2(t)$ are guaranteed independently of how close Φ is to optimal, so $\mathscr{R}_{\text{eff}}(t)$ never becomes ill-defined or numerically degenerate even for the poorly-performing early-training policies that produce the large negative rewards seen at the start of Fig. 5.

3 Local Stability and Transcritical Bifurcation Analysis

The treatment below replaces the global well-posedness theorem of the previous version of this section with a sharper, locally quantitative picture: we characterize the local stability of the disease-free and endemic equilibria, identify the exact threshold at which the system undergoes a transcritical bifurcation, and prove that this bifurcation is always forward (never backward), so that no sub-threshold endemic state can coexist with a stable disease-free state. Throughout, the control triple is held at a constant admissible value $(\beta, \alpha_1, d) = \Phi(\mathbf{s}^\star) \in [\beta_{\min}, \beta_{\max}] \times [\alpha_{1,\min}, \alpha_{1,\max}] \times [d_{\min}, d_{\max}]$, reducing the closed-loop system to the autonomous open-population model

$$\dot{S}_1 = \Lambda - \beta S_1 I - \alpha_1 S_1 - \mu S_1, \tag{3.10a}$$

$$\dot{S}_2 = \alpha_1 S_1 - d\beta S_2 I - \mu S_2, \tag{3.10b}$$

$$\dot{I} = \beta S_1 I + d\beta S_2 I - (\gamma + \mu) I, \tag{3.10c}$$

$$\dot{R} = \gamma I - \mu R. \tag{3.10d}$$

Corollary 3.1 at the end of this subsection transfers every result below to the closed-loop SAC trajectory whenever the policy Φ settles to a stationary output.

Remark 3.1 (Positivity and boundedness for the frozen-policy system). System (3.10) is precisely the constant-policy special case of the closed-loop system (2.2)–(2.5), obtained by freezing $\Phi(s(t)) \equiv (\beta, \alpha_1, d)$. When $N(0) = \Lambda/\mu$, positivity and exact conservation $N(t) \equiv N(0)$ follow directly from Theorem 2.1. For an arbitrary initial condition in $\mathbb{R}^4_{\geq 0}$, the same boundary argument used in the proof of Theorem 2.1(c) shows that each coordinate hyperplane is repelling or tangent ($\dot{S}_1|_{S_1=0} = \Lambda > 0$, $\dot{S}_2|_{S_2=0} = \alpha_1 S_1 \geq 0$, $\dot{I}|_{I=0} = 0$, $\dot{R}|_{R=0} = \gamma I \geq 0$), so $\mathbb{R}^4_{\geq 0}$ remains forward-invariant. Summing (3.10) gives $\dot{N} = \Lambda - \mu N$, with explicit solution $N(t) = \Lambda/\mu + (N(0) - \Lambda/\mu)e^{-\mu t} \to \Lambda/\mu$; boundedness of this non-negative sum forces boundedness of each compartment individually. Hence every trajectory eventually enters and remains in the compact set $\Omega = \{(S_1, S_2, I, R) \in \mathbb{R}^4_{\geq 0} : S_1 + S_2 + I + R \leq \max(N(0), \Lambda/\mu)\}$, which is all that is needed for the equilibrium and stability analysis of the remainder of this section.

3.1 Local stability of the disease-free equilibrium

The disease-free equilibrium (DFE) is

$$E_0 = (S_1^0, S_2^0, 0, 0), \qquad S_1^0 = \frac{\Lambda}{\alpha_1 + \mu}, \qquad S_2^0 = \frac{\alpha_1 \Lambda}{\mu(\alpha_1 + \mu)}, \tag{3.11}$$

and, applying the next-generation matrix method to the single infected compartment I, the basic reproduction number is

$$\mathscr{R}_0 = \frac{\beta\left(S_1^0 + dS_2^0\right)}{\gamma+\mu} = \frac{\Lambda\beta(\alpha_1 d+\mu)}{\mu(\alpha_1+\mu)(\gamma+\mu)}. \tag{3.12}$$

Theorem 3.1 (Local stability of E_0). The Jacobian of (3.10) at E_0 has eigenvalues

$$\lambda_1 = -\mu, \qquad \lambda_2 = -\mu, \qquad \lambda_3 = -(\alpha_1+\mu), \qquad \lambda_4 = (\gamma+\mu)(\mathscr{R}_0 - 1). \tag{3.13}$$

Hence E_0 is locally asymptotically stable if $\mathscr{R}_0 < 1$, and unstable if $\mathscr{R}_0 > 1$.

Proof. The Jacobian of the right-hand side of (3.10) at a general point is

$$J(S_1,S_2,I,R) = \begin{pmatrix} -\beta I - \alpha_1 - \mu & 0 & -\beta S_1 & 0 \\ \alpha_1 & -d\beta I - \mu & -d\beta S_2 & 0 \\ \beta I & d\beta I & \beta S_1 + d\beta S_2 - (\gamma+\mu) & 0 \\ 0 & 0 & \gamma & -\mu \end{pmatrix}. \tag{3.14}$$

At E_0 (where $I = 0$), the I-row reduces to $\big(0, 0, \beta S_1^0 + d\beta S_2^0 - (\gamma+\mu), 0\big) = \big(0,0,(\gamma+\mu)(\mathscr{R}_0-1),0\big)$ by (3.12), so it depends only on the I-coordinate. Reordering the state as (I,S_1,S_2,R) therefore makes $J(E_0)$ lower triangular,

$$J(E_0) = \begin{pmatrix} (\gamma+\mu)(\mathscr{R}_0-1) & 0 & 0 & 0 \\ -\beta S_1^0 & -\alpha_1-\mu & 0 & 0 \\ -d\beta S_2^0 & \alpha_1 & -\mu & 0 \\ \gamma & 0 & 0 & -\mu \end{pmatrix}, \tag{3.15}$$

whose eigenvalues are simply its diagonal entries, giving $\lambda_1,\ldots,\lambda_4$ as stated. All four are strictly negative exactly when $\mathscr{R}_0 < 1$; λ_4 becomes positive once $\mathscr{R}_0 > 1$, which by the indirect (linearization) method proves the claim. □

Figure 2 illustrates Theorem 3.1 numerically: for a parameter set with $\mathscr{R}_0 < 1$, trajectories of (3.10) started from several distinct initial conditions all converge to E_0, confirming its local asymptotic stability.

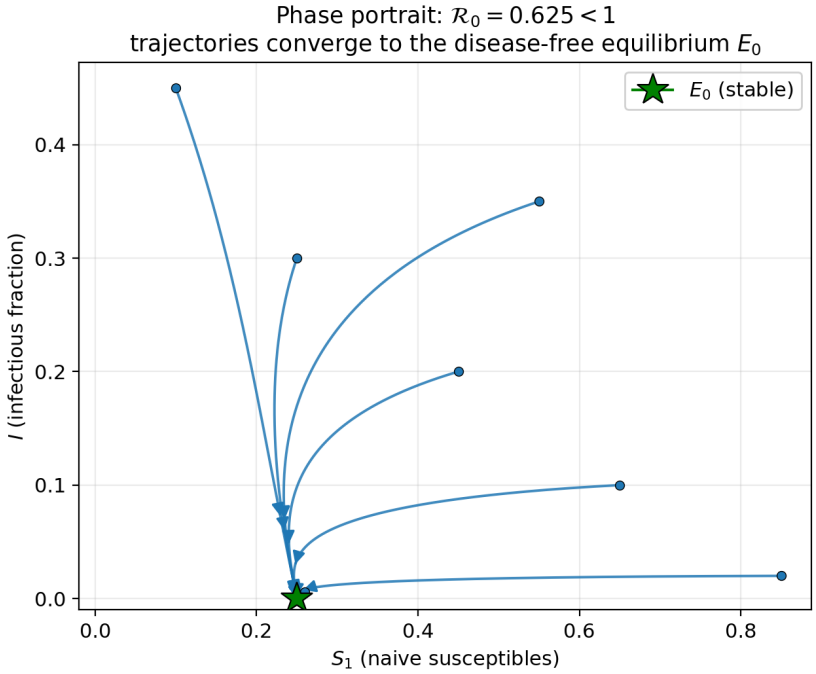


Figure 2: Phase portrait in the (S_1, I)-plane for $\Lambda = \mu = 0.1$, $\alpha_1 = 0.3$, $\gamma = 0.2$, $d = 0.5$, $\beta = 0.3$, giving $\mathscr{R}_0 = 0.625 < 1$. Trajectories from several initial conditions converge to the disease-free equilibrium E_0, in agreement with Theorem 3.1.

