

Resiliencia en sistemas multiagentes y la experiencia de un doctorado

Manuela Viviana Chacón Chamorro

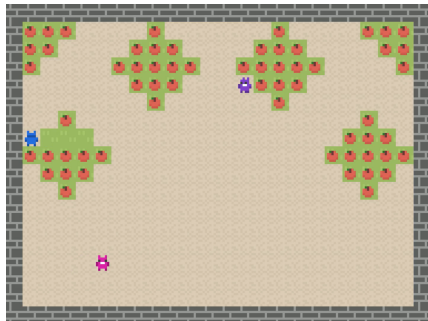
Candidata a Doctora en Ingeniería
Universidad de los Andes

Manizales, Noviembre 27, 2025

Motivación

Individual

→
Meta: Maximizar el número de manzanas recolectadas.



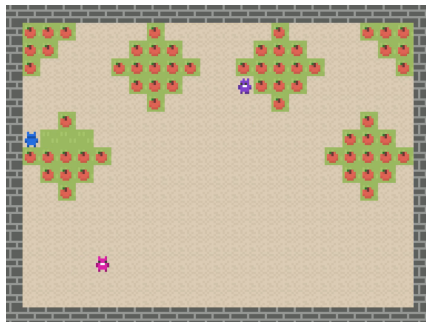
Collectiva

←
Meta: Evitar que se consuman los recursos compartidos.

Motivación

Individual

→
Meta: Maximizar el número de manzanas recolectadas.



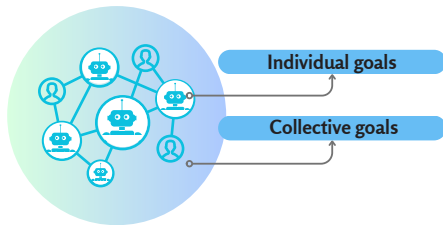
Collectiva

←
Meta: Evitar que se consuman los recursos compartidos.

Ahora pensemos aparece una **perturbación**. Si se desea mantener el bienestar conjunto y preservar la cooperación incluso con estas interrupciones entonces estamos hablando de **cooperative resilience**.

[1] Agapiou, J. P., Vezhnevets, A. S., Dueñez-Guzmán, E. A., Matyas, J., Mao, Y., Suneag, P., et al. (2022). *Melting Pot 2.0*. arXiv preprint arXiv:2211.13746.

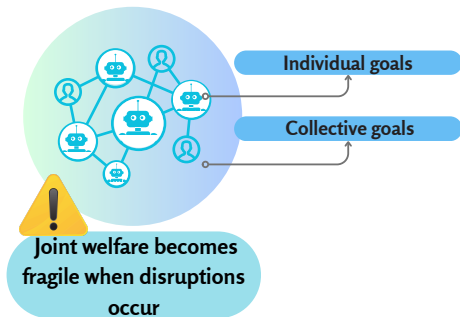
Motivación



[2] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., et al. (2020). *Open problems in cooperative AI*. arXiv preprint arXiv:2012.08630.

[3] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak et al., *Multi-agent risks from advanced AI*, arXiv preprint arXiv:2502.14143, 2025.

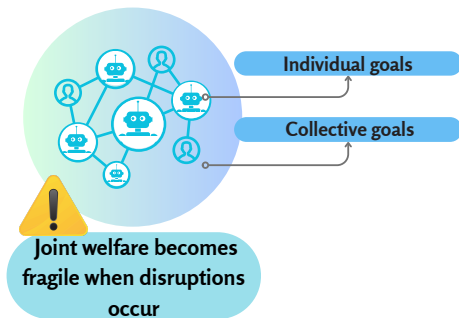
Motivación



[2] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., et al. (2020). *Open problems in cooperative AI*. arXiv preprint arXiv:2012.08630.

[3] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak et al., *Multi-agent risks from advanced AI*, arXiv preprint arXiv:2502.14143, 2025.

Motivación

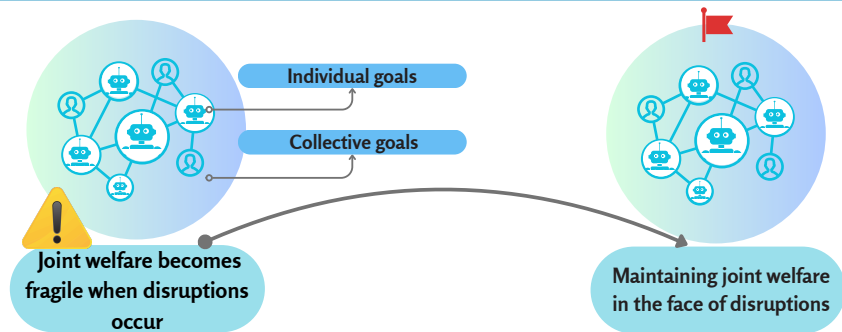


Reto clave: Generar sistemas multi-agente de IA que puedan tolerar las interrupciones y sostener el bienestar colectivo cuando esto suceda.

[2] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., et al. (2020). *Open problems in cooperative AI*. arXiv preprint arXiv:2012.08630.

[3] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak et al., *Multi-agent risks from advanced AI*, arXiv preprint arXiv:2502.14143, 2025.

Motivación



Reto clave: Generar sistemas multi-agente de IA que puedan tolerar las interrupciones y sostener el bienestar colectivo cuando esto suceda.

[2] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., et al. (2020). *Open problems in cooperative AI*. arXiv preprint arXiv:2012.08630.

