SAND2024-10202PE

Sandia National Laboratories

MADmax

Multi-Agent Trust Dynamics and Influence Maximization

Ace Sorensen, R&D Computer Science
ahsoren@sandia.gov
Sandia National Labs

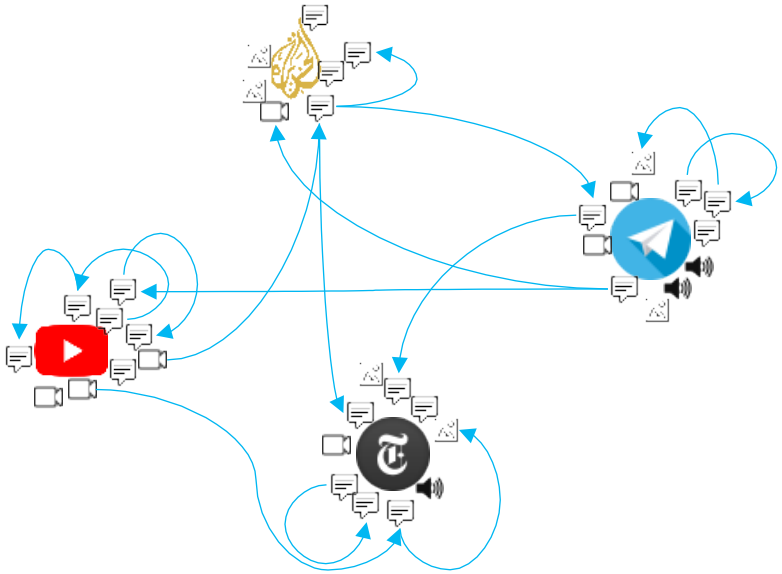
Matthew Sweitzer, Asmeret Naugle, Casey Doyle, Daniel Krofcheck

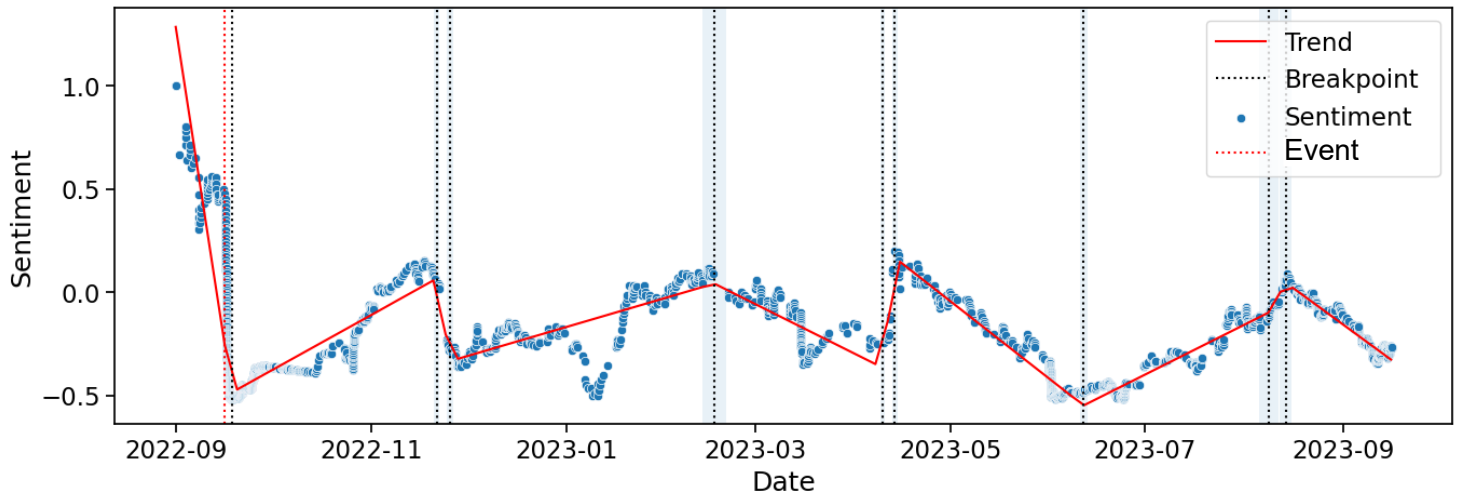
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

07.20.2024

Problem

Influence is more visible now than ever





What about the parts that aren’t?

What makes influencers successful?
How do they achieve their goals?

