AI Safety x Physics Grand Challenge Submission

Project track

Valentina Schastlivaia[*]
Molecular Bionics Labs
Institute for Bioengineering of Catalonia
Barcelona, Spain
vschastlivaia@ibecbarcelona.eu

Aray Karjauv[†]
XAI
Technical University of Berlin
Berlin, Germany
aray.karjauv@tu-berlin.de

With
PIBBSS, Timaeus, & Apart Research

July 28, 2025

Abstract

As AI systems scale into decentralized, multi-agent deployments, emergent vulnerabilities challenge our ability to evaluate and manage systemic risks. In this work, we adapt classical epidemiological modeling (specifically SEIR compartment models) to model adversarial behavior propagation in AI agents. By solving systems of ODEs describing the systems with physics-informed neural networks (PINNs), we analyze stable and unstable equilibria, bifurcation points, and the effectiveness of interventions. We estimate parameters from real-world data (e.g., adversarial success rates, detection latency, patching delays) and simulate attack propagation scenarios across 8 sectors (enterprise, retail, trading, development, customer service, academia, medical, and critical infrastructure AI tools). Our results demonstrate how agent population dynamics interact with architectural and policy design interventions to stabilize the system. This framework bridges concepts from dynamical systems and cybersecurity to offer a proactive, quantitative toolbox on AI safety. We argue that epidemic-style monitoring and tools grounded in interpretable, physics-aligned dynamics can serve as early warning systems for cascading AI agentic failures.

Keywords *Physics-informed AI safety · your specific physics approach · AI safety problem area · your methodology*

1 Introduction

Language-based AI agents are increasingly deployed across domains, from customer support to autonomous trading agents.

According to KMPG's survey 5,161 businesses with $1 billion or more in revenue reviles that 12% of respondents have deployed AI agents for use across their organizations, another 37% are piloting AI agents, and roughly half (51%) of organizations are exploring the use of AI agents. More than 80% of respondents identified risk management as a significant concern in their generative AI strategies.

However, the widespread deployment of these AI agents has also exposed fundamental vulnerabilities in their reliability [Boisvert et al., 2025]. They can be compromised by adversarial inputs (e.g. prompt injections), propagate misinformation learned from uncurated data, or miscommunicate in multi-agent settings [Nie et al., 2024, Lin et al., 2021, Sun et al., 2022]. While traditional robustness evaluations focus on static benchmarks or single-turn prompt testing, real-world deployments demand systematic, multi-faceted assessment under both intentional attacks and emergent errors [Boisvert et al., 2025]. Moreover, existing evaluation frameworks seldom address interactions among agents or leverage domain-driven priors to improve resilience [Sun et al., 2022].

[*] Research conducted at the AI Safety x Physics Grand Challenge, 2025

[†]

2 Methods

To model the propagation of adversarial behavior in large AI agent populations, we adapt a well-known technique from epidemiology: the SEIR compartmental model. We treat agents as elements of a dynamic system whose states change according to interactions with other agents, external adversaries, and intervention policies. See 2.1

We use Physics-Informed Neural Networks (PINNs) to learn the solution trajectories of the governing differential equations. PINNs are well-suited for this task because they allow us to encode known physical structure. See 2.2

Beyond simply tracking number of malignant agents, we investigate the systems' phase spaces and stability properties.

These help us answer key questions such as: What are current parameters of the system? At what parameter values does the system transition from a unsafe to safe regime? What are the long-term equilibrium states? How sensitive are these outcomes to intervention timing and scale? See **??**

2.1 Theoretical Framework: Epidemiological Model Adaptation

We reinterpret the SEIR model (originally developed for biological epidemics) as a way to track how adversarial behavior spreads among AI agents.