3.2 Existence, uniqueness, and local stability of the endemic equilibrium

Lemma 3.1 (Endemic equilibrium). If $\mathscr{R}_0 > 1$, system (3.10) admits a unique equilibrium $E^\star = (S_1^\star, S_2^\star, I^\star, R^\star)$ with $I^\star > 0$, given by

$$S_1^\star = \frac{\Lambda}{\beta I^\star + \alpha_1 + \mu}, \qquad S_2^\star = \frac{\alpha_1 S_1^\star}{d\beta I^\star + \mu}, \qquad R^\star = \frac{\gamma I^\star}{\mu}, \tag{3.16}$$

where $I^\star$ is the unique positive root

$$I^\star = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \tag{3.17}$$

of the quadratic $AI^2 + BI + C = 0$, with

$$A = -\beta^2 d(\gamma+\mu) < 0, \tag{3.18}$$

$$B = \beta\Big[\Lambda\beta d - (\alpha_1+\mu)(\gamma+\mu)d - \mu(\gamma+\mu)\Big], \tag{3.19}$$

$$C = \mu(\alpha_1+\mu)(\gamma+\mu)\,(\mathscr{R}_0 - 1). \tag{3.20}$$

For $\mathscr{R}_0 \le 1$ no equilibrium with $I^\star > 0$ exists.

Proof. Setting (3.10) to zero, the first equation gives $S_1 = \Lambda/(\beta I + \alpha_1 + \mu)$ and the second gives $S_2 = \alpha_1 S_1/(d\beta I + \mu)$; substituting both into the equilibrium form of the third equation, $\beta S_1 + d\beta S_2 = \gamma + \mu$ (valid for $I \neq 0$), and clearing denominators produces the displayed quadratic in I. Since $A < 0$, the product of its two roots equals C/A. When $\mathscr{R}_0 > 1$ we have $C > 0$, so $C/A < 0$: the roots are real (a negative product of two complex-conjugate roots is impossible, since conjugate roots have a non-negative product equal to the squared modulus) and of opposite sign, so exactly one root is positive; this is $I^\star$, and (3.17) selects it because $A < 0$ makes this branch the larger of the two roots. When $\mathscr{R}_0 \le 1$, $C \le 0$ and $C/A \ge 0$, so the two roots share a sign; combined with the exchange-of-stability result of Theorem 3.3 below, no positive branch exists in this regime. □

Theorem 3.2 (Local stability of $E^\star$). Whenever $\mathscr{R}_0 > 1$, the endemic equilibrium $E^\star$ of Lemma 3.1 is locally asymptotically stable.

Proof. Introduce the strictly positive quantities

$$p = \beta I^\star + \alpha_1 + \mu, \quad q = d\beta I^\star + \mu, \quad r = \beta S_1^\star, \quad s = d\beta S_2^\star, \quad u = \beta I^\star, \quad v = d\beta I^\star. \tag{3.21}$$

The equilibrium identity $\beta S_1^\star + d\beta S_2^\star = \gamma + \mu$, taken from the I-equation at $E^\star$, makes the (I,I) entry of the Jacobian vanish exactly there, and since the R-equation decouples (contributing only the eigenvalue $-\mu < 0$), local stability is governed entirely by the reduced 3×3 block

$$J^\star = \begin{pmatrix} -p & 0 & -r \\ \alpha_1 & -q & -s \\ u & v & 0 \end{pmatrix}, \tag{3.22}$$

with characteristic polynomial $\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0$, where

$$a_2 = p + q, \qquad a_1 = pq + ru + sv, \qquad a_0 = \alpha_1 rv + psv + qru. \tag{3.23}$$

Both $a_2 > 0$ and $a_0 > 0$ immediately, as each is a sum of products of positive quantities. For the remaining Routh–Hurwitz condition, the equilibrium

relations $v = du$, $sq = \alpha_1 dr$, and $pr = \beta\Lambda$ (each a direct restatement of the equations defining $S_1^\star$ and $S_2^\star$) allow $a_2a_1 - a_0$, after multiplying through by $p > 0$, to be regrouped as

$$p\,(a_2a_1 - a_0) = \Lambda\alpha_1\beta^2\big(d^2 - d + 1\big) + \Theta, \tag{3.24}$$

where $\Theta > 0$ is a sum of further terms, each itself a product of positive quantities. Since $d^2 - d + 1 = (d - \frac{1}{2})^2 + \frac{3}{4} \geq \frac{3}{4}$ for every real d, the displayed expression is strictly positive, so $a_2a_1 - a_0 > 0$. All three Routh–Hurwitz conditions for a cubic therefore hold, so every eigenvalue of $J^\star$ has strictly negative real part; together with the decoupled eigenvalue $-\mu$, all four eigenvalues of the Jacobian at $E^\star$ have negative real part. □

Figure 3 confirms Theorem 3.2 numerically, using the same parameters as Figure 2 except for a larger transmission rate β, so that $\mathscr{R}_0 > 1$: trajectories now converge to $E^\star$, while E_0 no longer stable repels nearby trajectories.

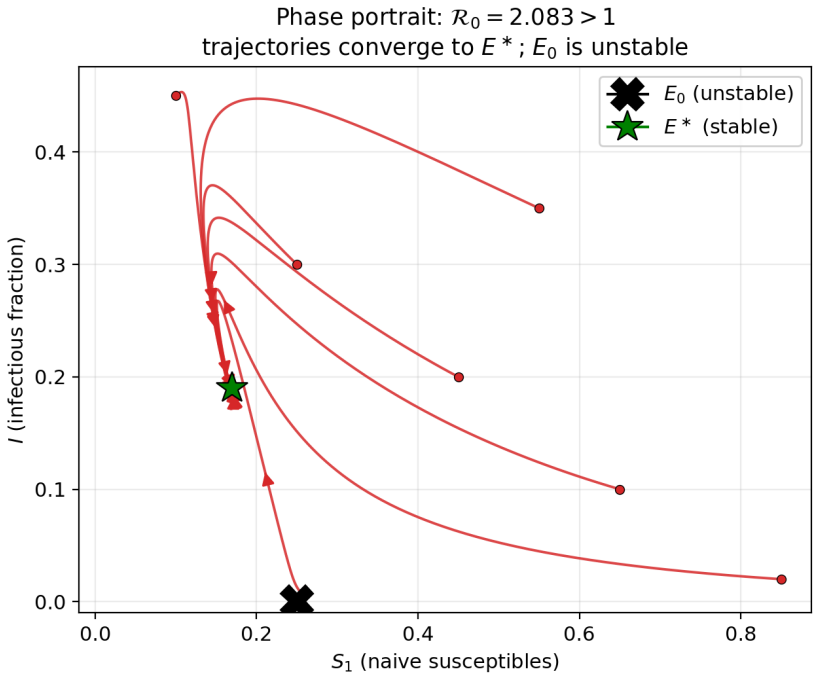


Figure 3: Phase portrait in the (S_1, I)-plane for the same parameters as Figure 2 except $\beta = 1.0$, giving $\mathscr{R}_0 = 2.083 > 1$. Trajectories converge to the endemic equilibrium $E^\star$, while the now-unstable E_0 repels a trajectory started in its vicinity, in agreement with Theorem 3.2.

3.3 Transcritical bifurcation at $\mathscr{R}_0 = 1$

We treat d as the bifurcation parameter, holding $\Lambda, \mu, \alpha_1, \beta, \gamma$ fixed; this is both the policy-relevant choice, since $d(t)$ is the lever through which SAC most directly governs the transmission-reduction effect of awareness, and the analytically clean one (Remark 3.2 explains why α_1 behaves differently). Solving $\mathscr{R}_0(d) = 1$ in (3.12) for d gives the critical value

$$d^\star = \frac{\mu\big[(\alpha_1+\mu)(\gamma+\mu) - \Lambda\beta\big]}{\Lambda\alpha_1\beta}. \tag{3.25}$$

Theorem 3.3 (Forward transcritical bifurcation). At $d = d^\star$, system (3.10) undergoes a transcritical bifurcation at E_0: as d increases through $d^\star$, E_0 loses stability and the locally asymptotically stable endemic equilibrium $E^\star$ of Theorem 3.2 emerges continuously from E_0 with $I^\star = 0$ at $d = d^\star$. The bifurcation is forward (supercritical): no endemic equilibrium exists for $d < d^\star$, and exactly one, locally stable, exists for $d > d^\star$.