Critical Elements

1 Trust. What is the impact of trust on opinion dynamics?

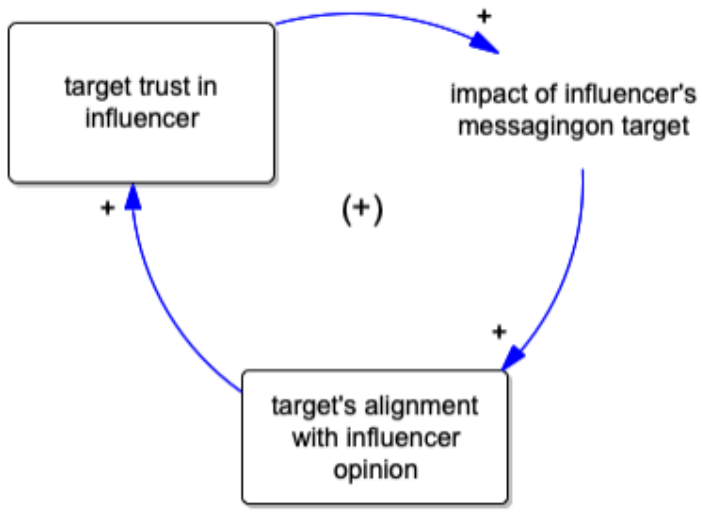
2 Strategies. What strategies do influencers employ to achieve their goals?

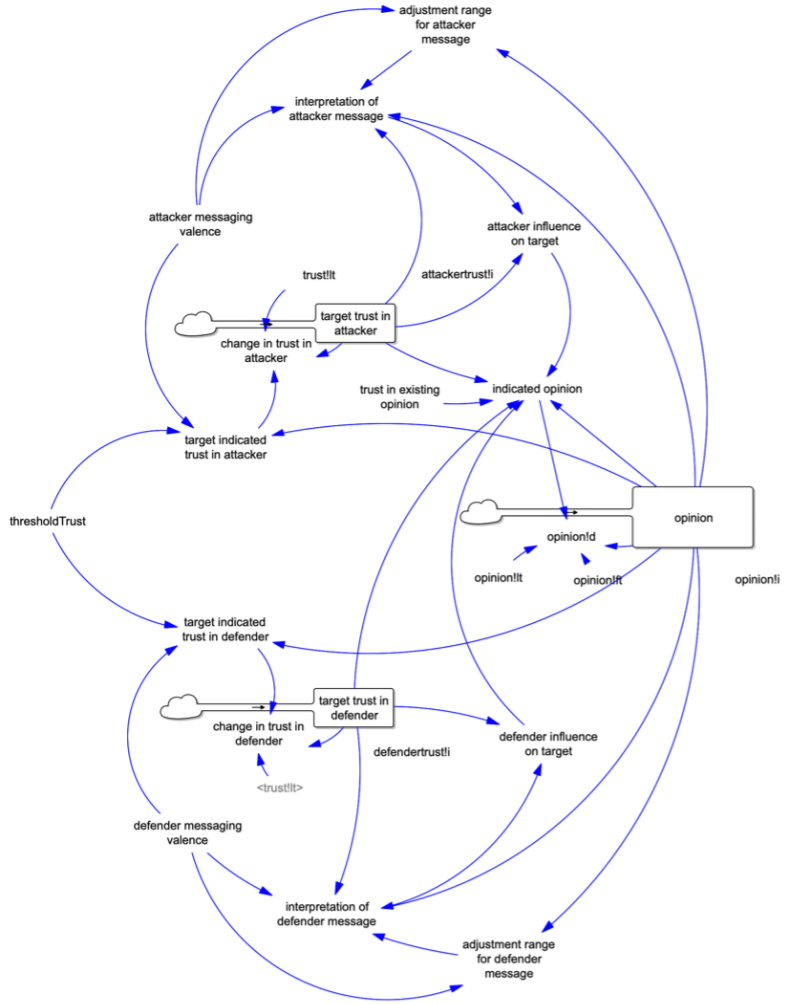
Simulation: Big Picture

Interactions: Agent Based Model

Dynamics: System Dynamics (trust and opinion)

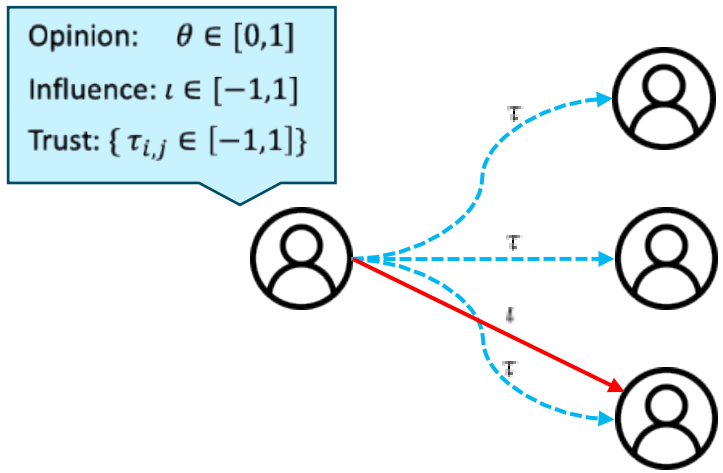
Strategy: Multi-agent Reinforcement Learning (MARL)