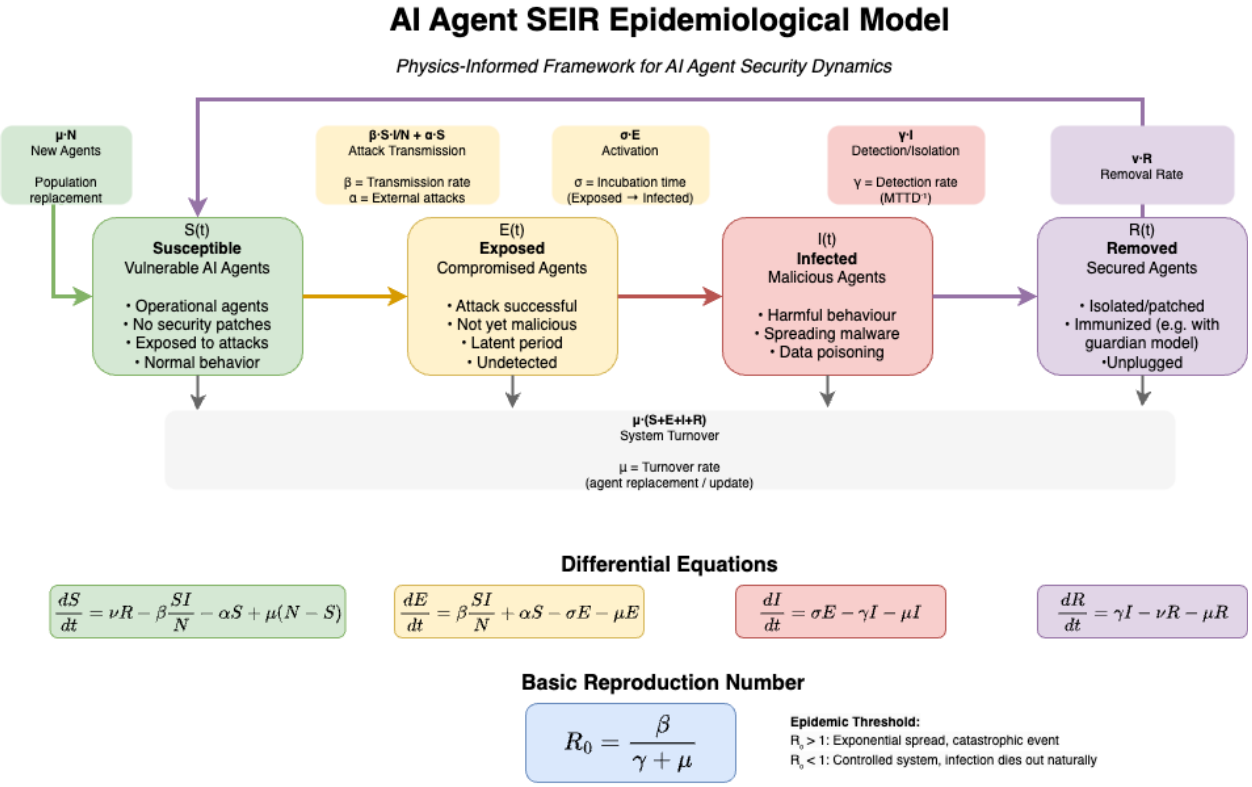


Figure 1: AI Agent SEIR Epidemiological Model

$$\frac{dS}{dt} = \nu R - \beta SI/N - \alpha S + \mu(N - S) \tag{1}$$

$$\frac{dE}{dt} = \beta SI/N + \alpha S - \sigma E - \mu E \tag{2}$$

$$\frac{dI}{dt} = \sigma E - \gamma I - \mu I \tag{3}$$

$$\frac{dR}{dt} = \gamma I - \nu R - \mu R \tag{4}$$

Where:

- $S(t)$: Susceptible agents (vulnerable to attacks)
- $E(t)$: Exposed agents (compromised but not actively malicious)
- $I(t)$: Infected agents (exhibiting malignant behavior)
- $R(t)$: Removed agents (isolated, patched, or immunized)

The key parameters for AI-specific dynamics:

- β: Attack transmission rate (depends on ASR and connectivity)
- σ: Incubation rate (exposed $\rightarrow$ infected transition)
- γ: Detection/isolation rate (mean time to detection)
- ν: Immunization/patching rate
- α: External attack pressure
- μ: Agent turnover rate (system refresh/replacement)

It's sometimes convenient to think of rates as probabilities of an agent transition from one state to another during time dt.

2.2 Technical Implementation: Physics-Informed Neural Network Implementation

We implement a SEIR-PINN solver using the PINNsFormer architecture [Zhao et al., 2023] to capture complex nonlinear dynamics while enforcing physical constraints. The loss function combines data fitting (when available) with physics constraints:

$$\mathcal{L} = \mathcal{L}_{data} + \lambda_{physics}\mathcal{L}_{physics} + \lambda_{boundary}\mathcal{L}_{boundary} \tag{5}$$

The physics loss enforces the SEIR differential equations:

$$\mathcal{L}_{physics} = \sum_{i=1}^{N_{physics}} \left[\left(\frac{\partial \hat{S}}{\partial t} - f_S\right)^2 + \left(\frac{\partial \hat{E}}{\partial t} - f_E\right)^2 + \left(\frac{\partial \hat{I}}{\partial t} - f_I\right)^2 + \left(\frac{\partial \hat{R}}{\partial t} - f_R\right)^2 \right] \tag{6}$$

Our numerical solver lets us explore the system's phase space, spot bifurcation points, and predict when things might tip into failure cascades.