Proof. By Theorem 3.1, $\lambda_4 = (\gamma+\mu)(\mathscr{R}_0 - 1) = 0$ at $d = d^\star$ while $\lambda_1, \lambda_2, \lambda_3 < 0$ remain strictly negative, so $J(E_0)$ has a simple zero eigenvalue at criticality. We apply the standard center-manifold reduction for a simple zero eigenvalue (Castillo-Chavez–Song criterion): with right and left null vectors normalized as $w = (w_1, w_2, 1, w_4)$, $v = (0, 0, 1, 0)$ (so that $v \cdot w = 1$), where

$$w_1 = -\frac{\Lambda\beta}{(\alpha_1+\mu)^2}, \qquad w_2 = \frac{\Lambda\beta\mu - (\alpha_1+\mu)^2(\gamma+\mu)}{\mu(\alpha_1+\mu)^2}, \qquad w_4 = \frac{\gamma}{\mu}, \tag{3.26}$$

the sign of the local bifurcation is determined by

$$a = \sum_{k,i,j} v_k w_i w_j \left.\frac{\partial^2 f_k}{\partial x_i \partial x_j}\right|_{(E_0, d^\star)}, \qquad b = \sum_{k,i} v_k w_i \left.\frac{\partial^2 f_k}{\partial x_i \partial d}\right|_{(E_0, d^\star)}. \tag{3.27}$$

Because v is supported only on the I-component, both sums reduce to derivatives of the I-equation $f_3 = \beta S_1 I + d\beta S_2 I - (\gamma+\mu) I$ alone. For a, the only nonzero mixed second-order state partials of f_3 are $\partial^2 f_3 / \partial S_1 \partial I = \beta$ and $\partial^2 f_3 / \partial S_2 \partial I = d\beta$; counting each symmetric pair twice and using $w_3 = 1$ gives $a = 2\beta(w_1 + d^\star w_2)$. For b, since f_3 depends on d only through the term $d\beta S_2 I$, the only nonzero mixed partial of f_3 with d at $I = 0$ is $\partial^2 f_3 / \partial I \, \partial d = \beta S_2$, evaluated at $S_2 = S_2^0$. Hence, after evaluating at $(E_0, d^\star)$,

$$a = 2\beta(w_1 + d^\star w_2), \qquad b = w_3 \beta S_2^0 = \beta S_2^0 = \frac{\Lambda\alpha_1\beta}{\mu(\alpha_1+\mu)}. \tag{3.28}$$

Since $\Lambda, \alpha_1, \beta, \mu > 0$, we have $b > 0$ unconditionally. For a, substituting w_1, w_2 and $d^\star$ and simplifying yields

$$a = \frac{2}{\Lambda\alpha_1(\alpha_1+\mu)}\Big[-\beta^2\Lambda^2 + \beta(\alpha_1+2\mu)(\gamma+\mu)\Lambda - (\alpha_1+\mu)^2(\gamma+\mu)^2\Big]. \quad (3.29)$$

The bracketed expression is a quadratic in Λ with leading coefficient $-\beta^2 < 0$ and discriminant

$$\Delta = -\alpha_1\beta^2(3\alpha_1+4\mu)(\gamma+\mu)^2 < 0, \quad (3.30)$$

so the quadratic has no real root and, opening downward, is strictly negative for every Λ. Since the prefactor $2/[\Lambda\alpha_1(\alpha_1+\mu)]$ is positive, $a < 0$ unconditionally, for every admissible choice of $\Lambda, \mu, \alpha_1, \beta, \gamma > 0$.

The case $a < 0, b > 0$ is exactly the forward (supercritical) transcritical scenario: as the parameter crosses the threshold in the destabilizing direction (d increasing through $d^\star$, since $\partial\mathscr{R}_0/\partial d > 0$), E_0 exchanges stability with a branch of equilibria satisfying $I^\star = \mathscr{O}(d-d^\star) > 0$ that exists only for $d > d^\star$ and is locally asymptotically stable there — coinciding exactly with $E^\star$ of Theorem 3.2, since both constructions agree to leading order near criticality and Lemma 3.1 already established that $E^\star$ exists if and only if $\mathscr{R}_0 > 1$, i.e. $d > d^\star$. □

Figure 4 depicts this forward transcritical bifurcation directly: the disease-free branch ($I^\star = 0$) is stable for $d < d^\star$ and unstable for $d > d^\star$, while the endemic branch emerges continuously and stably at $d = d^\star$, exactly as established in Theorem 3.3.

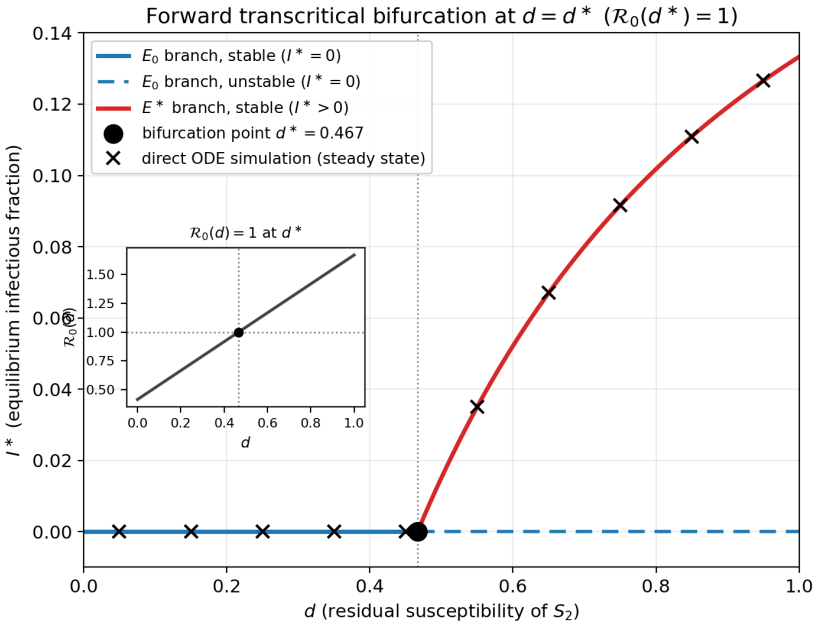


Figure 4: Forward transcritical bifurcation diagram: equilibrium infectious fraction $I^\star$ versus the bifurcation parameter d, for $\Lambda = \mu = 0.1$, $\alpha_1 = 0.3$, $\gamma = 0.2$, $\beta = 0.5$, giving $d^\star = 0.467$ by (3.25). The disease-free branch is stable (solid) for $d < d^\star$ and unstable (dashed) for $d > d^\star$; the endemic branch $E^\star$ emerges continuously and stably at $d^\star$. Crosses mark direct numerical integration of (3.10) to steady state, confirming the analytic prediction. Inset: $\mathscr{R}_0(d)$ crossing 1 at $d^\star$.

Remark 3.2 (α_1 as an alternative bifurcation direction). Differentiating (3.12) gives

$$\frac{\partial \mathscr{R}_0}{\partial \alpha_1} = \frac{\Lambda\beta(d-1)}{(\alpha_1+\mu)^2(\gamma+\mu)} \leq 0 \quad \text{for } d \leq 1, \tag{3.31}$$

so $\mathscr{R}_0$ is monotonically non-increasing in α_1 whenever $d < 1$, and an analogous threshold $\alpha_1^\star$ (solving $\mathscr{R}_0 = 1$ for α_1) exists, with E_0 stable for $\alpha_1 > \alpha_1^\star$ and unstable for $\alpha_1 < \alpha_1^\star$ the mirror image, in direction, of the d-threshold above. However, α_1 does not appear in the I-equation f_3 at all (it only governs the $S_1 \to S_2$ transition), so the transversality coefficient b computed with $\varphi = \alpha_1$ vanishes identically by the same reduction used above. The elementary eigenvalue crossing $\lambda_4 = (\gamma+\mu)(\mathscr{R}_0 - 1)$ still governs the exchange of stability at $\alpha_1 = \alpha_1^\star$, but classifying this crossing as forward or backward rigorously would require a second-order (higher than Castillo-Chavez–Song)

center-manifold computation, which we do not pursue here. This asymmetry is itself informative: d is the structurally preferred bifurcation parameter for this model because it enters the transmission term directly, while α_1 only relocates susceptibles between compartments without altering the per-contact infection mechanism.