Simulation: Influence

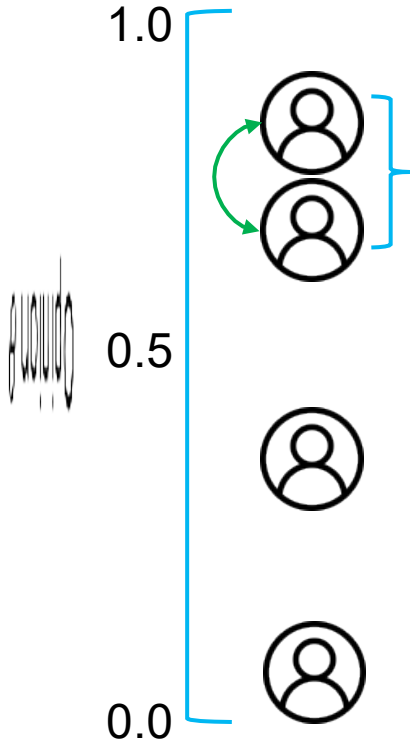
Direct



Intentional optimal control formulation with influence as an offset from the target's opinion.

$$\theta_{expressed} = \theta_{target} + \iota$$

Indirect



Peer influence $|\theta_j - \theta_i| < \epsilon$

Interesting side effect – stubborn agents because individuals gravitate towards peer opinions.

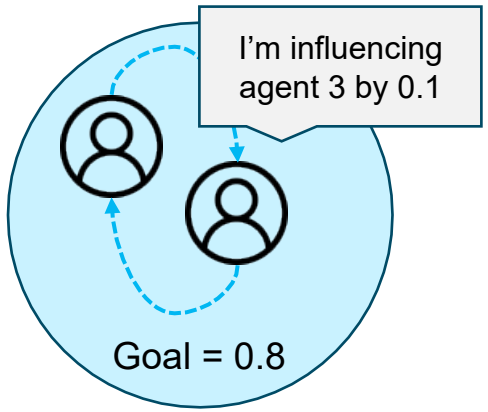
Stubborn Agents

Simulation: Agent Environment Cycle

Each Step:

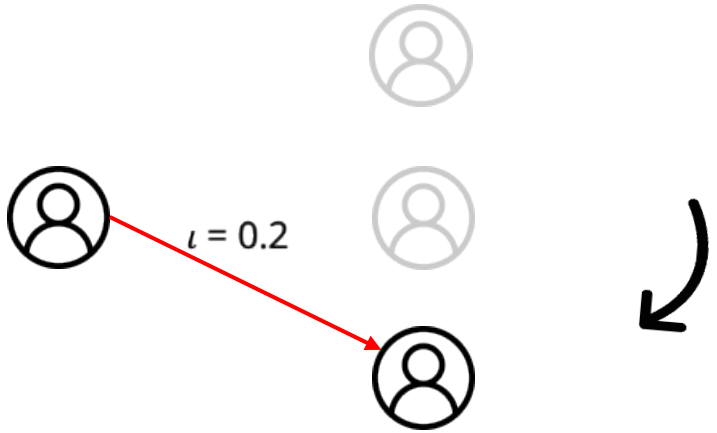
1. Observe All agent opinions

Team: Shared goal + message passing



2. Act

Action: Target + Influence



3. Reward

Simulation: Configuration Training and Evaluation

Configuration

Default agents: random target, smoothed influence

Simulation Setup:
Interested in **intra-team coordination** and **inter-team competition**

- Population size: 8
- Opinions initialized random uniform
- Trust is 0

Training

Curriculum: Trust → Individual Influence → Community Influence

Autocurricula: compete vs self, random goals

Evaluation

Consistency: MARL opinions 0.8. Opposition 0.2

Simulation: Goal

Challenge

- $1.8\,x\,10^{132}$ **possible** action sequences
- **Training directly** on the goal **fails**

Approach

- **SB3 PPO** with agent perspective reward
- Dynamic rewards for **curriculum learning** in stages