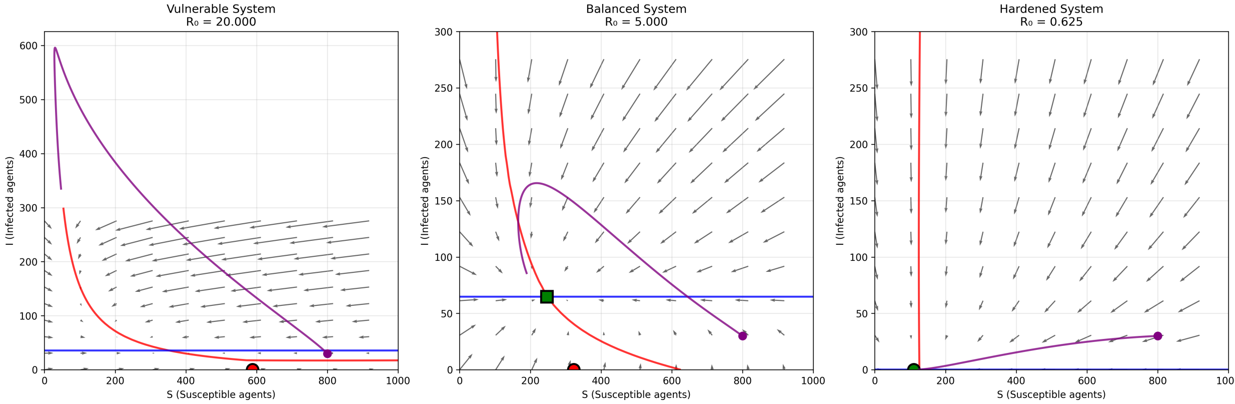


Figure 2: Systems' dynamics analysis of their phase spaces

Most importantly, it can allow to monitor and implement effective interventions when basic reproduction rate R_0 exceeds the critical threshold.

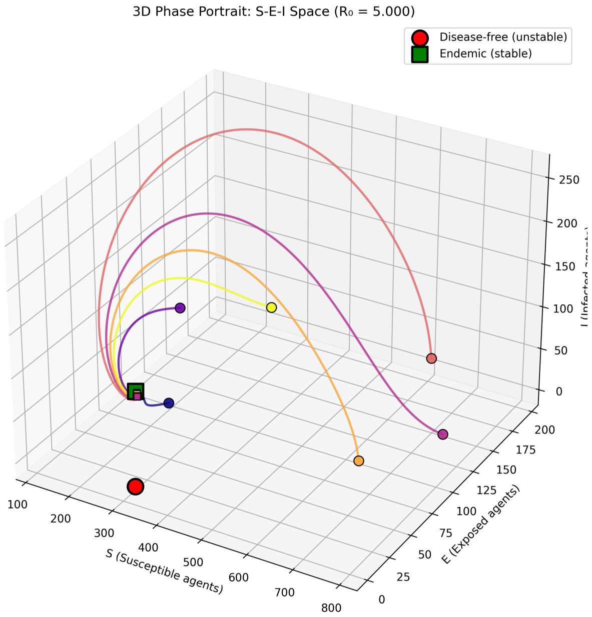


Figure 3: System's phase space S-E-I prjection ($R_0 = 5.00$) Disease-free (unstable) Endemic (stable) states