[3] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčič et al., *Multi-agent risks from advanced AI*, arXiv preprint arXiv:2502.14143, 2025.

Outline

- 1 ¿Qué queremos resolver?
- 2 ¿En que parte estamos ahora?
 - Definir resiliencia cooperativa
 - Medir resiliencia cooperativa
 - Promoviendo resiliencia cooperativa
- 3 ¿Qué hemos logrado?
- 4 ¿Qué tal la experiencia?
 - Experiencia Ph.D

¿Qué queremos resolver?

General

Pregunta

¿Cómo emerge la resiliencia cooperativa en MAS-IA?

Objetivo

Descubrir los mecanismos que permiten a los MAS-IA ser cooperativamente resilientes.

1

¿Qué queremos resolver?

2

¿En que parte estamos ahora?

3

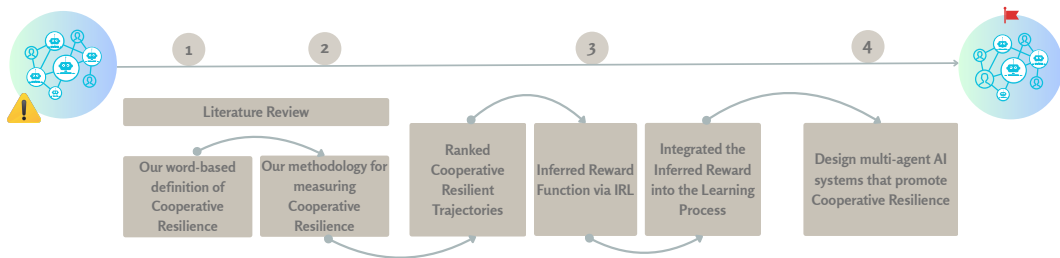
¿Qué hemos logrado?

4

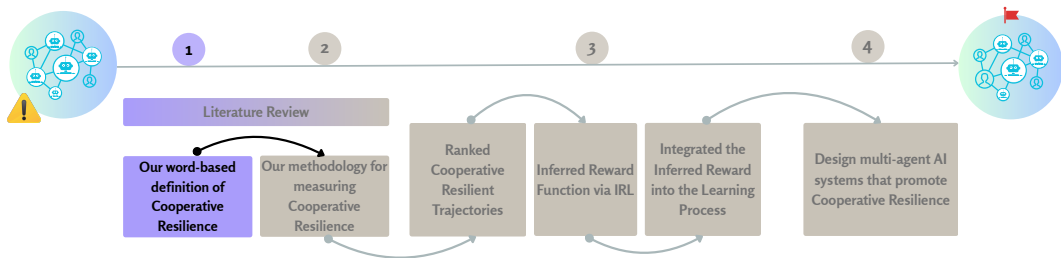
¿Qué tal la experiencia?

¿En que parte estamos ahora?

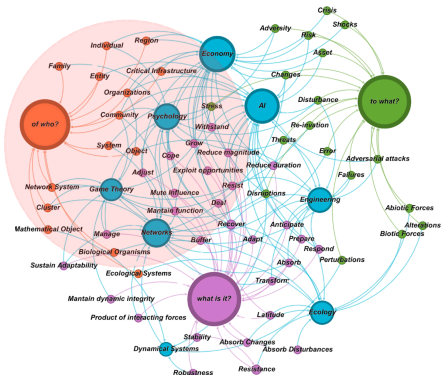
¿Dónde estamos?



¿Dónde estamos?

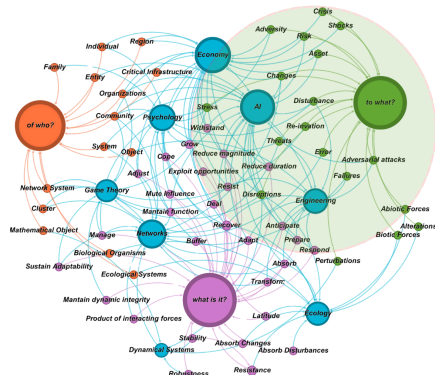
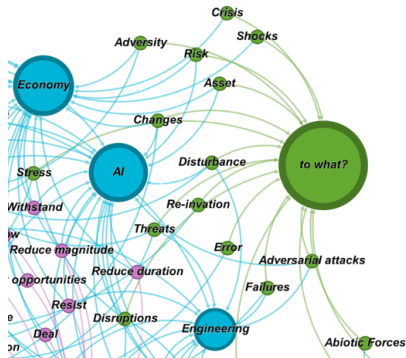


Manuela Chacon-Chamorro



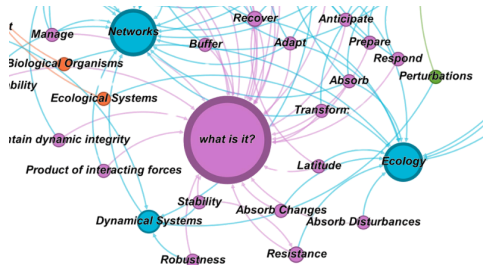
- 6 / 37

Resiliencia en otras áreas



- [9] A. Rose, *Economic resilience to natural and man-made disasters: Multidisciplinary origins and contextual dimensions*, Environmental Hazards, vol. 7, no. 4, pp. 383–398, 2007.
- [10] L. Briguglio, G. Cordina, N. Farrugia, and S. Vella, *Economic vulnerability and resilience: concepts and measurements*, in Measuring Vulnerability in Developing Countries. Routledge, 2014, pp. 47–65.
- [11] J. Simmie and R. Martin, *The economic resilience of regions: towards an evolutionary approach*, Cambridge Journal of Regions, Economy and Society, vol. 3, no. 1, pp. 27–43, 2010.