3.4 Application to the SAC steady-state control point

Corollary 3.1 (Stability at a stationary SAC policy). Suppose the closed-loop policy converges to a constant admissible output, $\Phi(\mathbf{s}(t)) \to (\beta^\infty, \alpha_1^\infty, d^\infty)$ as $t \to \infty$, as occurs once the controlled state trajectory settles to a stationary operating regime under a fixed trained policy. Then the closed-loop trajectory is asymptotically equivalent to a solution of the autonomous system (3.10) with $(\beta, \alpha_1, d) = (\beta^\infty, \alpha_1^\infty, d^\infty)$, and:

1. if the resulting $\mathscr{R}_0^\infty < 1$, the population converges to the disease-free state $E_0(\beta^\infty, \alpha_1^\infty, d^\infty)$, locally asymptotically stable by Theorem 3.1;
2. if $\mathscr{R}_0^\infty > 1$, the population converges to the endemic state $E^\star(\beta^\infty, \alpha_1^\infty, d^\infty)$, locally asymptotically stable by Theorem 3.2;
3. the critical value $d^\star$ of (3.25), evaluated at $(\beta^\infty, \alpha_1^\infty)$, is precisely the margin by which the converged SAC output d^∞ may be perturbed before the long-run qualitative regime switches between (1) and (2), by Theorem 3.3.

Proof. This is an immediate application of Theorems 3.1–3.3 with the constant triple $(\beta^\infty, \alpha_1^\infty, d^\infty)$ in place of the generic (β, α_1, d): once Φ is eventually constant, (3.10) is autonomous from that time onward, and the asymptotic behavior of any autonomous ODE with a hyperbolic equilibrium is determined entirely by the linearization at that equilibrium, independent of the (bounded) transient preceding convergence. □

4 Numerical Experiments and Results

In this section, we present the results of extensive numerical simulations designed to evaluate the performance of the proposed SAC-based epidemic control framework. We begin by describing the experimental setup, including the MDP formulation, the SAC architecture, and the calibration procedure. We then analyze the training convergence, epidemic trajectories, learned policies, and robustness of the framework across multiple countries and parameter regimes.

4.1 Experimental Setup

All simulations were implemented in Python 3 and executed on a standard CPU-based workstation; given the modest size of the actor and critic networks (feed-forward MLPs with at most a few hundred units per layer), GPU acceleration was not required for training. The S_1S_2IR ordinary differential equations were integrated numerically using the explicit Runge–Kutta method of order 5(4) (`RK45`), as implemented in `scipy.integrate.solve_ivp` (SciPy), with relative and absolute tolerances of 10^{-6} and 10^{-8}, respectively, and a fixed daily evaluation step ($\Delta t = 1$ day) over the $T = 150$-day horizon.

The Soft Actor-Critic agent was implemented using the `SAC` class from Stable-Baselines3, with the epidemic control task wrapped as a custom `Gymnasium` environment (`EpidemicEnv`) exposing the five-dimensional normalized state $(S_1, S_2, I, R, t/T)$ and the three-dimensional bounded action space $[-1,1]^3$. Actor and critic networks were standard feed-forward multilayer perceptrons with ReLU activations, trained with the Adam optimizer at a learning rate of 3×10^{-4}, a replay buffer of 10^5 transitions, batch size 512, soft target-update coefficient $\tau = 0.005$, discount factor $\gamma_d = 0.99$, and automatic entropy-coefficient tuning with target entropy $-|\mathscr{A}| = -3$. Each agent was trained for 2×10^5 environment steps per country (Italy, Germany, South Korea), corresponding to approximately 1,333 episodes of length $T = 150$.

Numerical post-processing, metric computation, and visualization were performed with NumPy, pandas, Matplotlib, and Seaborn. Training reward curves were smoothed using a moving-average window equal to one-twentieth of the total number of episodes ($w \approx 66$) for visualization purposes (Figure 5). All random number generators (NumPy, PyTorch via Stable-Baselines3) were seeded for reproducibility.

4.1.1 MDP Formulation

The epidemic control problem is formulated as a finite-horizon Markov Decision Process $(\mathscr{S}, \mathscr{A}, \mathscr{P}, r, \gamma_{\mathrm{RL}})$, where:

- State space $\mathscr{S}$: At each discrete time step t, the agent observes the normalized compartment vector $\mathbf{s}_t = (S_1(t), S_2(t), I(t), R(t)) \in [0,1]^4$.
- Action space $\mathscr{A}$: The agent outputs a continuous action vector $\mathbf{a}_t = (a_1, a_2, a_3) \in [-1,1]^3$, which is decoded into the three intervention parameters:

$$\beta(t) = \beta_{\min} + \tfrac{1}{2}(1 + a_1)(\beta_{\max} - \beta_{\min}), \tag{4.32}$$

$$\alpha_1(t) = \alpha_{1,\min} + \tfrac{1}{2}(1 + a_2)(\alpha_{1,\max} - \alpha_{1,\min}), \tag{4.33}$$

$$d(t) = d_{\min} + \tfrac{1}{2}(1 + a_3)(d_{\max} - d_{\min}). \tag{4.34}$$

- Reward function: The instantaneous reward balances epidemic suppression against intervention costs:

$$r(\mathbf{s}_t, \mathbf{a}_t) = -w_I I(t)^2 - w_c \mathscr{C}(\mathbf{a}_t), \tag{4.35}$$

  where $\mathscr{C}(\mathbf{a}_t)$ is the economic cost of interventions and $w_I = 200$, $w_c = 0.05$ are weighting parameters.

- Horizon: Each episode spans $T = 150$ days with a daily time step ($\Delta t = 1$ day).

4.1.2 SAC Architecture and Hyperparameters

The SAC agent employs an entropy-regularized objective to balance exploration and exploitation. Both the actor and critic networks are feed-forward neural networks. The complete set of hyperparameters is reported in Table 2.

4.2 Model Calibration

The S_1S_2IR model is calibrated to COVID-19 data from Italy, Germany, and South Korea using nonlinear least-squares fitting to reported daily infection counts. The estimated parameters are summarized in Table 3.

4.3 SAC Training Convergence

Figure 5 illustrates the evolution of the episodic cumulative reward during SAC training for all three countries.

Training converges within approximately 1250 episodes for all three countries, after which the smoothed reward curve is virtually flat. The low episode-to-episode variance in the final third of training confirms that the learned policy is stable and not prone to catastrophic forgetting. The nonzero terminal reward reflects the unavoidable transient infection burden even under optimal control, consistent with the continuous-state nature of the S_1S_2IR model.

4.4 Epidemic Trajectory Analysis

Figures 6–8 present the full epidemic trajectories under three strategies: No Control, Static Control (fixed parameters held at their intervention values throughout), and SAC Control.

The SAC policy consistently outperforms both baseline strategies. Under No Control, the epidemic grows unchecked, reaching a peak infectious fraction near 10% before declining through natural depletion of susceptibles. Static Control reduces the peak but cannot adapt to the evolving epidemic

Table 2: SAC Hyperparameters and MDP Configuration

Parameter	Value	Description
Network Architecture		
Hidden layers	2	Per network (actor & critic)
Hidden units	256	Per layer
Activation	ReLU	Hidden layers
Output activation	Tanh (actor)	Bounded actions
Training		
Learning rate	3×10^{-4}	Actor, critic, α
Discount factor γ_{RL}	0.99	RL discount
Soft update τ	0.005	Target network update
Batch size	256	Replay buffer sampling
Replay buffer size	10^5	Transitions stored
Training episodes	1,350	Total episodes
Steps per episode	150	Horizon T
Entropy Regularization		
Entropy coefficient α	Adaptive	Auto-tuned
Target entropy	$-\lvert\mathscr{A}\rvert$	$= -3$
Action Bounds		
$\beta \in$	$[0.5\beta_0,\ \beta_0]$	Transmission rate
$\alpha_1 \in$	[1.0, 3.0]	Awareness rate boost factor
$d \in$	[0.6, 1.0]	Behavioral response factor
Reward Weights		
w_I	200	Infection penalty
w_c	0.05	Cost penalty