Goal

- Sway opinion to your team's goal, **end with largest cluster** (DBSCAN)
- **Average opinion closer** to your team than any other

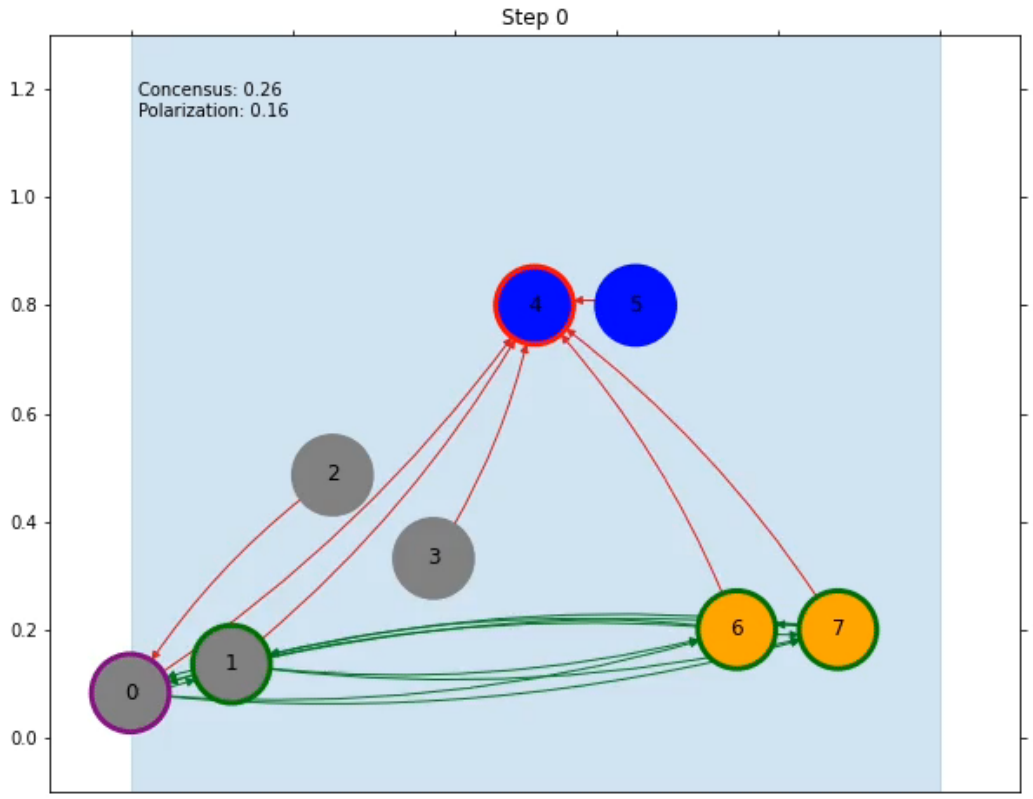
Simulation: In Action

Nodes

- **Blue:** goal 0.8
- **Orange:** goal 0.2
- **Grey:** no affiliation

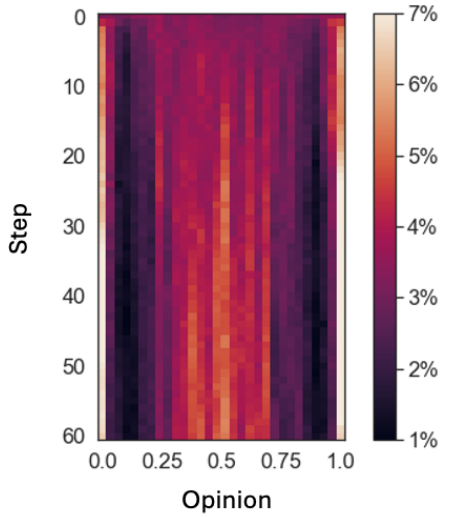
Edges (and node border)

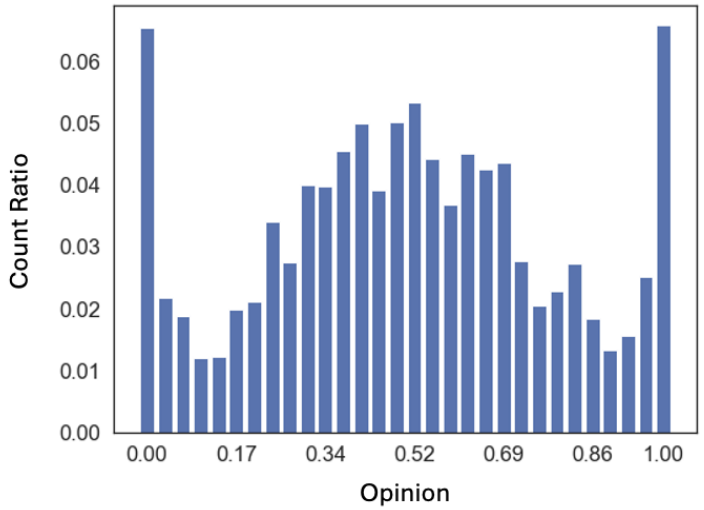
- **Green:** peer influence
- **Red:** direct influence