Resiliencia en otras áreas



- [11] J. Gao, B. Barzel, and A.-L. Barabási, *Universal resilience patterns in complex networks*, Nature, vol. 530, no. 7590, pp. 307–312, 2016.
- [12] X. Liu, D. Li, M. Ma, B. K. Szymanski, H. E. Stanley, and J. Gao, *Network resilience*, Physics Reports, vol. 971, pp. 1–108, 2022.
- [13] G. Como, *On resilient control of dynamical flow networks*, Annual Reviews in Control, vol. 43, pp. 80–90, 2017.
- [14] C. S. Holling, *Resilience and stability of ecological systems*, Annual Review of Ecology and Systematics, vol. 4, no. 1, pp. 1–23, 1973.
- [15] A. Hastings, *Transient dynamics and persistence of ecological systems*, Ecology Letters, vol. 4, no. 3, pp. 215–220, 2001.

Nuestra definición

Resiliencia cooperativa

Resiliencia cooperativa es la habilidad de un sistema, que involucra la acción colectiva de individuos, los cuales pueden ser artificiales o humanos, para anticiparse, prepararse, resistir, recuperarse, resistir y transformarse en respuesta a eventos disruptivos que amenazan su bienestar conjunto.

Nuestra definición

Cooperative Resilience

Resiliencia cooperativa es la habilidad de un sistema, **que involucra la acción colectiva de individuos, los cuales pueden ser artificiales o humanos**, para anticiparse, prepararse, resistir, recuperarse y transformarse en respuesta a eventos disruptivos que amenazan su bienestar conjunto.

Nuestra definición

Cooperative Resilience

Resiliencia cooperativa es la habilidad de un sistema, que involucra la acción colectiva de individuos, los cuales pueden ser artificiales o humanos, para anticiparse, prepararse, resistir, recuperarse y transformarse en respuesta a **eventos disruptivos que amenazan su bienestar conjunto**.

Nuestra definición

Cooperative Resilience

Resiliencia cooperativa es la habilidad de un sistema, que involucra la acción colectiva de individuos, los cuales pueden ser artificiales o humanos, **para anticiparse, prepararse, resistir, recuperarse y transformarse** en respuesta a eventos disruptivos que amenazan su bienestar conjunto.

1

¿Qué queremos resolver?

2

¿En que parte estamos ahora?
– Medir resiliencia cooperativa –

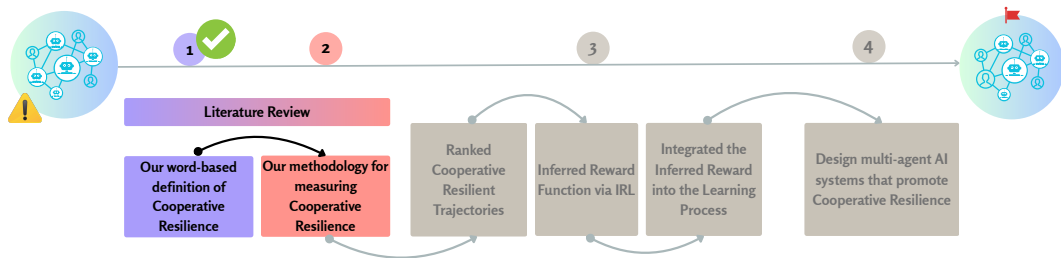
3

¿Qué hemos logrado?

4

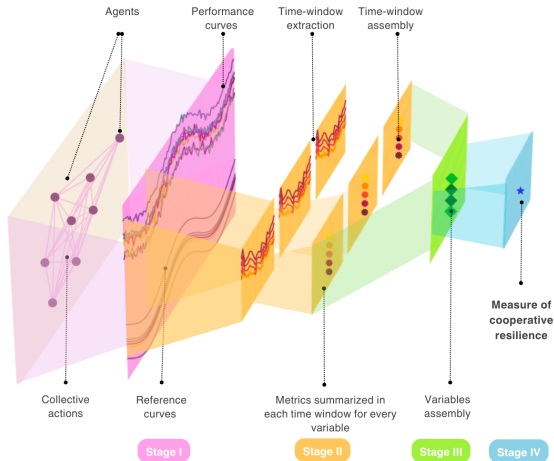
¿Qué tal la experiencia?

¿Dónde estamos?



Metodología

- **Etapa I.** Capturar curvas de referencia y desempeño.
- **Etapa II.** Calcular metricas de resumen.
- **Etapa III.** Agregar medidas en el tiempo.
- **Etapa IV.** Agregar medidas en los indicadores.

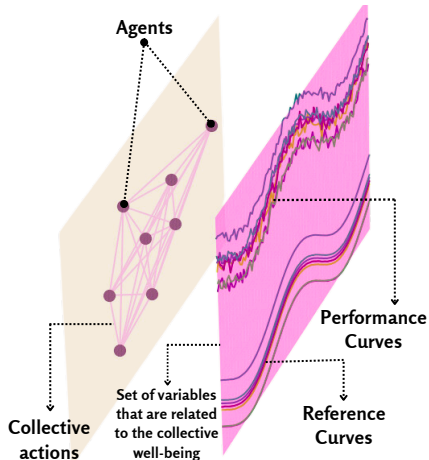


Metodología: Etapa I

En la primera etapa el proceso es definir un conjunto de indicadores¹⁶ que se relacionen con el bienestar colectivo de los agentes en el sistema¹⁷. Luego se estiman las curvas de estos indicadores sin interrupción (*reference*) con interrupción (*performance*).

