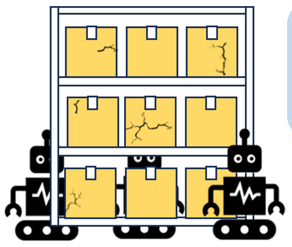
Mitigating Negative Side Effects in Multi-Agent Systems using Blame Assignment

Pulkit Rustagi and Sandhya Saisubramanian
Oregon State University

Motivation



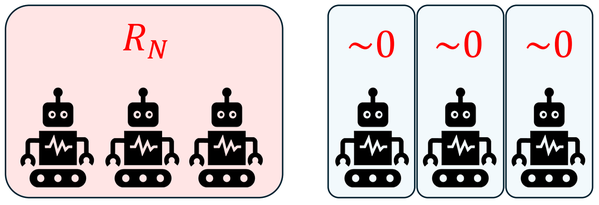
Objectives and rewards in multi-agent systems are rarely completely specified

May produce negative side effects after deployment

Negative Side Effects (NSE)

Unanticipated, undesirable consequences of multiple agents acting together

- Objective specifications are always *incomplete*
- NSEs are often discovered *after* deployment
- Associated penalties are reported *collectively*

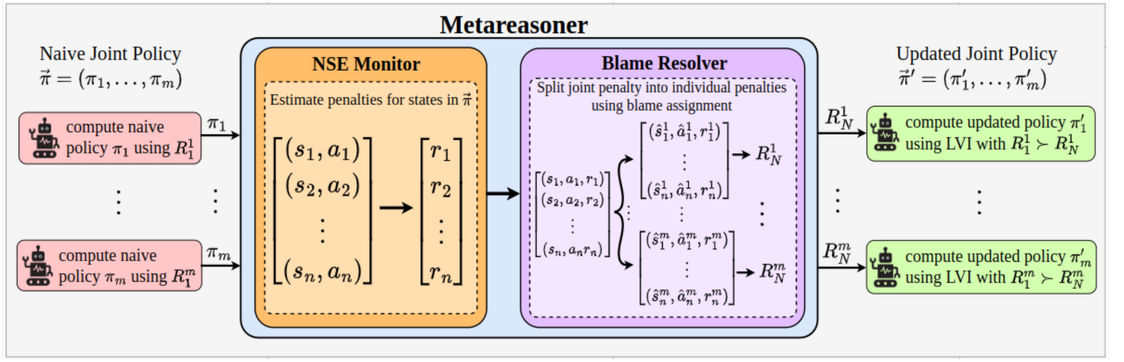


Challenge:

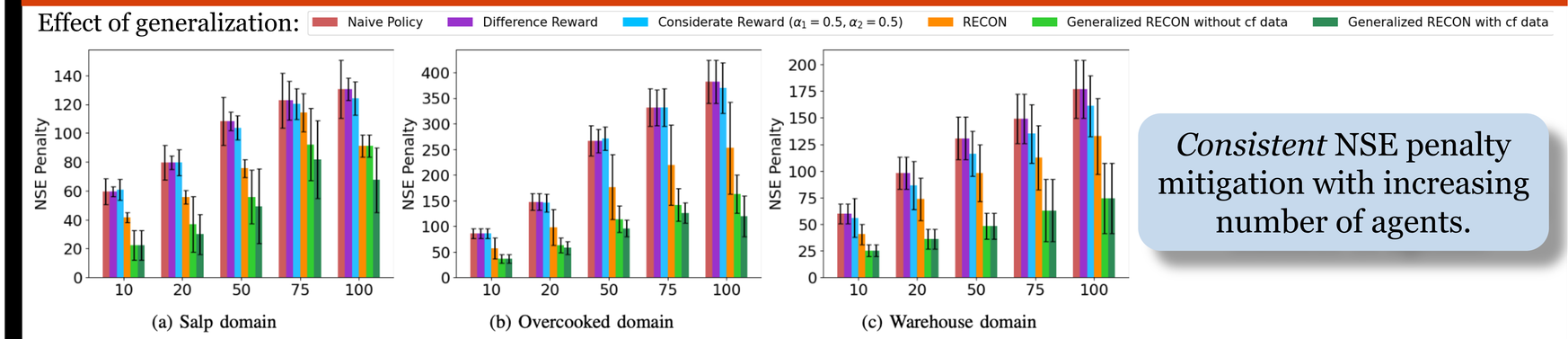
- Mitigating jointly reported penalty requires solving a *coordination* problem, traditionally done using:
  - ➢ Centralized computation
    - *Not scalable* to higher number of agents
  - ➢ Communication
    - *Not feasible* in every setting

Approach

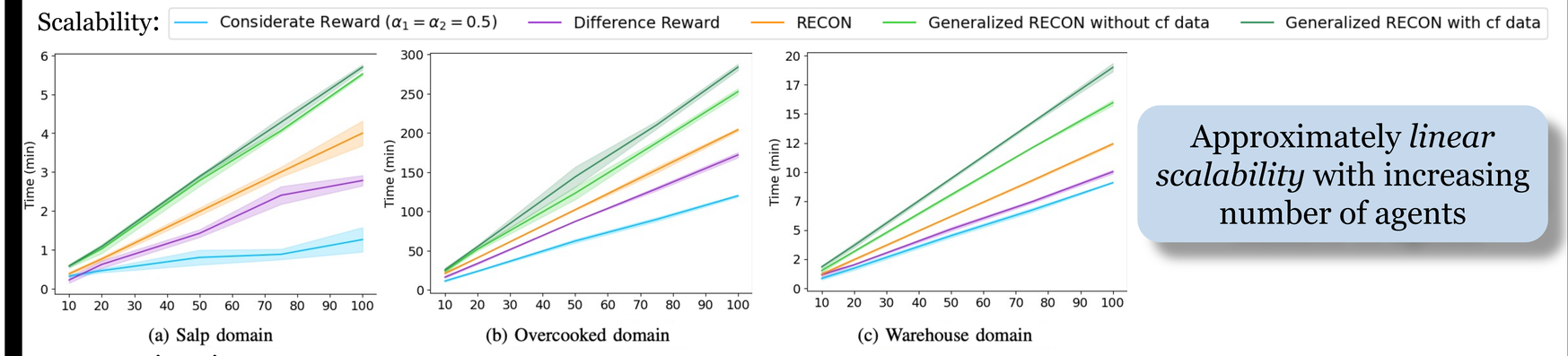
1. Compute naïve policy
2. Metareasoner
   - *i.* *Collect* policy
   - *ii.* *Estimate* joint NSE penalty
   - *iii.* *Attribute* blame to each agent from estimated joint penalty
   - *iv.* *Compute* penalty function for each agent using attributed blames
3. Compute new policy



Results and Discussion



Consistent NSE penalty mitigation with increasing number of agents.



Approximately *linear scalability* with increasing number of agents

Future Directions:

- Extending approach to *tightly coupled* tasks
- Exploiting agent dependencies to leverage *complimentary skills to mitigate* NSE.