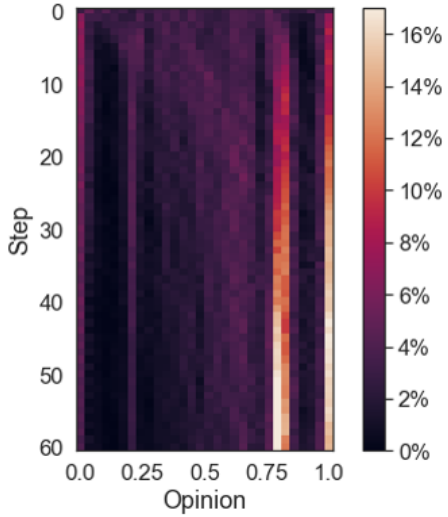
Simulation: Results

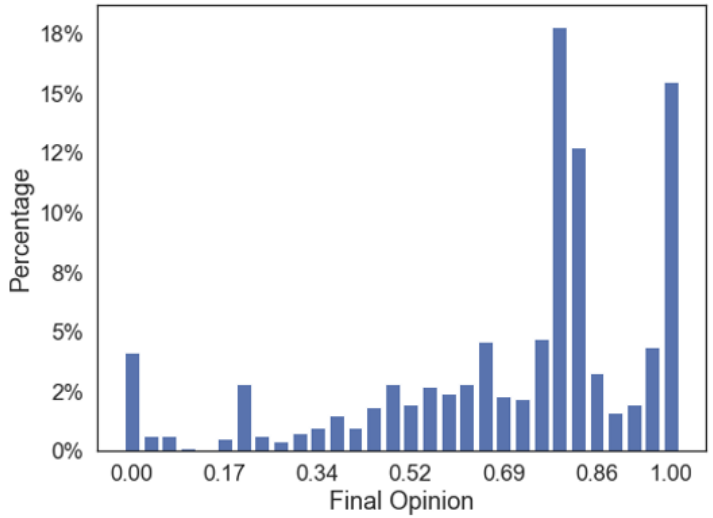
Without MARL agents



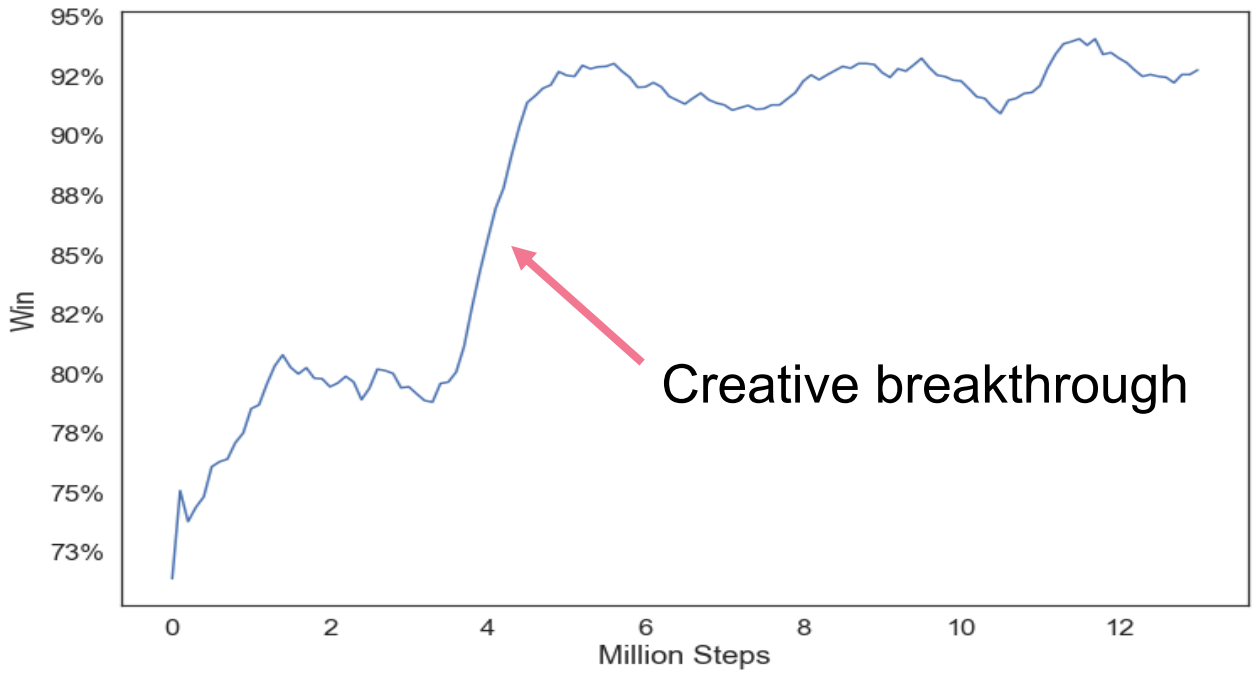


With MARL agents