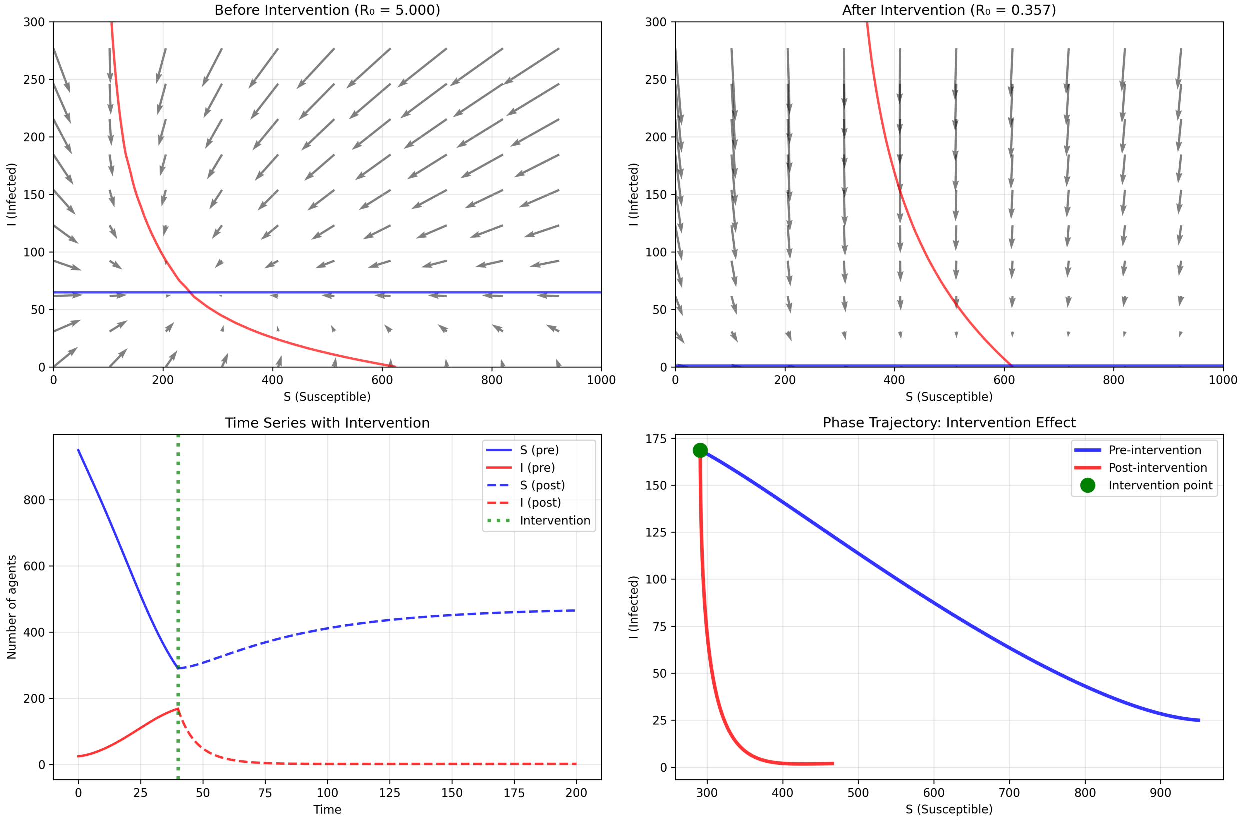


Figure 4: System's reaction on intervention. Before Intervention $R_0 = 5.00$, after intervention $R_0 = 0.357$

2.3 Experimental Design: Emperical Parameter Estimation

We combined vulnerability data from multiple sources to adjust our model to real-world observations.

- **DoomArena** Boisvert et al. [2025]: for GPT-4o from 22.7% fro airlines scenario with defence to 78.6%Zhang et al. [2024] vulnerability on OSWorld (Computer-Use), and for Claude-3.5 from 0.7% to 22.9%Zhang et al. [2024] respectively
- **Web3 Context Manipulation**Patlan et al. [2025]: 65% vulnerability (ASR) across 500+ test cases
- **Medical AI Research**Qiu et al. [2025]: 55% vulnerability in healthcare agents
- **Industry Cybersecurity Reports**Edgesca [2023], Chakrabarty [2025]: Mean Time to Detect (MTTD), Mean Time to Remediate (MTTR), and breach statistics

We estimated parameters by weighting vulnerability sources by confidence and sample size, convert vulnerability rates to transmission rates using connectivity factors (guestimation here), detection rates were taken from cybersecurity reports (MTTR, MTTD), progression rates from cyber kill chain modelHoffmann [2019], and constrained parameters against epidemic thresholds and realism

$$0.1 \leq R_0 = \frac{\beta}{\gamma + \mu} \leq 10.0 \tag{7}$$

$$\tag{8}$$

Table 1: Empirical Parameter Estimates for AI Agent Epidemiology

Parameter	Range	Interpretation	Data Source
β	0.002-0.055 day^{-1}	Daily transmission probability	DoomArena, Web3 studies
γ	0.01-0.3 day^{-1}	Daily detection rate	Industry MTTD benchmarks
σ	0.02-1.0 day^{-1}	Activation rate	Cyber kill chain model
ν	0.0005-0.05 day^{-1}	Patching rate	Software lifecycle
α	0.0001-0.005 day^{-1}	External attack rate	Threat intelligence
μ	0.0001-0.01 day^{-1}	System turnover rate	Infrastructure data

Based on analysis of publicly available data from DoomArena, academic research, and industry reports, we've derived realistic epidemiological parameters for AI agent security modeling.

Table 2: Population of different AI Agent Deployment Scenarios

AI agent purpose	Population
Enterprise Assistants	4,855 +[3]
Development Tools	85[4]
Retail bots	2M[5]
Customer Service	17,333 [6]
Research/Academic	3,000[7]
Web3/Blockchain/Autonomous Trading	200K[8]
Medical AI	223[9]
Critical Infrastructure (airlines, banks, telecoms)[10]	32,000

3 Results

Using the trained PINNs, we simulated time-series curves, estimated long-term prevalence under no intervention, and tested the effectiveness of countermeasures (such as increasing γ or ν). The results helped visualize when a given system might approach criticality and how to reduce the risk.

Analysis of 8 realistic deployment AI agentic scenarios reveals significant variation in epidemic potential.

Our empirical analysis revealed that some agent deployments (especially in research and medical contexts) lie close to or above the $R_0 = 1$ threshold, meaning the need monitoring tools and risk-mitigation frameworks.

3.1 Empirical Findings: System Dynamics and Phase Portraits

Using the trained PINNs, we plotted system trajectories in the S-I and S-E-I phase space to better understand the structure of the dynamical system. We observed that:

- In low R_0 regimes ($R_0 < 1$), the system tends toward disease-free equilibria.
- In high R_0 regimes ($R_0 > 1$), adversarial behavior persists and may saturate large parts of the agent population.
- Some systems exhibit bifurcation behavior: a critical point in the parameters where stability flips.