Ejemplo:

En un ambiente donde la meta es el consumo de recursos, las variables pueden incluir disponibilidad de recursos, igualdad en el acceso a los recursos, entre otras.



[16] E. Serfilippi and G. Ramnath, *Resilience measurement and conceptual frameworks: A review of the literature*, Annals of Public and Cooperative Economics, vol. 89, no. 4, pp. 645–664, 2018.

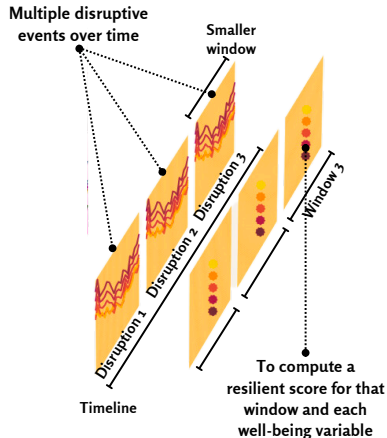
[17] G. P. Cimellaro, C. Renschler, A. M. Reinhorn, and L. Arendt, *Peoples: A framework for evaluating resilience*, Journal of Structural Engineering, vol. 142, no. 10, p. 04016063, 2016.

Metodología: Etapa II

- Los sistemas no enfrentan solo una disrupción, pueden tener varias ¹⁸.
- Segmentamos la línea temporal en ventanas aislando el efecto de cada disrupción.
- En cada ventana comparamos el comportamiento de las curvas de desempeño.
- Con la comparación calculamos, **un score de resiliencia** ¹⁹.

[18] G. E. Richardson, *The metatheory of resilience and resiliency*, Journal of Clinical Psychology, vol. 58, no. 3, pp. 307–321, 2002.

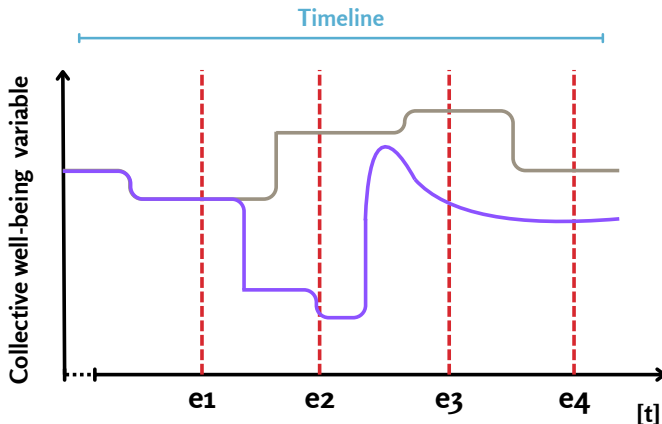
[19] Ayyub, B. M. (2014). *Systems resilience for multihazard environments: Definition, metrics, and valuation for decision making*. Risk Analysis, 34(2), 340–355. Wiley Online Library.



Metodología: Etapa II

Example:

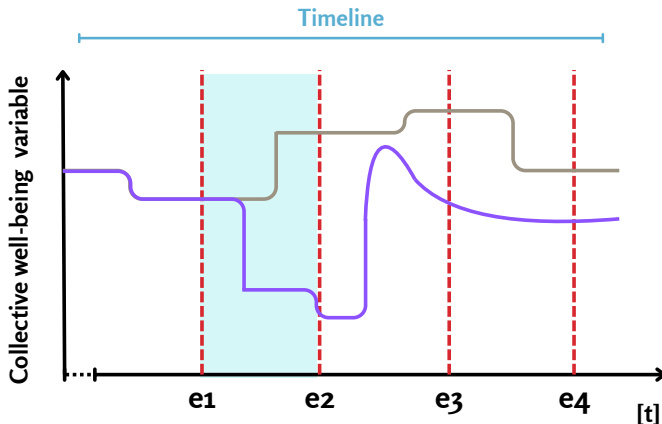
- Reference curve
- Performance curve
- - Disruptive events



Metodología: Etapa II

Ejemplo:

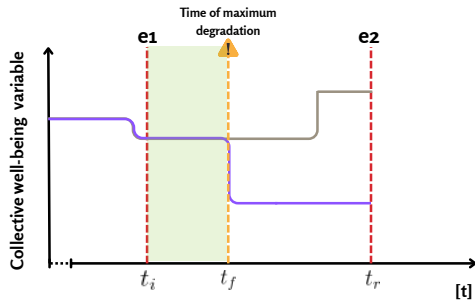
- Referencia
- Desempeño
- - Evento disruptivo
- Ventana para e_1



Metodología: Etapa II

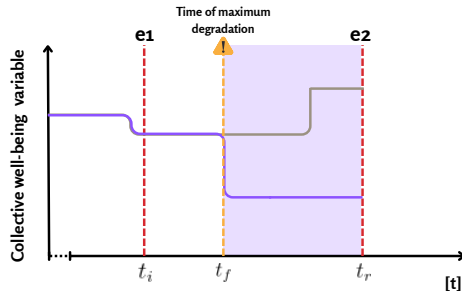
Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t) dt \left(\int_{t_i}^{t_f} R_{jl}(t) dt \right)^{-1}$$



Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t) dt \left(\int_{t_f}^{t_r} R_{jl}(t) dt \right)^{-1}$$



Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t)dt \left(\int_{t_i}^{t_f} R_{jl}(t)dt \right)^{-1}$$

Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t)dt \left(\int_{t_f}^{t_r} R_{jl}(t)dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl}\Delta t_f + G_{jl}\Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t) dt \left(\int_{t_i}^{t_f} R_{jl}(t) dt \right)^{-1}$$

Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t) dt \left(\int_{t_f}^{t_r} R_{jl}(t) dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl} \Delta t_f + G_{jl} \Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t)dt \left(\int_{t_i}^{t_f} R_{jl}(t)dt \right)^{-1}$$

Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t)dt \left(\int_{t_f}^{t_r} R_{jl}(t)dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl}\Delta t_f + G_{jl}\Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t)dt \left(\int_{t_i}^{t_f} R_{jl}(t)dt \right)^{-1}$$

Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t)dt \left(\int_{t_f}^{t_r} R_{jl}(t)dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl}\Delta t_f + G_{jl}\Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

- Δt_f duración de la falla, desde el inicio de la disrupción hasta el punto de mayor degradación.
- Δt_r duración de recuperación, desde el punto más bajo hasta el inicio de la siguiente disrupción.

Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t)dt \left(\int_{t_i}^{t_f} R_{jl}(t)dt \right)^{-1}$$

Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t)dt \left(\int_{t_f}^{t_r} R_{jl}(t)dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl}\Delta t_f + G_{jl}\Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

El **numerador** captura qué tan bien resistió el sistema durante la falla y cuánto se recuperó.

Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t) dt \left(\int_{t_i}^{t_f} R_{jl}(t) dt \right)^{-1}$$

Perfil de recuperación:

$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t) dt \left(\int_{t_f}^{t_r} R_{jl}(t) dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl}\Delta t_f + G_{jl}\Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

El **denominador** es solo el intervalo de tiempo total que estamos considerando.

Metodología: Etapa II

Perfil de falla:

$$F_{jl} = \int_{t_i}^{t_f} P_{jl}(t)dt \left(\int_{t_i}^{t_f} R_{jl}(t)dt \right)^{-1}$$

Perfil de recuperación:

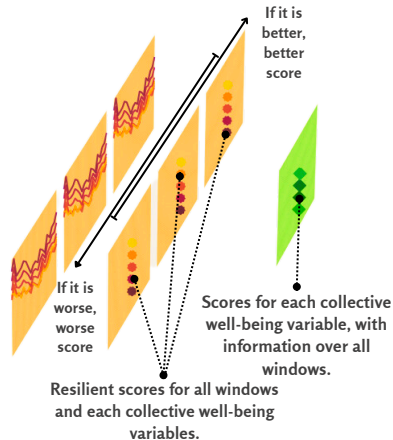
$$G_{jl} = \int_{t_f}^{t_r} P_{jl}(t)dt \left(\int_{t_f}^{t_r} R_{jl}(t)dt \right)^{-1}$$

Score total:

$$J_{jl} = \frac{t_i + F_{jl}\Delta t_f + G_{jl}\Delta t_r}{t_i + \Delta t_f + \Delta t_r}$$

Metodología: Etapa III

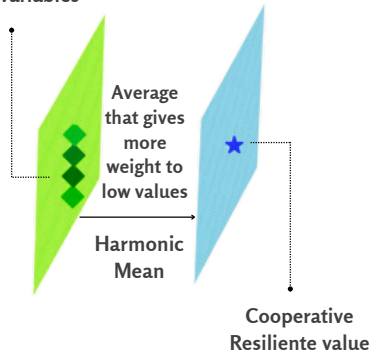
- Hemos calculado el *score* de resiliencia para múltiples ventanas y la variable de bienestar colectivo.
- Para cada variable, **consolidamos la puntuación resiliente en ventanas de tiempo**.
- Utilizamos un promedio ponderado para tener en cuenta la tendencia del sistema.
 - Si el sistema empeora, penalizamos la puntuación.
 - Si el sistema mejora, premiamos la puntuación.



Metodología: Etapa IV

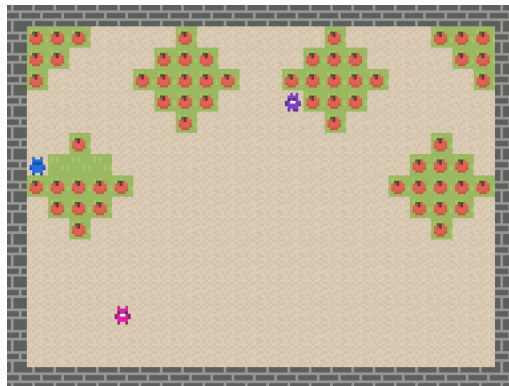
- **Reunimos las puntuaciones de resiliencia de todas las variables en una única métrica global.**
- Para combinar todas las puntuaciones de las variables en un solo valor, utilizamos la media armónica.
- De esta manera, un desempeño deficiente en cualquier variable reduce la puntuación general.

Consolidated resilience
for collective well-
being variables



Medir resiliencia cooperativa

- **Enfoques de toma de decisiones**
 - PPO (Reinforcement Learning), 22²⁰
 - LLM-Based Agent (Text-to-Action)²¹
- **Meta:** Medir la resiliencia cooperativa bajo diferentes enfoques de agentes artificiales.