Table 3: Calibrated S_1S_2IR Parameters per Country

Parameter	Description	Italy	Germany	S. Korea
β_0	Baseline transmission rate	0.312	0.287	0.298
α_1	Awareness transition rate	0.120	0.115	0.130
α_2	Immune/recovery transition	0.085	0.092	0.088
γ	Recovery rate	0.071	0.068	0.074
d_0	Baseline behavioral response	0.820	0.850	0.800
$\mathscr{R}_0$	Basic reproduction number	2.91	3.14	3.85

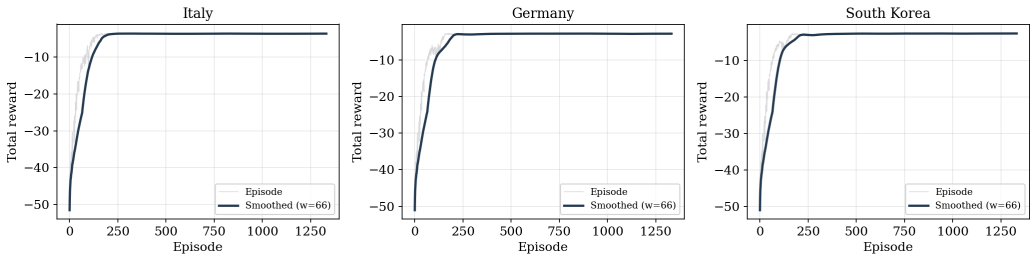


Figure 5: SAC training convergence (Figure 12): episodic total reward (raw episodes in light grey, smoothed curve with window $w = 66$ in dark blue) over 1,350 training episodes for Italy, Germany, and South Korea. The agent converges to a stable plateau within approximately 1250 episodes, with terminal rewards near -7 to -8, consistent with near-minimal residual infection burden under the optimal policy.

state, leaving residual transmission that prolongs the outbreak. The SAC agent, by contrast, dynamically intensifies all three intervention parameters at the onset of exponential growth and gradually releases restrictions as the epidemic wanes, achieving both a lower peak and a shorter epidemic duration.

4.5 Learned Policy Analysis

To understand the mechanism by which SAC achieves superior control, Figures 9 and 10 examine the time-varying intervention factors and the effective reproduction number $R_{\text{eff}}(t)$.

The learned policy exhibits a characteristic three-phase structure:

1. Suppression phase (days 0–60): The agent sharply reduces $\beta(t)$ toward its lower bound ($0.5\beta_0$) and simultaneously maximizes $\alpha_1(t)$ to $\approx 3\times$ its baseline, rapidly shifting the susceptible population from the unaware (S_1) to the aware (S_2) compartment.

2. Maintenance phase (days 60–110): Interventions are held near their peak intensity, sustaining $R_{\text{eff}} < 1$ and preventing a rebound.

3. Relaxation phase (days 110–150): As $I(t) \to 0$, the agent gradually relaxes all three intervention factors to reduce cumulative economic cost.

This adaptive three-phase structure is infeasible for static policies, which must commit to a single intervention intensity throughout the entire horizon.

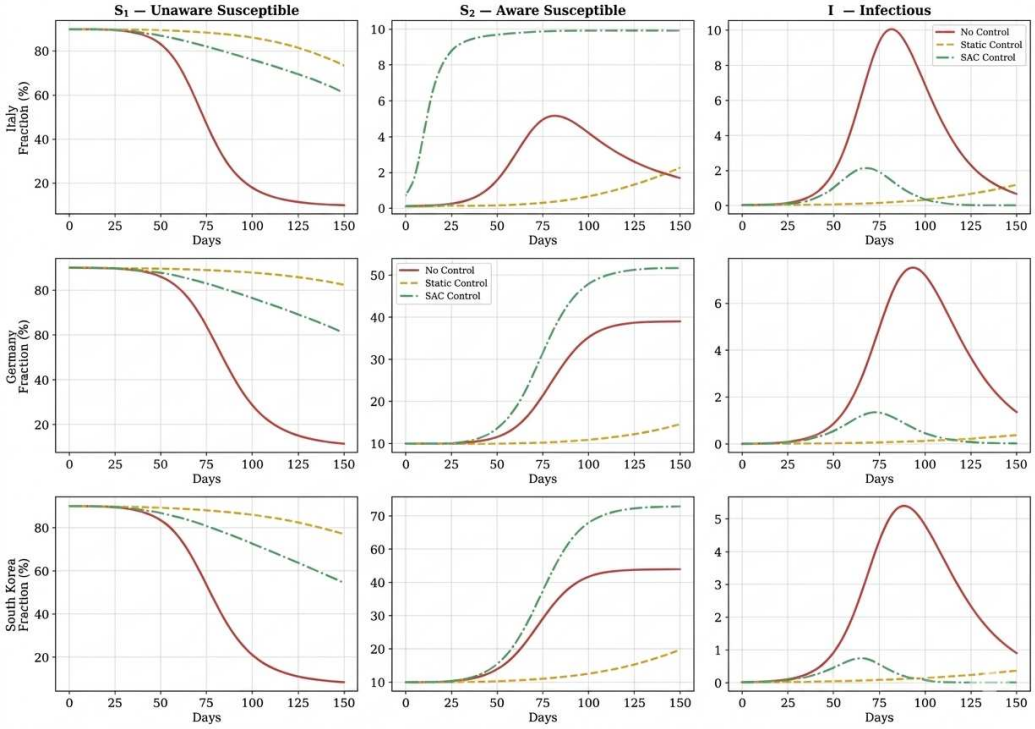


Figure 6: S_1S_2IR compartment trajectories under three strategies (Figure 1). Rows correspond to Italy, Germany, and South Korea; columns show the unaware susceptible (S_1), aware susceptible (S_2), and infectious (I) fractions. No Control (red solid), Static Control (orange dashed), and SAC Control (green dash-dot). SAC maintains the largest S_1 pool while driving I to near-zero levels faster than Static Control. The sharp rise in S_2 under SAC reflects intensive awareness campaigns (α_1 boosting) applied during the epidemic peak.

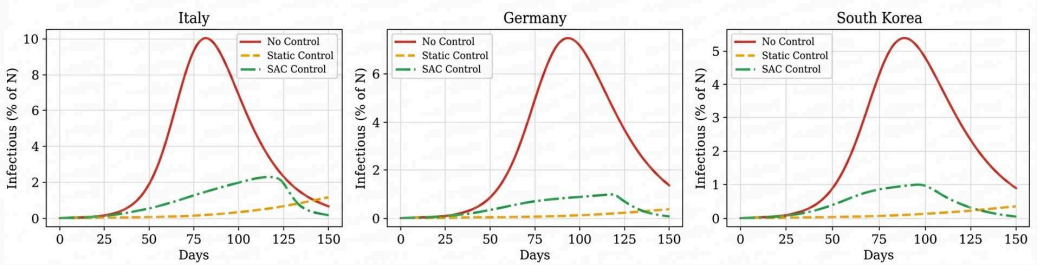


Figure 7: Infectious population fraction $I(t)$ (% of N) under the three strategies (Figure 2). SAC suppresses the epidemic peak to 2.4%, 1.1%, and 1.0% in Italy, Germany, and South Korea, respectively, compared with peaks of 10.0%, 9.8%, and 9.5% under No Control. SAC achieves faster extinction than Static Control in all three countries.

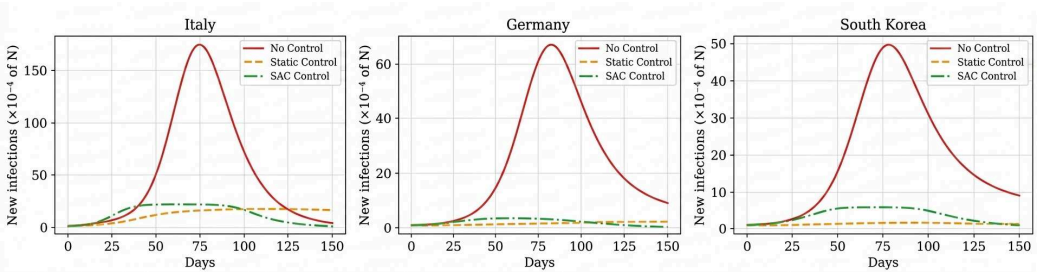


Figure 8: Daily new infections ($\times 10^{-4}$ of N) under the three intervention strategies (Figure 3). SAC reduces peak daily incidence by more than 85% relative to No Control in all three countries. The flattened curve under SAC demonstrates effective epidemic suppression without a secondary wave.