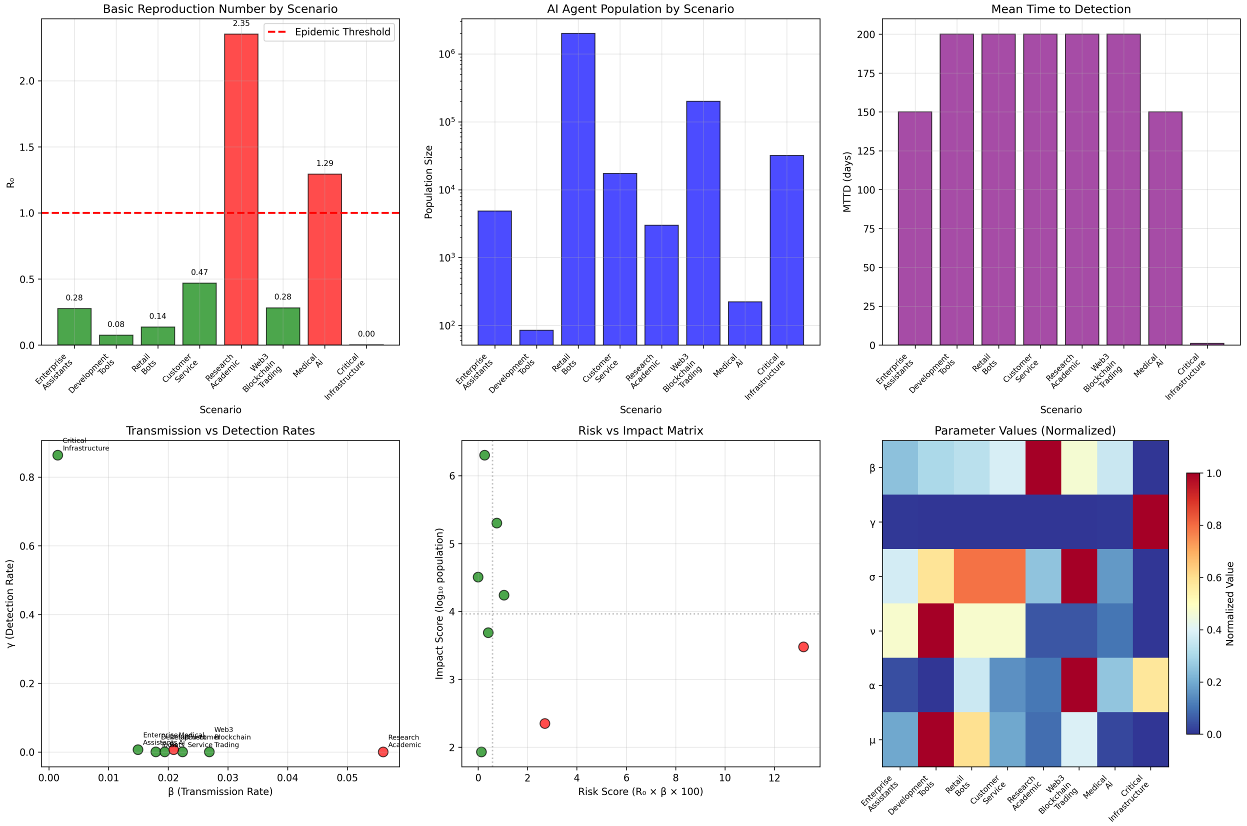


Figure 5: Estimates for current AI agent epidemiological parameters based on empirical data from cybersecurity research and industry deployments statistics

Table 3: Risk Assessment Across AI Agent Deployment Scenarios

AI Agent Purpose	Population	R_0	Risk Level	Data Source
Enterprise Assistants	4,855	0.276	LOW	DoomArena airline scenarios
Development Tools	85	0.075	LOW	DoomArena computer-use
Retail Bots	2M	0.136	LOW	DoomArena retail scenarios
Customer Service	17,333	0.469	LOW	DoomArena retail with defense
Research/Academic	3,000	2.353	LOW	DoomArena web navigation
Web3/Blockchain/Trading	200K	0.282	LOW	Web3 context manipulation
Medical AI	223	1.293	MODERATE	Medical AI vulnerability
Critical Infrastructure	32,000	0.002	LOW	NIST cybersecurity

Enhance monitoring and detection (γ) is most effective parameter for reducing R_0, network segmentation and model isolation (β) reduce attack transmission between agents, immunization by introducing guardian model ν, as well and regular updates reduce susceptible population.

