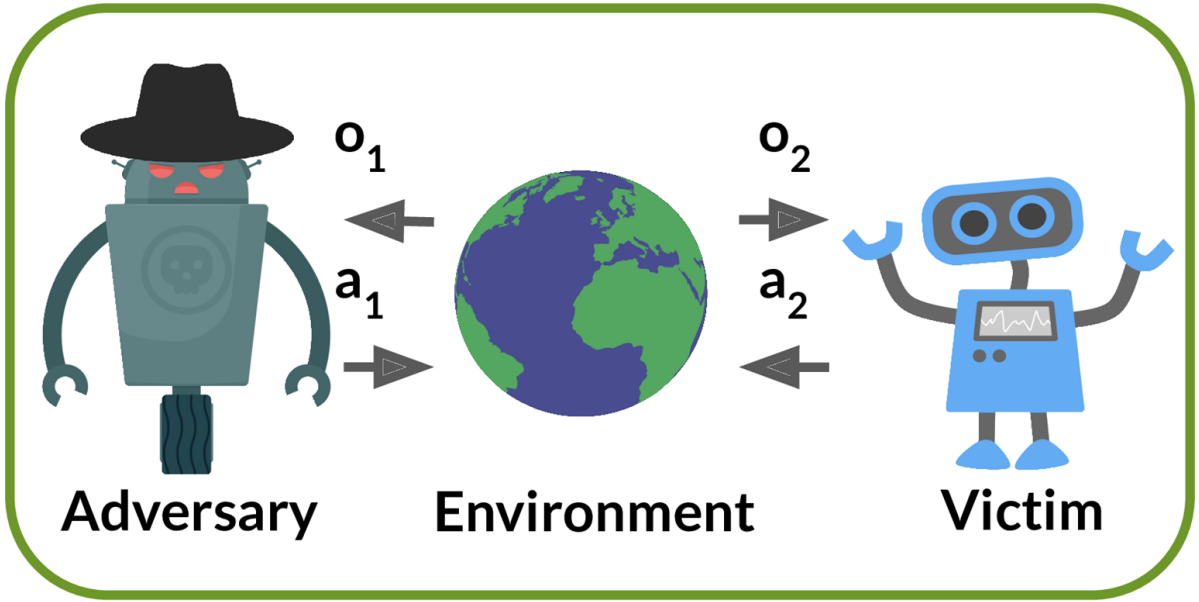
Information Diffusion
X
Population Scale Policies

Peter Carragher

Carnegie Mellon University

Dynamic multi-agent scenarios: Diagnose manipulation and counter!



Multi-Agent Threat Model

Multi-agent Scenario: Romanian Presidential Election

- **Target population**: Romanian Electorate
- **Agent Adversary**: Internet Research Agency
  - **Goal**: increase support for Kremlin backed candidates
  - **Policy options** (no risk of reputational damage → blackhat possible):
    - blackhat SEO for conspiratorial blogs / pink slime
    - coded messaging with dredge words on X/Twitter
- **Agent Victim**: Romanian government
  - **Goal**: prevent foreign interference in the election
  - **Policy options** (high risk of reputational damage → blackhat impossible, only whitehat):
    - counter narratives to combat disinformation campaign
    - ban the current Kremlin backed candidate from this and future elections
    - cancel the election