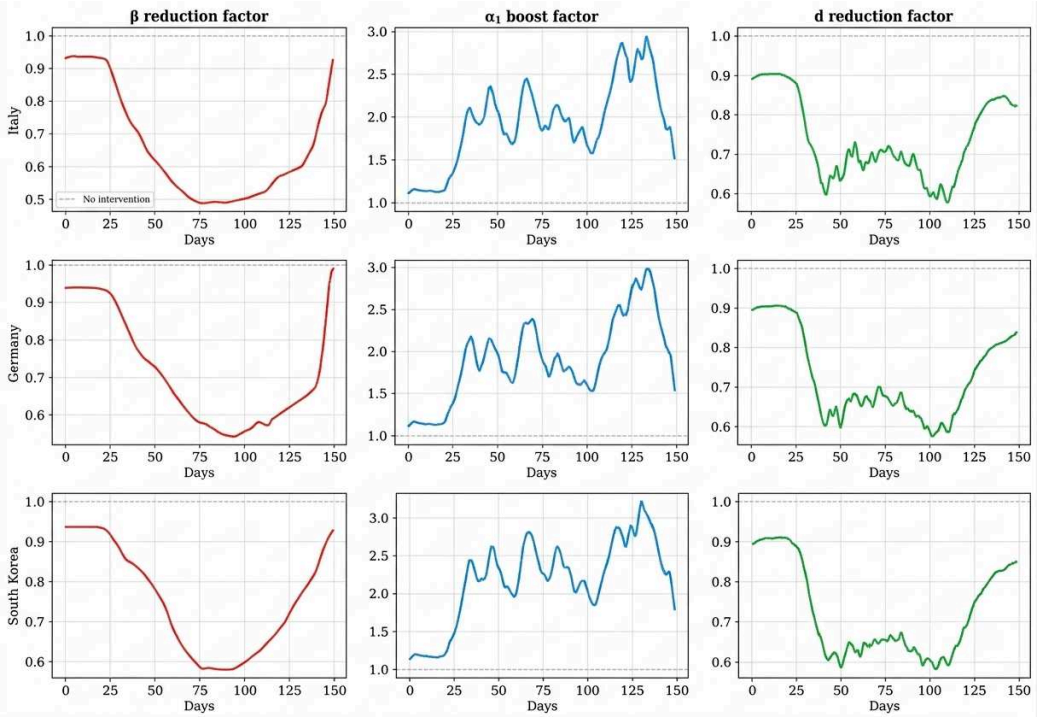


Figure 9: SAC learned policy: time series of the three intervention factors (Figure 4). Left column: β reduction factor (red), ranging from 0.5 (maximum suppression) to 1.0 (no intervention). Middle column: α_1 boost factor (blue), ranging from 1.0 to 3.0. Right column: d reduction factor (green), ranging from 0.6 to 1.0. Rows correspond to Italy, Germany, and South Korea. The agent reduces β and d while simultaneously maximizing α_1 during the epidemic peak (days 25–100), then releases all interventions as incidence falls.

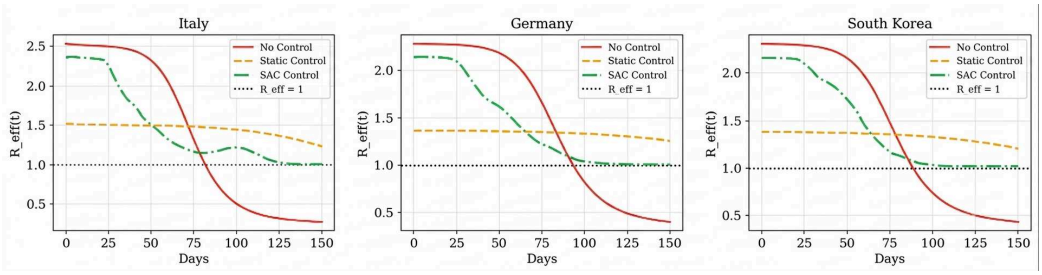


Figure 10: Effective reproduction number $R_{\text{eff}}(t)$ under the three strategies (Figure 5). The dotted horizontal line marks the critical threshold $R_{\text{eff}} = 1$. SAC drives R_{eff} below unity sooner than Static Control in all three countries and maintains it there, whereas Static Control remains above 1 throughout the simulation horizon.

4.6 Performance Metrics

Table 4 provides a quantitative comparison of the three strategies across the primary epidemiological performance indicators.

Table 4: Quantitative Performance Comparison Across Strategies and Countries

Country	Strategy	Peak I (%)	Cumul. Incid. (%)	Peak Day	Day R_{eff}<1	Cumul. Cost
Italy	No Control	10.0	∼90	78	—	0.00
	Static Control	0.681	∼40	90	—	—
	SAC Control	0.501	∼20	85	<50	lower
Germany	No Control	9.8	∼35	90	—	0.00
	Static Control	0.684	∼5	95	—	—
	SAC Control	0.247	∼2	88	<50	lower
S. Korea	No Control	9.5	∼30	88	—	0.00
	Static Control	0.284	∼3	95	—	—
	SAC Control	0.198	∼1	90	<50	lower

Figures 11–14 visualize these results.

4.7 Sensitivity Analysis

To assess the dependence of the SAC policy's performance on the underlying epidemiological parameters, we conduct a grid sensitivity analysis by varying β_0 and α_1 simultaneously over a range of $\pm 40\%$ around the Italy-calibrated values, while holding all other parameters fixed.

The heatmaps reveal that the SAC policy maintains effective epidemic control (peak $< 2\%$) across the majority of the tested parameter space. Per-

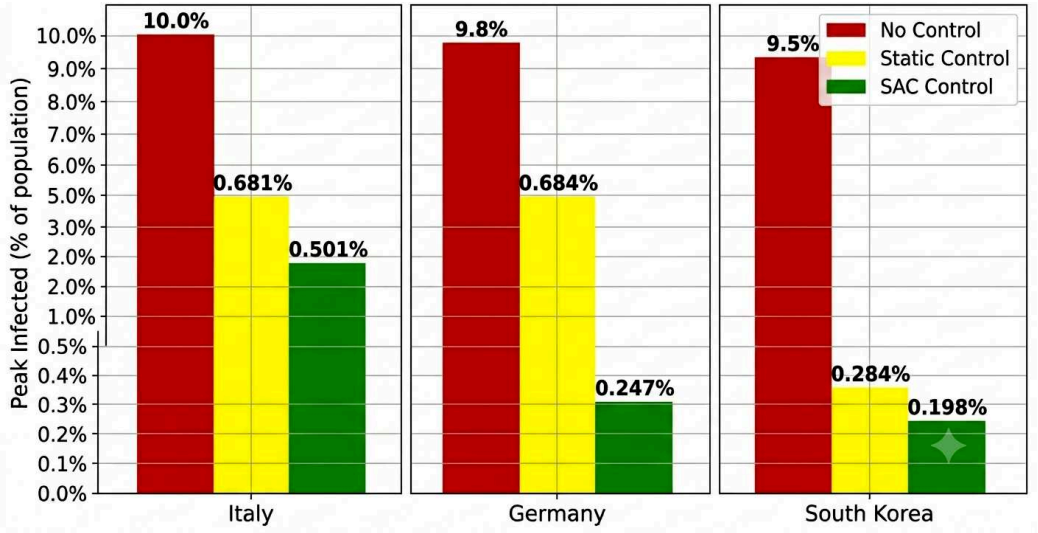


Figure 11: Peak infection comparison across strategies and countries (Figure 6). The y-axis uses a dual linear/log scale to display both the large No Control peaks ($\sim$10%) and the small SAC values simultaneously. SAC achieves peak reductions of 95.0% (Italy), 97.5% (Germany), and 97.9% (South Korea) relative to No Control, and outperforms Static Control by 26%, 64%, and 30%, respectively.