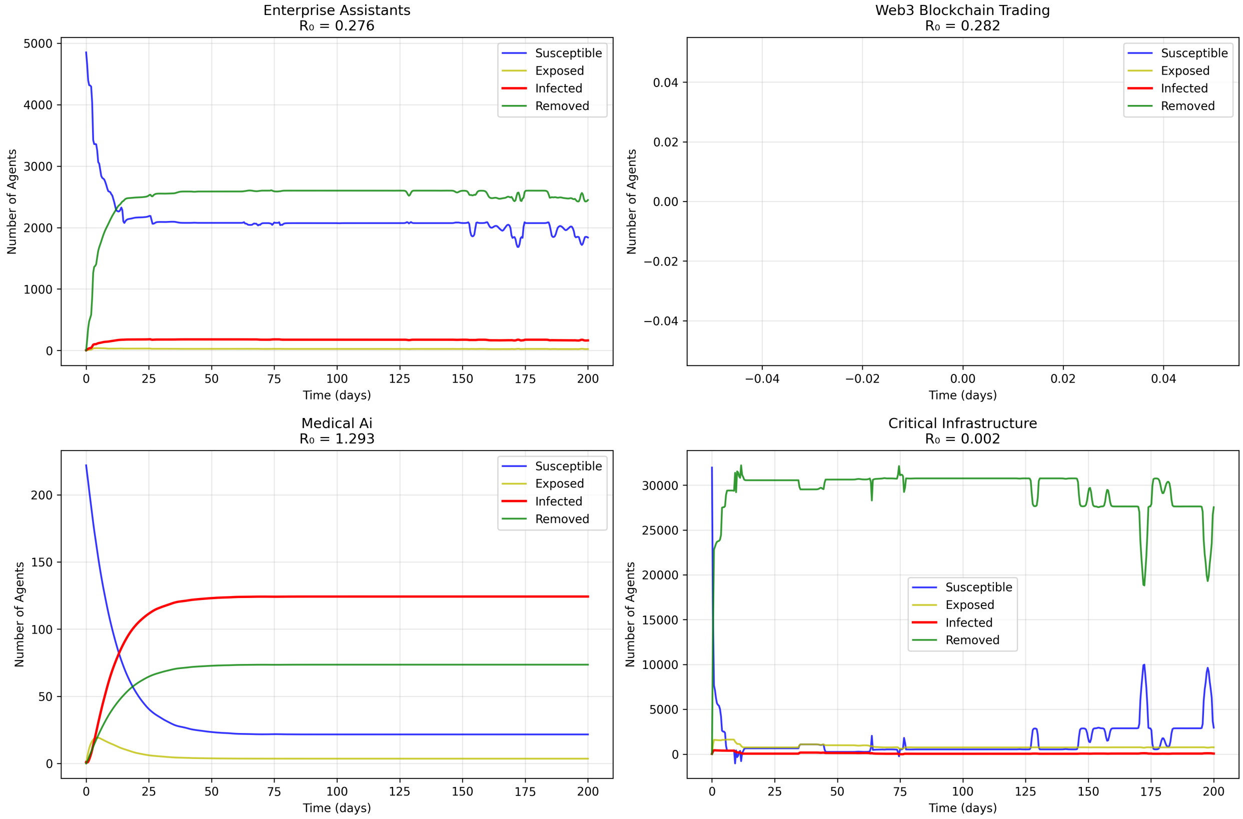
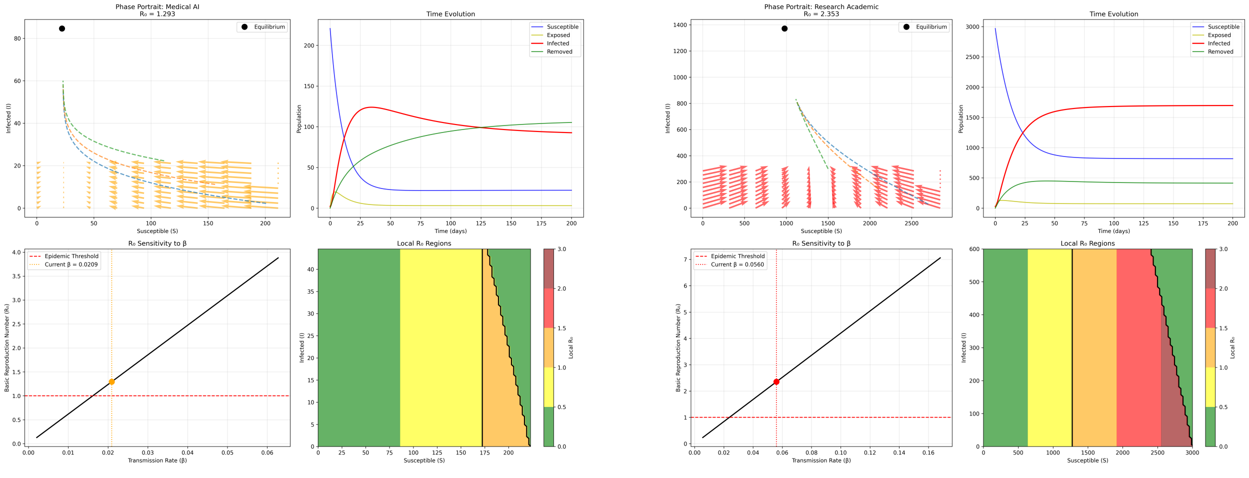