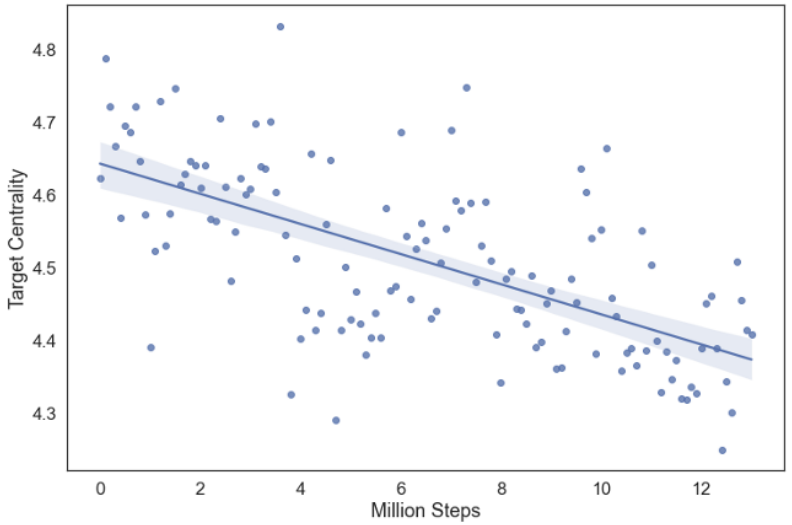


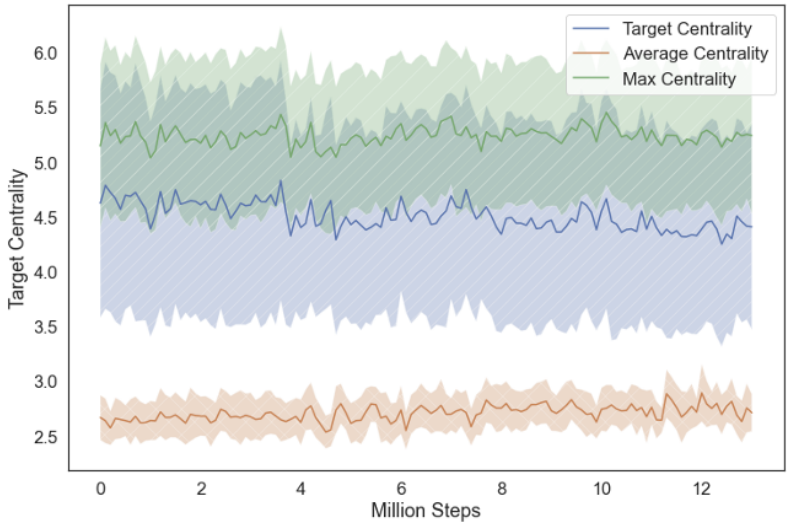
Simulation: Results

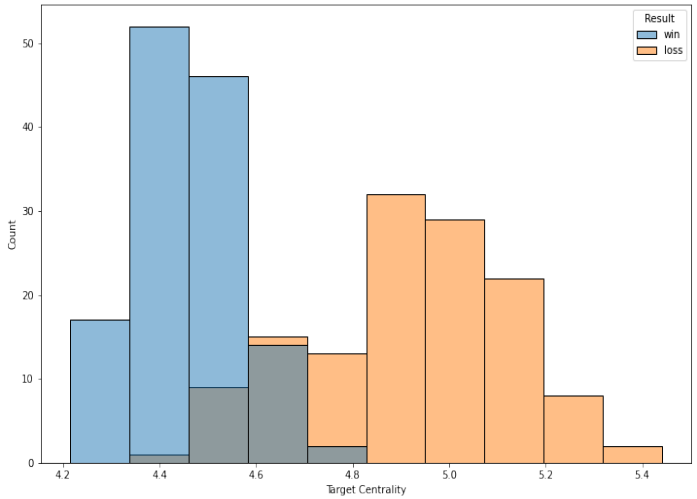


Do creative breakthroughs mark identifiable shifts in strategies?