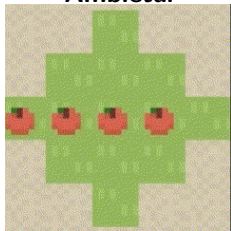
[20] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv preprint arXiv:1707.06347.

[21] Mosquera, M., Pinzón, J. S., Ríos, M., Fonseca, Y., Giraldo, L. F., Quijano, N., and Manrique, R. (2024). *Can LLM-augmented autonomous agents cooperate? An evaluation of their cooperative capabilities through Melting Pot*. arXiv preprint arXiv:2403.11381.

[22] J. Perolat, J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel, *A multi-agent reinforcement learning model of common-pool resource appropriation*, Advances in Neural Information Processing Systems, vol. 30, 2017.

Medir resiliencia cooperativa

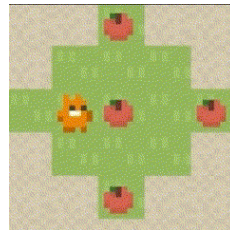
Disrupción tipo 1: Ambietal



Riesgo: Número de eventos disruptivos (1, 2, 3).

Vulnerabilidad: Número de manzanas eliminadas (bajo, medio, alto).

Disrupción tipo 2: Social

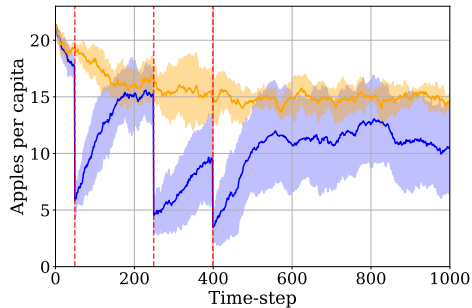


Riesgo: Fijo (1 inyección del bot)

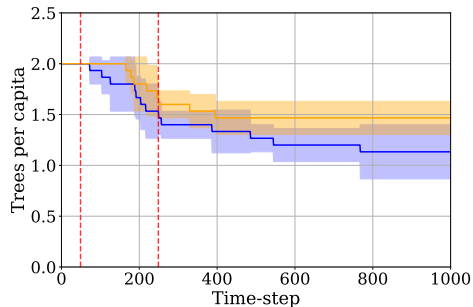
Vulnerabilidad: Duración de la actividad del bot (corta, media, larga)

Indicadores de bienestar conjunto

Manzanas disponibles *per cápita*



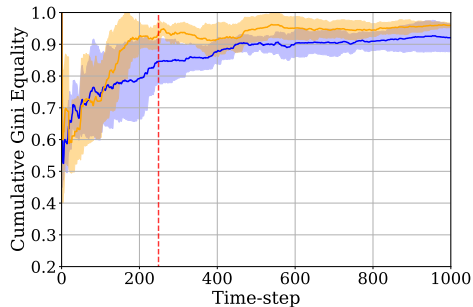
Árboles disponibles *per cápita*



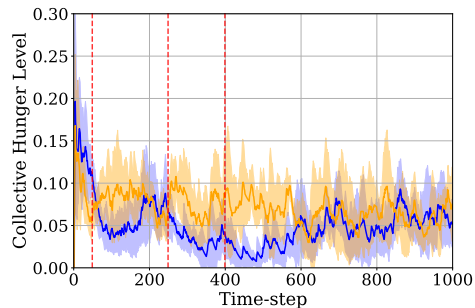
Convenciones: ■ Referencia ■ Disempeño - - - Eventos disruptivos

Indicadores de bienestar conjunto

Indice GINI de igualdad



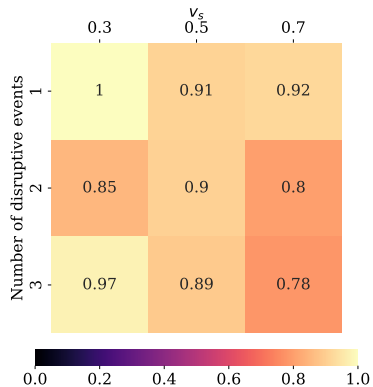
Indice colectivo de saciedad



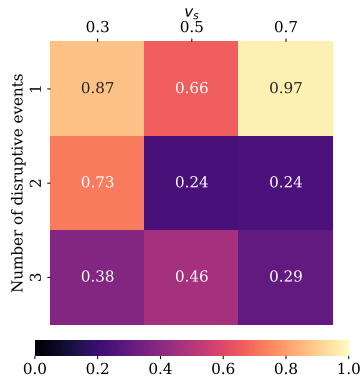
Convenciones: ■ Referencia ■ Disempeño - - - Eventos disruptivos