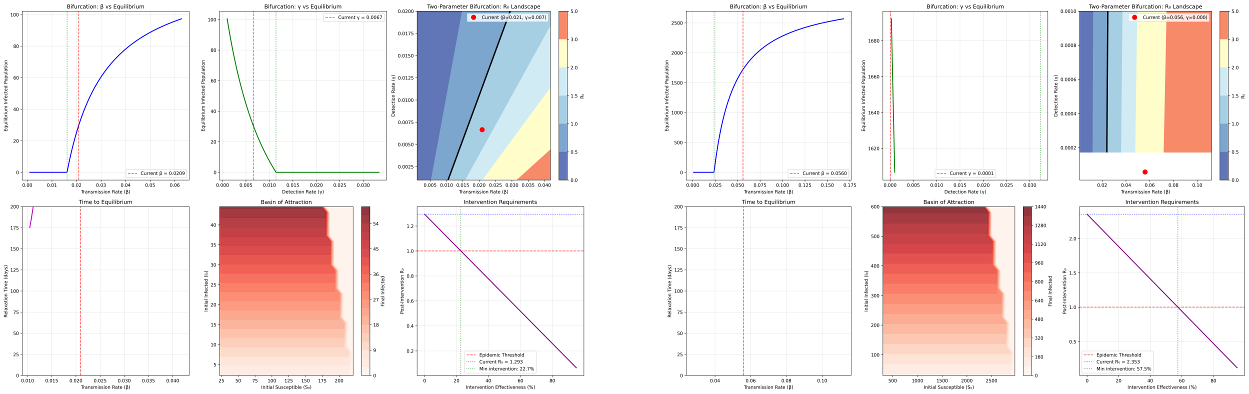
Figure 6: trained PINNs predictions of AI agent epidemiological situation evolvement



(a) Phase space for Medical AI agents

(b) Phase space for Research AI agents

Figure 7: Phase portraits of AI systems with top 2 reproduction rates



(a) Bifurcation analysis of Medical AI agents

(b) Bifurcation analysis of Research AI agents

Figure 8: Bifurcation analysis and intervention planning for AI systems with top 2 reproduction rates: Medical AI and Research AI agents

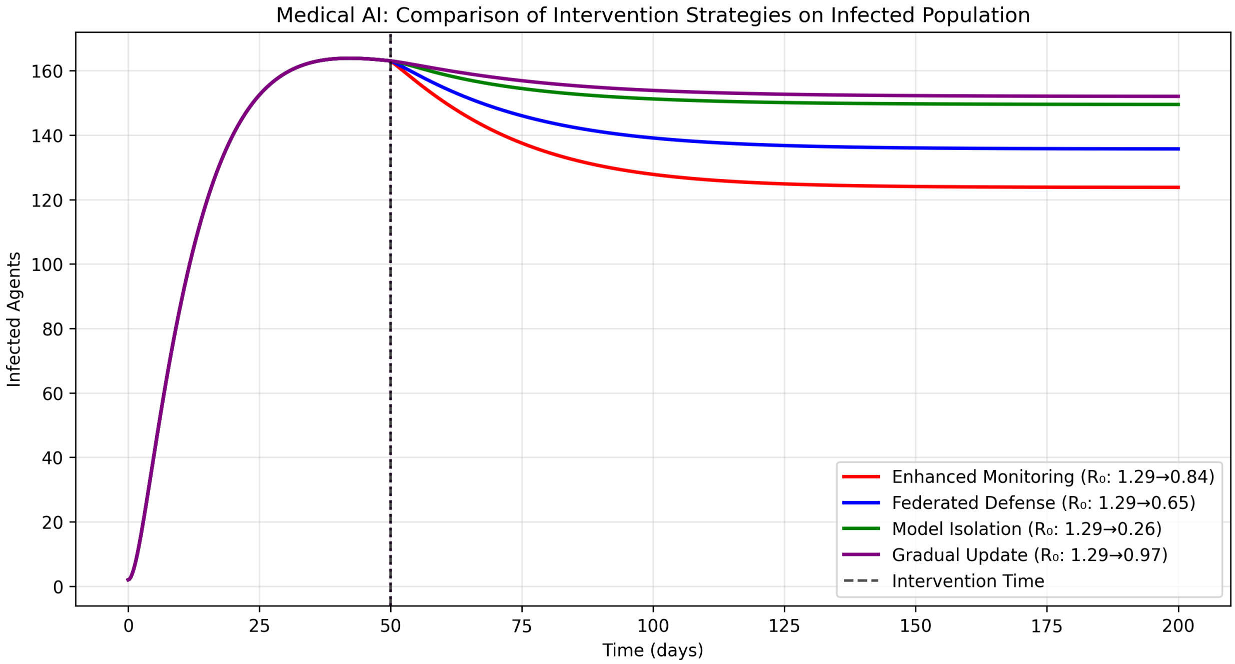


Figure 9: Comparison of intervention strategies effect on number of infected agents