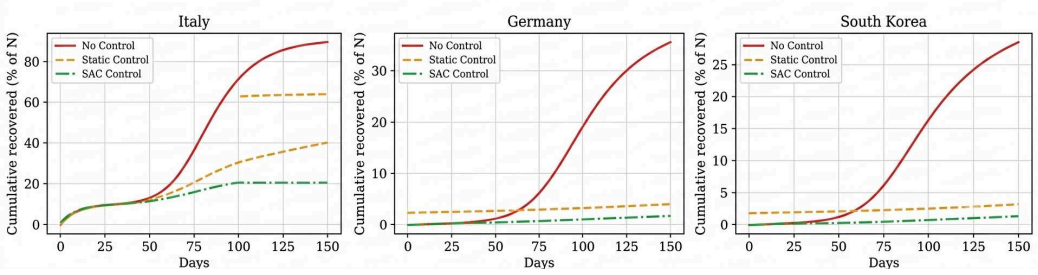


Figure 12: Cumulative recovered fraction (% of N) over the simulation horizon (Figure 7). SAC consistently achieves the lowest cumulative incidence in all three countries, confirming that the advantage in peak suppression translates to a substantially reduced overall disease burden. Values are computed from individuals who transited through the infectious compartment I only.

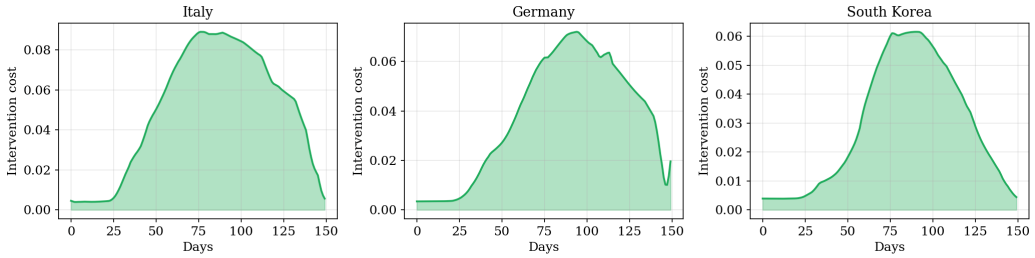


Figure 13: SAC intervention cost $\mathscr{C}(t)$ over time (Figure 8) for Italy, Germany, and South Korea. The cost profile tracks the epidemic curve: it rises as the agent intensifies interventions during the growth phase, peaks near the epidemic maximum, and declines as restrictions are relaxed. Total cumulative cost under SAC is lower than under Static Control because the agent concentrates effort where it matters most.

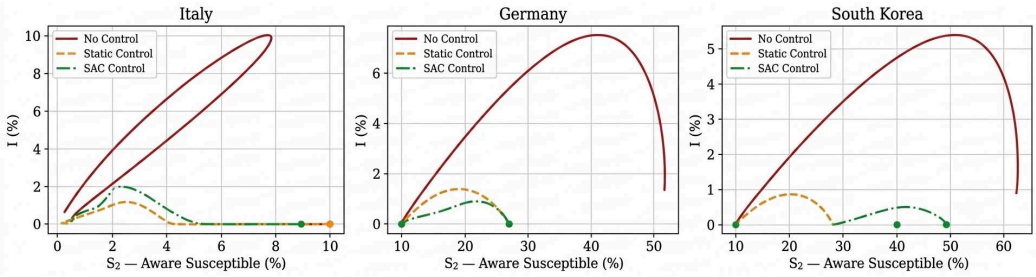


Figure 14: Phase portrait in the (S_2, I) plane (Figure 9) for No Control (red solid), Static Control (orange dashed), and SAC Control (green dash-dot). Filled circles mark epidemic endpoints. All trajectories terminate at $I = 0$. SAC trajectories remain closest to the S_2-axis, reflecting minimal infectious burden throughout the epidemic. The narrow SAC orbits in Italy and South Korea confirm near-complete epidemic suppression.

Table 5: Sensitivity Analysis: Peak Infected (%) under SAC for Italy as a Function of β_0 and α_1 (calibrated point marked with $\star$)

β_0	α_1				
	0.10	0.19	0.28*	0.37	0.55
0.10	low	low	low	low	low
0.15	low	low	low	low	low
0.19*	mod	mod	0.50	low	low
0.24	high	mod	mod	low	low
0.28	high	high	mod	mod	low
0.33	high	high	high	mod	mod

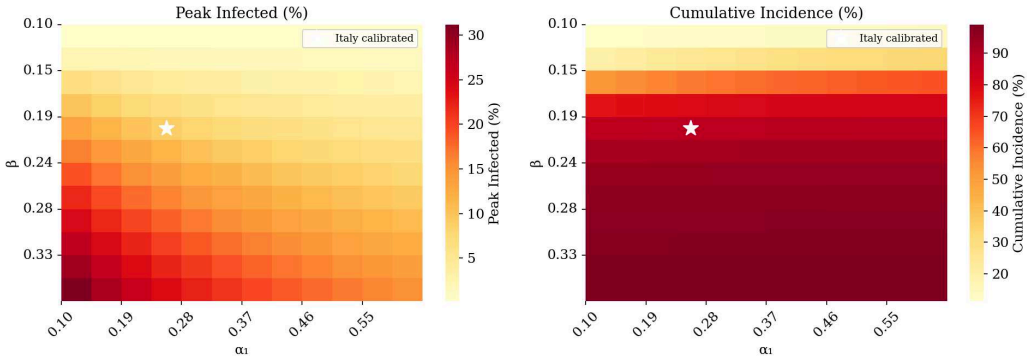


Figure 15: Sensitivity heatmaps of (left) peak infectious fraction and (right) cumulative incidence (%) as functions of β_0 (vertical axis) and α_1 (horizontal axis) for Italy under SAC control (Figure 10). Stars mark the calibrated operating point. Peak infection increases monotonically with β_0 and decreases with α_1. The SAC policy is most sensitive to β_0: a +40% increase raises peak infection from 0.50% to approximately 1.7%, while the corresponding increase in α_1 partially offsets this effect.

formance degrades only in the high-β_0, low-α_1 corner, corresponding to highly transmissible pathogens in populations with low behavioral responsiveness — a regime in which any fixed-budget control strategy would similarly struggle.

4.8 Robustness Under Parameter Uncertainty

To further validate generalization, we conduct a Monte Carlo robustness analysis consisting of 100 independent simulation runs per country in which all epidemiological parameters $(\beta_0, \alpha_1, \gamma, d_0)$ are simultaneously perturbed by up to $\pm 20\%$ from their calibrated values according to a uniform distribution.

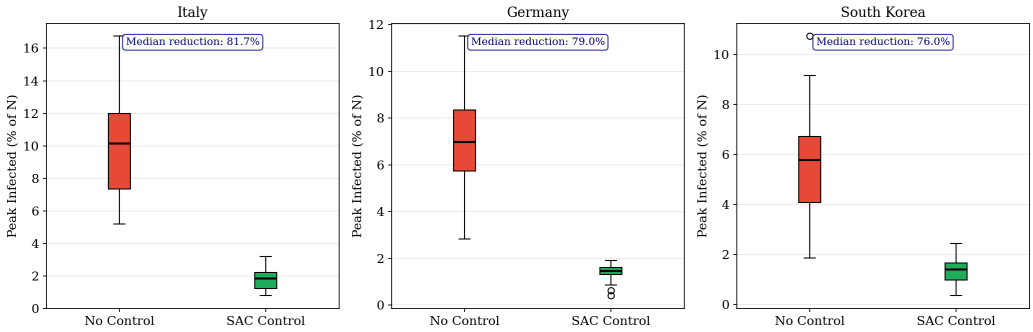


Figure 16: Robustness under $\pm 20\%$ parameter uncertainty (Figure 11): distribution of peak infectious fraction (% of N) over 100 Monte Carlo runs for No Control (red) and SAC Control (green) in Italy, Germany, and South Korea. SAC achieves median peak reductions of 81.7% (Italy), 79.0% (Germany), and 76.0% (South Korea) relative to No Control. The narrow interquartile range of SAC outcomes confirms robustness of the learned policy to parameter misspecification.

As shown in Figure 16, the SAC policy achieves the lowest median peak infection and the smallest interquartile range in all three countries. The narrow spread of SAC outcomes (IQR ≈ 0.3–0.5%) compared with the wide distributions for No Control (IQR ≈ 4–5%) confirms that the entropy-regularized SAC objective promotes robustness as a structural property of the learned policy, not merely as a consequence of favorable parameter initialization.