Mapas de resiliencia cooperativa: Tipo 1

Agentes PPO

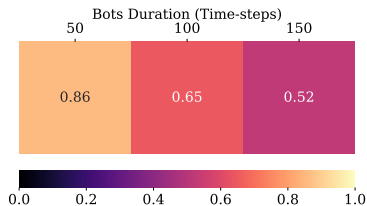


Agentes LLM

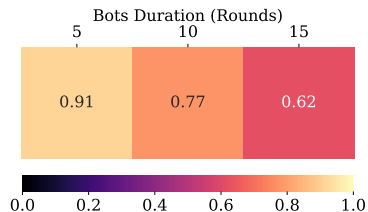


Mapas de resiliencia cooperativa: Tipo 2

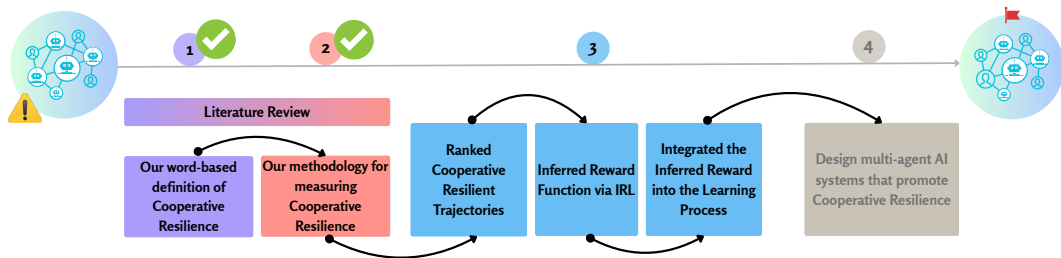
Agentes PPO



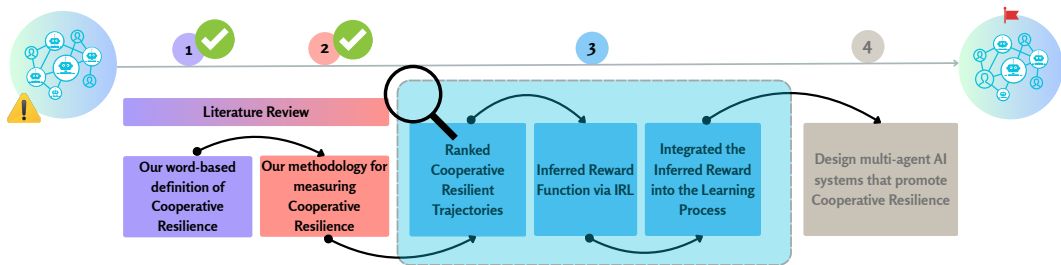
Agentes LLM



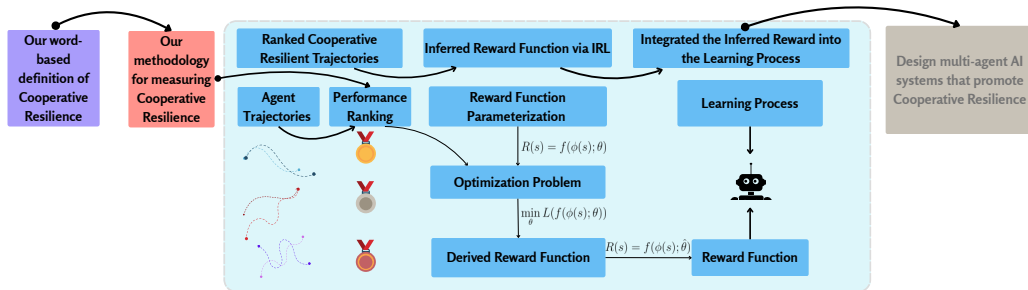
¿Dónde estamos?



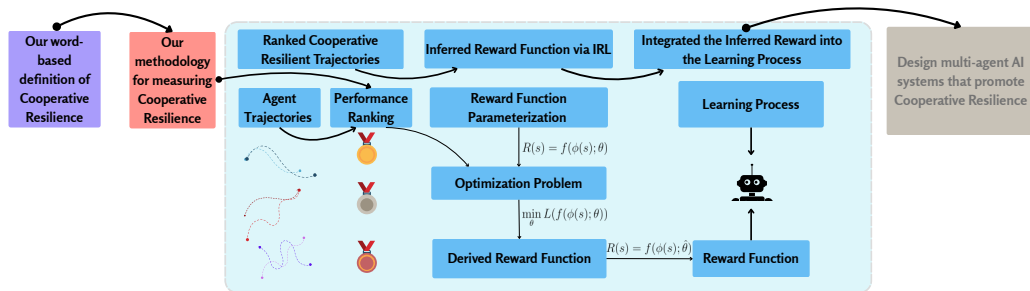
¿Dónde estamos?



Aprendiendo de trayectorias resilientes



Aprendiendo de trayectorias resilientes



¿Qué función de recompensa debieron haber optimizado los agentes para comportarse de esta manera?

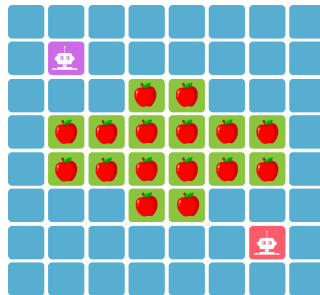
[23] Brown, D., Goo, W., Nagarajan, P., and Niekum, S. (2019). *Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations*. In Proceedings of the International Conference on Machine Learning (ICML), pp. 783–792. PMLR.

[24] D. Goktas, A. Greenwald, S. Zhao, A. Koppel, and S. Ganesh, *Efficient inverse multiagent learning*, arXiv preprint arXiv:2502.14160, 2025.

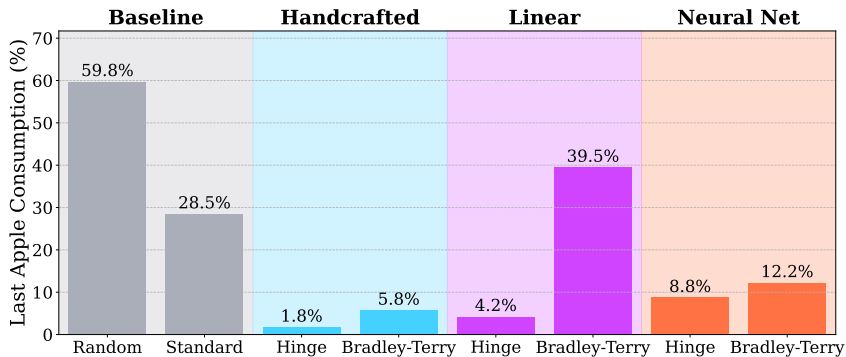
[25] J. Wu, W. Shen, F. Fang, and H. Xu, *Inverse game theory for Stackelberg games: The blessing of bounded rationality*, Advances in Neural Information Processing Systems, vol. 35, pp. 32186–32198, 2022.