4 Discussion and Conclusion

4.1 Future Directions

We have demonstrated that physics-informed epidemiological modeling provides a powerful framework for understanding and managing security risks in large-scale AI agent deployments. Our empirical analysis reveals significant variation in epidemic potential across deployment contexts, with research environments requiring immediate attention due to high R_0 values.

We would like to explore percolation theory to characterize the spread of malignant behavior. The attractiveness of percolation theory is that it exhibits power law behavior, which might be interesting to look at.

On the practical side development of a monitoring system might enable AI companies to:

1. **Continuous R_0 Calculation**: Real-time basic reproduction number monitoring
2. **Epidemic Alert System**: Automated alerts when $R_0 > 1$
3. **Time-to-Saturation Prediction**: Calculate hours until 90% infection
4. **Intervention Strategy Optimization**: Cost-benefit analysis for different responses

The real-time monitoring will enable AI companies to transition from reactive to proactive security postures, providing quantitative guidance for intervention strategies. By establishing the basic reproduction number (R_0) as a key metric for AI system health, we provide a universal language for discussing and managing AI security risks with executive management.

This work opens a new direction for physics-informed AI safety research while providing immediately actionable tools for securing the rapidly growing population of AI agents across diverse deployment contexts.

References

Leo Boisvert, Mihir Bansal, Chandra Kiran Reddy Evuru, Gabriel Huang, Abhay Puri, Avinandan Bose, Maryam Fazel, Quentin Cappart, Jason Stanley, Alexandre Lacoste, et al. Doomarena: A framework for testing ai agents against evolving security threats. *arXiv preprint arXiv:2504.14064*, 2025.

Yuzhou Nie, Zhun Wang, Ye Yu, Xian Wu, Xuandong Zhao, Wenbo Guo, and Dawn Song. Privagent: Agentic-based red-teaming for llm privacy leakage. *arXiv preprint arXiv:2412.05734*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust policy learning against adversarial communication in multi-agent systems. *arXiv preprint arXiv:2206.10158*, 2022.

Zhiyuan Zhao, Xueying Ding, and B. Aditya Prakash. PINNsFormer: A Transformer-Based Framework For Physics-Informed Neural Networks. *arXiv preprint arXiv:2307.11833*, 7 2023. URL http://arxiv.org/abs/2307.11833.

Yanzhe Zhang, Tao Yu, and Diyi Yang. Attacking vision-language computer agents via pop-ups, 2024.

Atharv Singh Patlan, S Ashwin Hebbar, Prateek Mittal, and Pramod Viswanath. Real ai agents with fake memories: Fatal context manipulation attacks on web3 agents. *arXiv preprint arXiv:2503.1624*, 7 2025. URL https://arxiv.org/abs/2503.16248.

Jianing Qiu, Lin Li, Jiankai Sun, Hao Wei, Zhe Xu, Kyle Lam, and Wu Yuan. Emerging cyber attack risks of medical ai agents. *arXiv preprint arXiv:2504.03759*, 4 2025. URL https://arxiv.org/pdf/2504.03759.

Edgesca. Vulnerability statistics report | mean time to remediate data (mttr), 2023. URL https://info.edgescan.com/vulnerability-statistics-li23.

Pradipta Kishore Chakrabarty. Adversarial attacks on agentic ai systems: Mechanisms, impacts, and defense strategies. *International Journal of Science and Research (IJSR)*, 14:1367–1369, 4 2025. doi:10.21275/SR25417074844.

Romuald Hoffmann. Markov models of cyber kill chains with iterations. *2019 International Conference on Military Communications and Information Systems, ICMCIS 2019*, 5 2019. doi:10.1109/ICMCIS.2019.8842810.

Appendix

n*Code and Implementation

Complete implementation available at: `https://github.com/GingerSpacetail/pinnsformer`

Key components:

- `ai_epidemiology_model.py`: Core SEIR-PINN implementation
- `bifurcation`$_{a}nalysis.py$`empirical_parameter_estimation.py` : $Parameterestimationframework$
- `realistic_ai_epidemiology_scenarios.py`: Scenario analysis tools
- `real_time_monitoring.py`: Monitoring framework (future work)

LLM Usage Declaration

This research was conducted with assistance from Claude 3.5 Sonnet for:

- Sources summarization, introduction improvement
- Code debugging for PINNs implementation
- Extensive technical documentation

The core theoretical insights, empirical analysis, and framework development represent original research contributions by the authors.