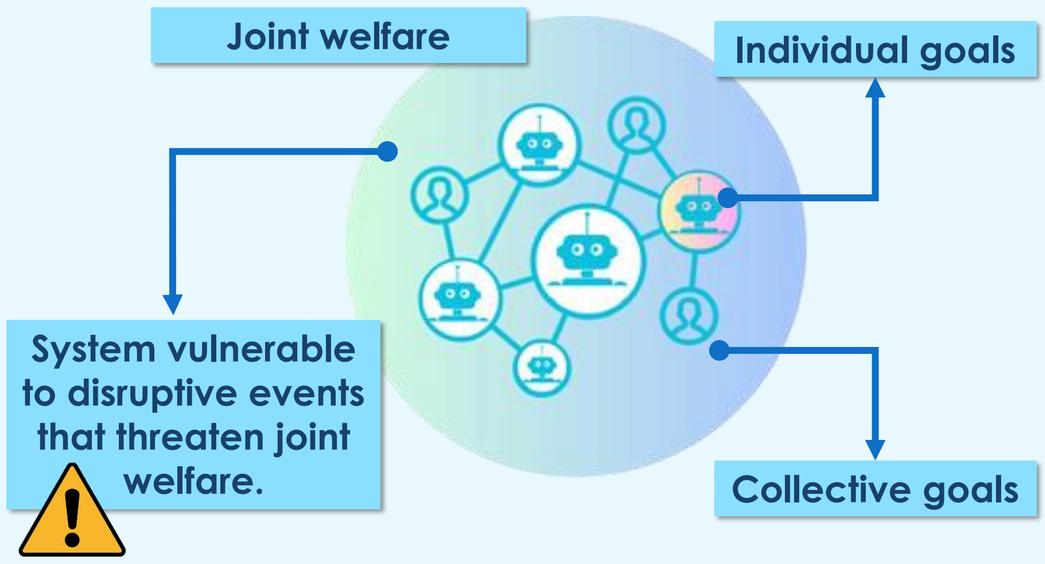
Cooperative Resilience in Multi-Agent AI Systems:

From Definition to Reward Inference

Universidad de los Andes Colombia

Manuela Chacon-Chamorro, Luis Felipe Giraldo and Nicanor Quijano

Introduction



Multi-agent AI systems with mixed motives often face disruptions that threaten collective welfare. Designing systems that can protect joint welfare under such disruptions requires **cooperative resilience.**

Key Challenge: To enable multi-agent AI systems to withstand disruptions and sustain collective, well-being even in the face of these disruptions.

To address this, we propose a framework to **define** and **measure** cooperative resilience, identify the conditions that **cause** it, and use this understanding to **design** systems that promote it.

Our Proposal: Define. Measure. Understand causes. Design for resilience.

Definition

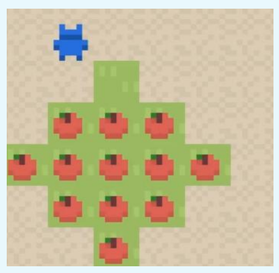
Cooperative Resilience is the ability of a system that involves the collective action of individuals, be they humans, machines, or both, to anticipate, prepare for, resist, recover from, and transform in face to disruptive events that pose a risk to their joint welfare.

Measure

Our methodology evaluates how well agents sustain collective welfare across disruptions by tracking performance over time.

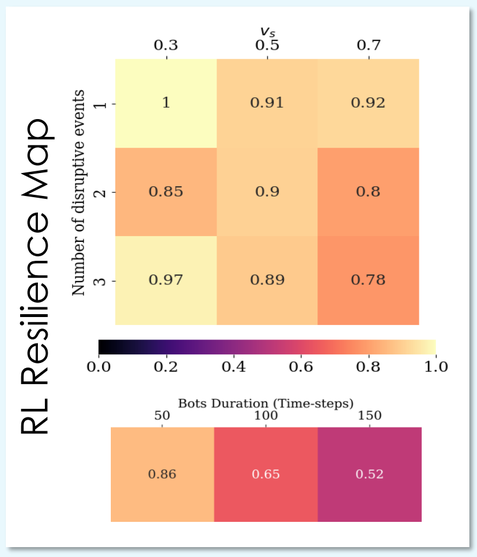
- **Stage I**: Capture agents' behaviors and outcomes over time, under normal and disrupted conditions, for key indicators (e.g., resource use, equality, coordination).
- **Stage II**: We compared the reference and performance curves. With the comparison, we compute a resilient score.
- **Stage III**: Organize metrics across time to analyze recovery and adaptation dynamics.
- **Stage IV**: Aggregate the information into a single Cooperative Resilience Score.

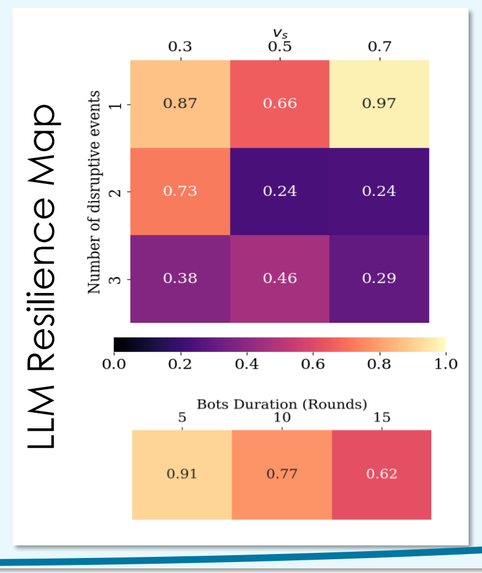
Experiments



Experiments are conducted in Melting Pot, we focus on the Common Harvest scenario, where agents must coordinate to sustainably harvest a shared resource.