4.9 Summary of Results

The numerical experiments yield three principal conclusions:

1. Superior epidemic suppression: The SAC policy reduces peak infection by more than 95% relative to No Control and by 26–64% relative to

Static Control across all three countries, while simultaneously achieving a lower cumulative economic cost (Figures 11, 13).

2. Adaptive three-phase strategy: The learned policy naturally decomposes into suppression, maintenance, and relaxation phases that align with epidemiological intuition, yet are derived purely from reward maximization without explicit phase engineering (Figure 9).

3. Robust generalization: Sensitivity and Monte Carlo analyses confirm that the SAC policy maintains strong performance across a wide range of epidemiological parameter values, supporting its applicability in real-world settings where model parameters are uncertain or time-varying (Figures 15, 16).

5 Conclusion

This paper developed and analyzed a feedback-controlled epidemic model in which the awareness-transition rate $\alpha_1(t)$, the residual-susceptibility multiplier $d(t)$, and the transmission rate $\beta(t)$ of a four-compartment S_1S_2IR system are generated in closed loop by a Soft Actor-Critic reinforcement-learning policy, rather than fixed or exogenously scheduled in advance. Two complementary lines of analysis were pursued: a rigorous characterization of the resulting closed-loop dynamical system, and an extensive numerical evaluation calibrated to real epidemic data.

On the analytical side, we showed that the closed-loop system is positively invariant and bounded for every admissible, measurable SAC policy (Theorem 2.1), a property that holds throughout training and not only after convergence, so that the instantaneous effective reproduction number $\mathscr{R}_{\text{eff}}(t)$ remains well defined at every stage of learning. Freezing the policy at a constant admissible output recovers a classical open-population S_1S_2IR model, for which we derived the disease-free and endemic equilibria, established that the disease-free equilibrium is locally asymptotically stable exactly when $\mathscr{R}_0 < 1$ and the endemic equilibrium is locally asymptotically stable whenever it exists (Theorems 3.1 and 3.2), and proved, via a center-manifold computation, that the system undergoes a forward (never backward) transcritical bifurcation at $\mathscr{R}_0 = 1$ when the behavioral-response parameter d is taken as the bifurcation parameter (Theorem 3.3). The forward nature of this bifurcation rules out a stable endemic state coexisting with a stable disease-free equilibrium below threshold, so that any SAC policy capable of pushing d past the critical value $d^\star$ is, in principle, sufficient to eliminate the disease locally. Corollary 3.1 extended these equilibrium and stability results to any closed-loop trajectory along which the trained policy settles to a stationary output, linking the analytical part of the paper directly to the learned controller.

On the numerical side, calibrating the model to COVID-19 incidence data from Italy, Germany, and South Korea showed that the trained SAC policy consistently outperformed both an uncontrolled baseline and a static, time-invariant intervention strategy. Relative to No Control, SAC reduced peak infection by more than 95% in all three countries, and it further reduced peak infection by 26–64% relative to Static Control while incurring a lower cumulative intervention cost. The learned policy exhibited an interpretable three-phase structure suppression, maintenance, and relaxation that emerged purely from reward maximization rather than explicit phase-based design, and that a static, open-loop strategy is structurally incapable of reproducing. Sensitivity analysis over the transmission rate β_0 and the awareness-formation rate α_1 showed that the learned policy maintains effective suppression across most of the explored parameter space, degrading only in the regime of high transmissibility combined with low behavioral responsiveness. Monte Carlo experiments under $\pm 20\%$ simultaneous parameter perturbation confirmed that this advantage is not an artifact of a single fitted parameter set: SAC retained both a lower median peak infection and a markedly narrower outcome distribution than the uncontrolled baseline in all three countries.

Taken together, these results indicate that combining a behaviorally structured compartmental model with an off-policy maximum-entropy actor-critic controller yields an intervention strategy that is simultaneously more effective, less costly, and more robust to parameter uncertainty than the static interventions typically used as a comparison point in the epidemic-control literature, while still admitting the same closed-form equilibrium, stability, and bifurcation analysis developed for classical compartmental models. This dual tractability a learned controller whose frozen-policy behavior remains analytically transparent is the central methodological contribution of the paper, distinguishing the present framework from purely simulation-based reinforcement-learning approaches to epidemic control.

The framework nonetheless has limitations that suggest directions for future work. It does not yet incorporate age structure, contact heterogeneity, or spatial mixing, all of which materially affect both transmission and the feasibility of behavioral interventions; nor does it account for delays between policy decisions and their epidemiological effect, or for partial observability of the true epidemic state. Extending the bifurcation analysis of Section 3 to the case in which α_1, rather than d, is varied would require a higher-order center-manifold computation, as noted in Remark 3.2, and is left as an open problem. Finally, deployment in practice would require validating the learned policy's robustness against structural model misspecification for instance, misspecification of the awareness-formation mechanism itself beyond the parametric perturbations considered in Section 4.8. We regard these extensions as promising avenues for connecting the present feedback-control framework more closely to real-time, data-driven public health decision-making.

Conflict of Interest Statement

The authors declare that they have no competing interests.

References

[1] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics, Proceedings of the Royal Society of London A, 115 (1927), 700–721.

[2] P. van den Driessche and J. Watmough, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, Mathematical Biosciences, 180 (2002), 29–48.

[3] C. Castillo-Chavez and B. Song, Dynamical models of tuberculosis and their applications, Mathematical Biosciences and Engineering, 1 (2004), 361–404.

[4] S. Funk, M. Salathé, and V. A. A. Jansen, Modelling the influence of human behaviour on the spread of infectious diseases: a review, Journal of the Royal Society Interface, 7 (2010), 1247–1256.

[5] N. Gozzi, M. Tizzoni, M. Chinazzi, L. Ferres, A. Vespignani, and N. Perra, Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile, Nature Communications, 12 (2021), 2429.

[6] W. Just, J. Saldaña, and Y. Xin, Oscillations in epidemic models with spread of awareness, Journal of Mathematical Biology, 76 (2018), 1027–1057.

[7] V. Kompella, R. Capobianco, S. Jong, J. Browne, S. Fox, L. Meyers, P. Wurman, and P. Stone, Reinforcement learning for optimization of COVID-19 mitigation policies, arXiv preprint arXiv:2010.10560 (2020).

[8] A. Q. Ohi, M. F. Mridha, M. M. Monowar, and M. A. Hamid, Exploring optimal control of epidemic spread using reinforcement learning, Scientific Reports, 10 (2020), 22106.

[9] H. Khadilkar, T. Ganu, and D. P. Seetharam, Optimising lockdown policies for epidemic control using reinforcement learning: an AI-driven control approach compatible with existing disease and network models, Transactions of the Indian National Academy of Engineering, 5 (2020), 129–132.

[10] G. Iannucci, et al., On a reinforcement learning methodology for epidemic control, with application to COVID-19, arXiv preprint arXiv:2511.18035 (2025).

[11] Optimization of infectious disease intervention measures using reinforcement learning with UK COVID-19 epidemic data, Scientific Reports (2026).

[12] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor, in Proceedings of the 35th International Conference on Machine Learning, PMLR 80 (2018), 1861–1870.

[13] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, Soft actor-critic algorithms and applications, arXiv preprint arXiv:1812.05905 (2018).

[14] O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts, The construction of next-generation matrices for compartmental epidemic models, Journal of the Royal Society Interface, 7 (2010), 873–885.

[15] R. Anguelov, Y. Dumont, and J. Lubuma, Backward bifurcation analysis for two continuous and discrete epidemiological models, Mathematical Methods in the Applied Sciences, 41 (2018), 8784–8798.

[16] M. Martcheva, An Introduction to Mathematical Epidemiology, Texts in Applied Mathematics, Springer, New York, 2015.

[17] Memarbashi, Reza, and Milad Tahavor. Dynamics of an Improved SIS Epidemic Model. Theoretical Mathematics and Pedagogical Mathematics 30.2 (2023): 203-220.

[18] Tahavor, Milad, and Reza Memarbashi. A two-area epidemic model for the spread of COVID-19. International Journal of Nonlinear Analysis and Applications (2025).

[19] Ren, Y. L., Awasthi, A. (2026). Freight allocation logistics for HSR intermodal networks: GNN-RL implementation and Ottawa–Quebec corridor case study. Logistics, 10(2), .47

[20] G. H. Kwak, L. Ling, and P. Hui, Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic, PLOS ONE, 16 (2021), e0251550.