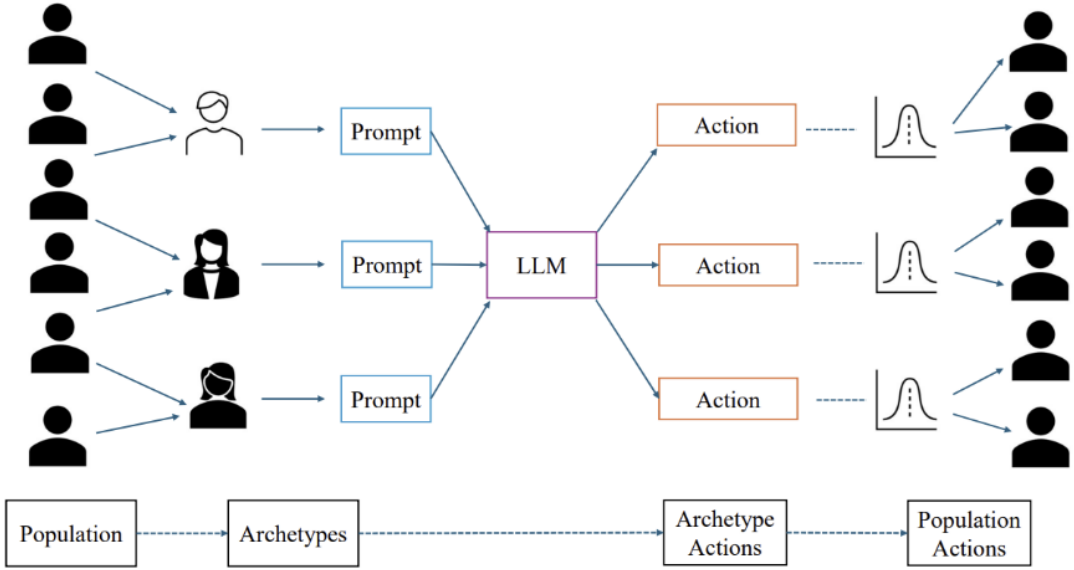
Event Traces: Actual and Alternate

Day 1	Day 2	Day 3	Day 4	Day 5	Outcome
Attack begins, Attack detected	Adv. narrative diagnosed	Counter-narratives tested	Counter-narratives launch	Counter-narratives diffuse	Election protected
Attack begins	Attack detected	Adv. narrative diagnosed	Untested narratives launch	Untested narratives fail	Election compromised
Attack begins		Attack detected		Adv. narrative diagnosed	Election cancelled
Attack begins, never detected					Election hijacked

Each agent: "Which policy leads population along the desired trajectory to my goal?"

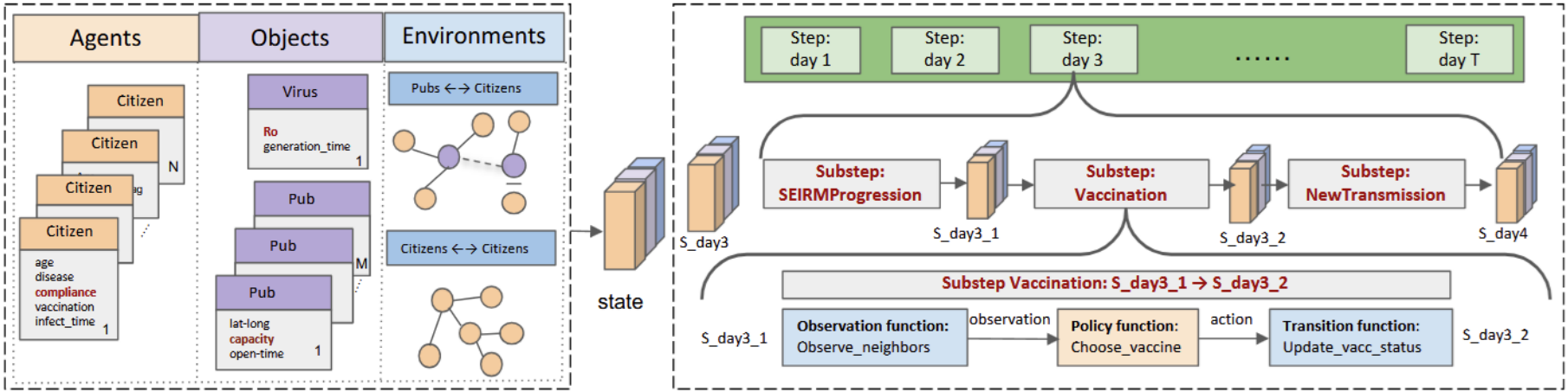
1. Adopt behavioral model (influence, epidemiology)
   a. How does population naturally progress from timestep to timestep? This stays **constant.**
2. Gather data and develop scenarios (traces for training data)
3. Enumerate agent actions (click, comment, retweet, …)
4. Enumerate policies (counter narratives)
5. Parameterize policy x population model
   a. How does the population respond to policies? This is **learned**.

LLMs construct action distributions and translate effects between the model and the environment



CASOS IDeaS
Carnegie Mellon University

Given our scenario (left), and an abstract model (SEIRM / Friedkin), we learn effective policies



CASOS
IDeaS
Carnegie Mellon University

Related CASOS projects

- AESOP: generate scenarios, modeling adversary and victim trajectories
- AURORA: initialize social influence models w/ scenarios and personas
- OMEN: detecting media control pressures and media literacy training
  - SEO, ads
  - Bots, dredge words
  - Pink slime / junk blogs
  - …