- **Two types of agents:** RL and LLM
- **Two set of disruptive events:**
  (I) Apples disappear.
  (II) Introduction of a bot with unsustainable behavior.

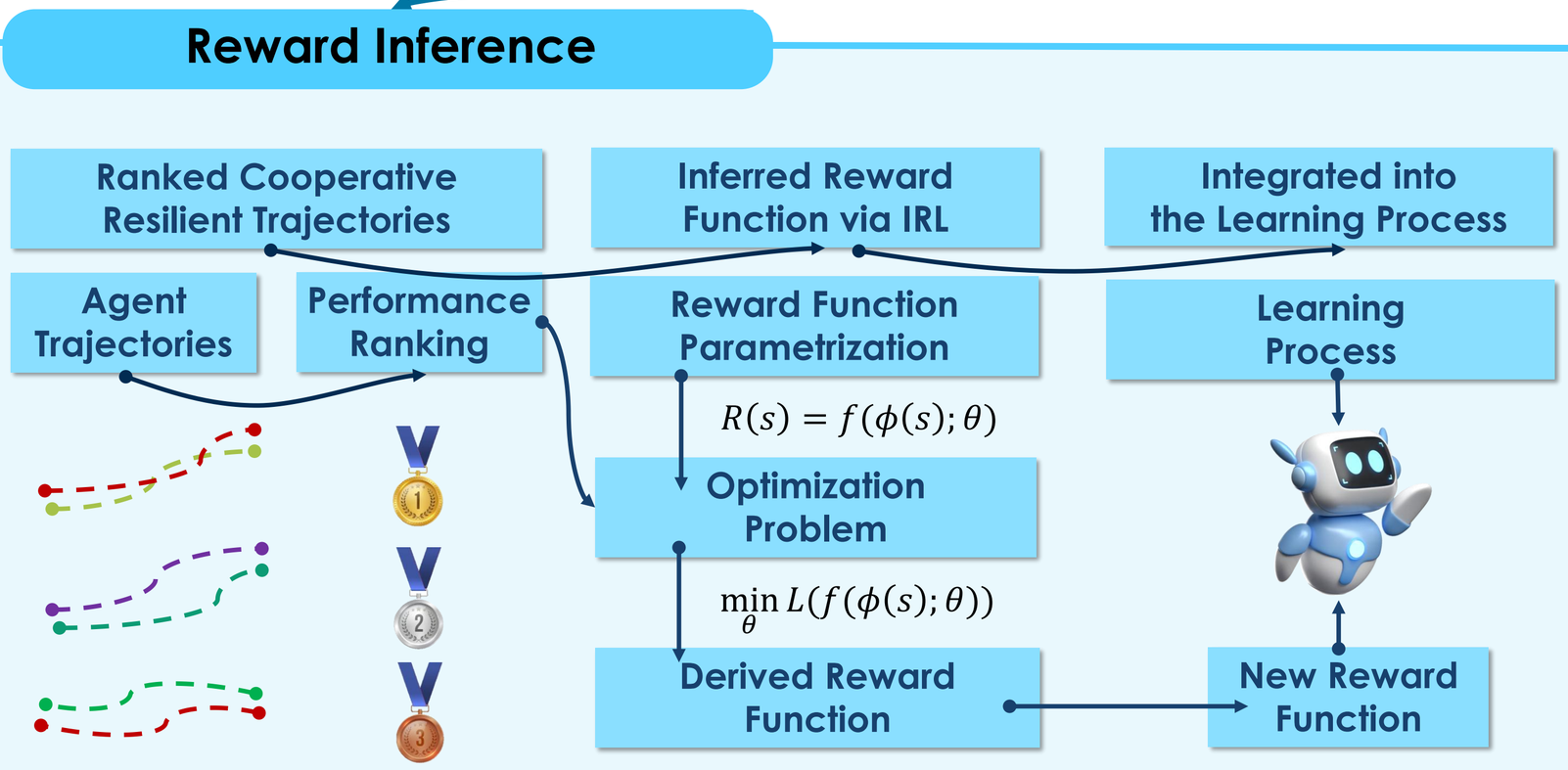




Using scores to infer rewards

Using our cooperative resilience metric, we rank agent trajectories and apply inverse reinforcement learning to uncover the reward signals behind resilient behaviors.

Reward Inference



Conclusions

- We proposed a definition and a metric to quantify cooperative resilience in multi-agent systems.
- We used it to create **resilience maps** comparing how RL and LLM agents respond to two set of disruptive events:
- We aim to infer reward functions from behaviors ranked using our resilience metric, to support the design of **resilient cooperative** multi-agent systems.

Acknowledgments: We thank the UniAndes–DeepMind Scholarship 2023, Cooperative AI Foundation (CAIF) for supporting participation in the summer school, and the Universidad de los Andes for institutional support.

References: This poster is based on the paper *Chacón-Chamorro, M., Giraldo, L. F., Quijano, N., Vargas-Panesso, V., González, C., Pinzón, J. S., Manrique, R., Ríos, M., Fonseca, Y., Gómez-Barrera, D., & Perdomo-Pérez, M. (2025). Cooperative Resilience in Artificial Intelligence Multiagent Systems. IEEE Transactions on Artificial Intelligence, Early Access. Full paper QR code.*