Simulation: Centrality in Moderation

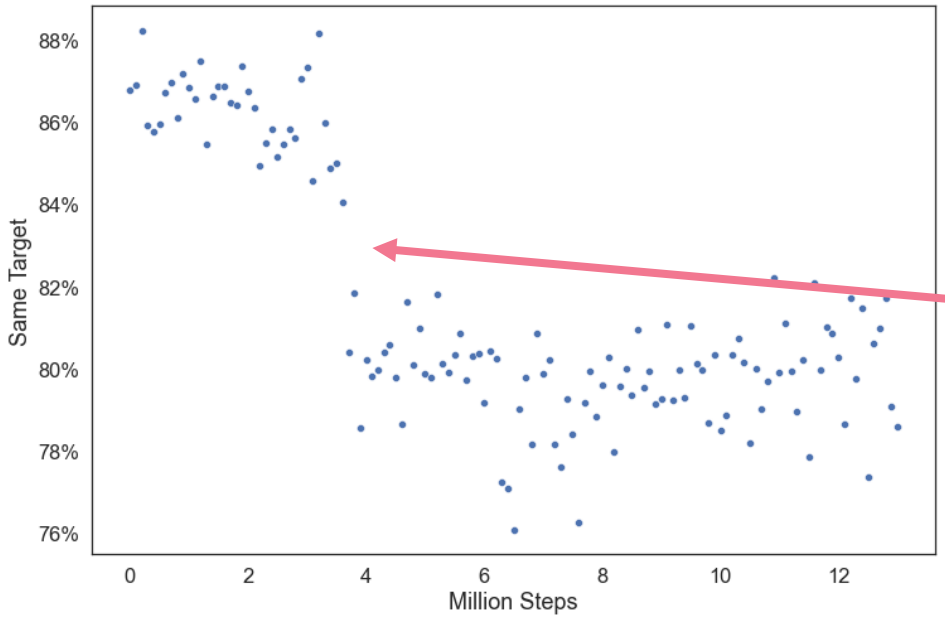


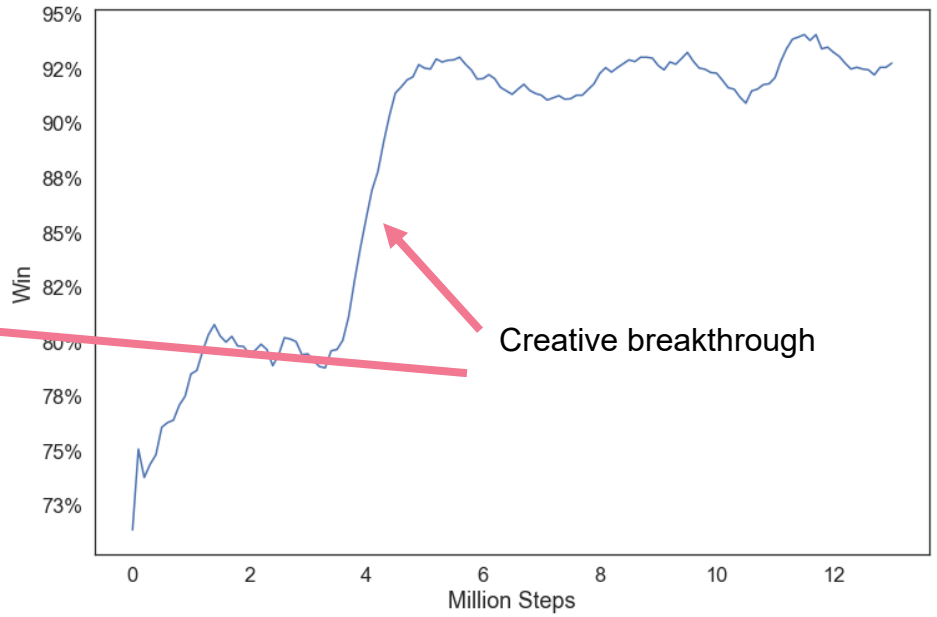




Gradual reduction in target centrality doesn't explain creative breakthrough.
Trend of targeting moderately high centrality agents is worth further investigation

Simulation: Divide and Conquer





A change in coordination strategies resulted in the creative breakthrough. Agents became better at influencing simultaneously

Simulation: In Action

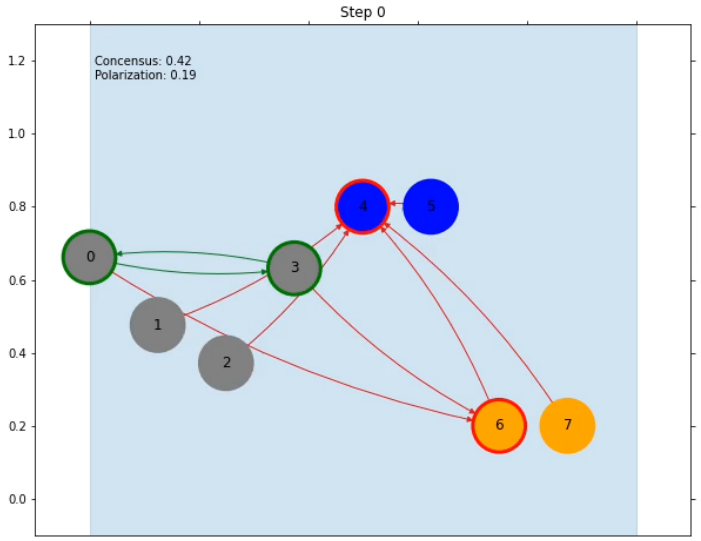
After 10 million steps

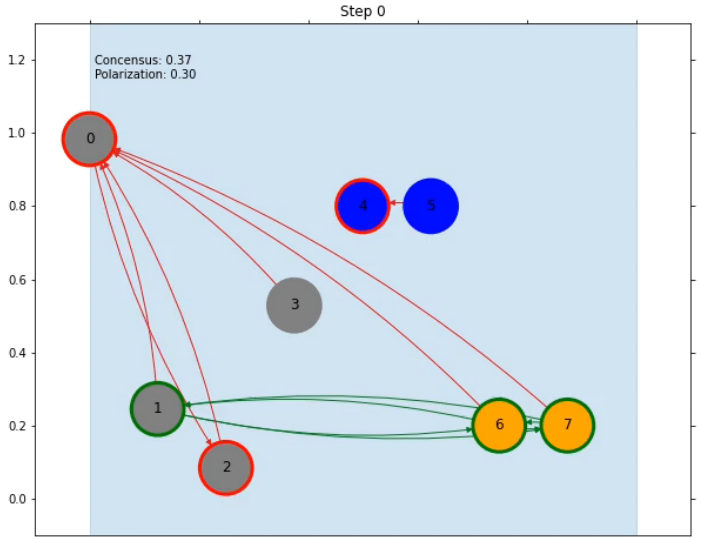
Nodes

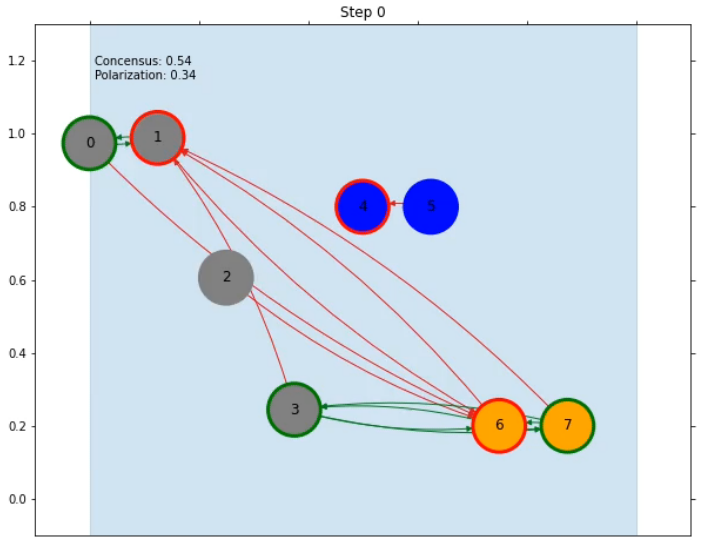
- **Blue:** goal 0.8
- **Orange:** goal 0.2
- **Grey:** no affiliation

Edges (and node border)

- **Green:** peer influence
- **Red:** direct influence







Simulation: Discussion

Trust insights: Trust decreased for the influencers, but increased between the other agents

Strategy insights: We successfully trained agents who exhibited believable behaviors we can use as markers to look for influence

System insights: We observed a rise in extremism when influencers participated in contrast so simulations without influencers

Future work: We plan to use insights to improve fidelity of real world analysis

Questions?