Mejorar la resiliencia cooperativa mediante trayectorias jerarquizadas

- Nos centramos en una versión simplificada del dilema social:
- Extraemos trayectorias de este entorno y las clasificamos utilizando la métrica de resiliencia cooperativa.
- Usamos esas clasificaciones para inferir una nueva función de recompensa.

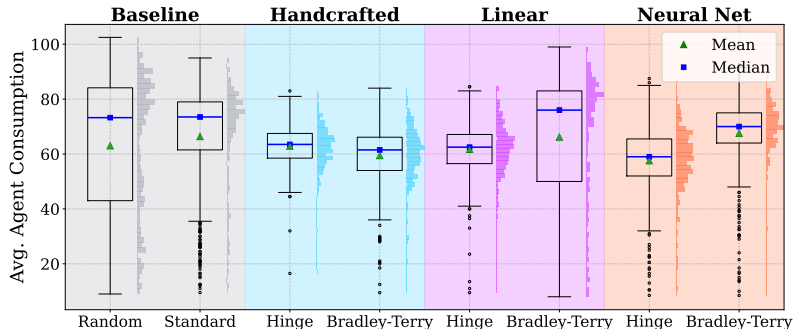


Hallazgos clave: Episodios con el último consumo de manzana



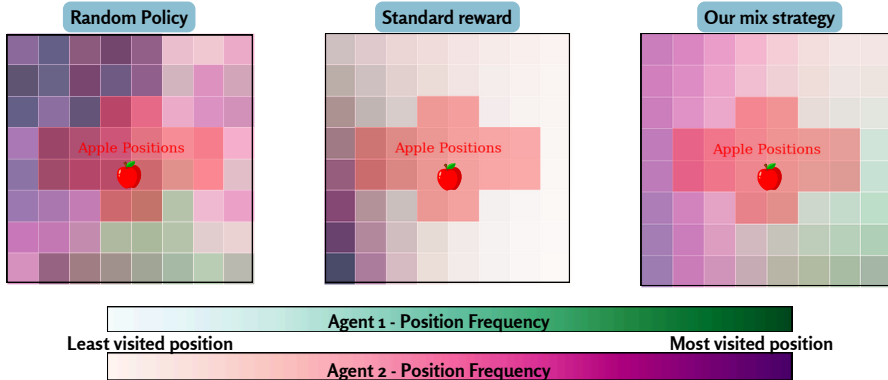
El porcentaje de simulaciones en las que se consumió la última manzana. Un porcentaje alto indica un comportamiento más egoísta. La mayoría de nuestros métodos redujeron significativamente este comportamiento.

Hallazgos clave: Consumo total promedio de manzanas



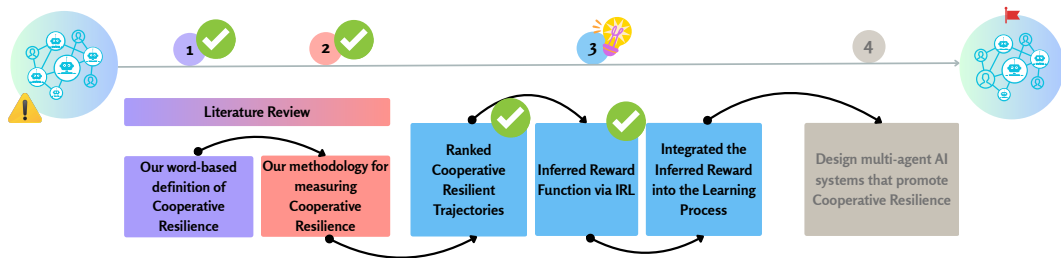
Observamos que el consumo promedio de manzanas de ambos agentes se mantiene estable en todos nuestros métodos. Si bien nuestros modelos redujeron el comportamiento egoísta, el consumo total se mantuvo estable, lo que significa que la cooperación mejoró sin perder eficiencia.

Resumen de los hallazgos clave: mapas de frecuencia de posición



Al utilizar nuestra estrategia, los agentes se distribuyen más eficientemente en el entorno y cada uno se ubica en un rincón, a diferencia de la mala distribución con la recompensa estándar o la ubicación aleatoria observada en la política aleatoria.

¿Dónde estamos?



1

¿Qué queremos resolver?

2

¿En que parte estamos ahora?

3

¿Qué hemos logrado?

4

¿Qué tal la experiencia?

¿Qué hemos logrado?

Resultados

- **Artículo en Journal IEEE Transactions on Artificial Intelligence:** Título “*Cooperative Resilience in Artificial Intelligence Multiagent Systems*”.



Resultados

- Participación en el Concurso Concordia y NeurIPS 2025 (Datasets and Benchmarks): Contribuimos al artículo “*Evaluating generalization capabilities of LLM-based agents in mixed-motive scenarios using Concordia*” , aceptado en NeurIPS 2025, una de las conferencias más importantes y prestigiosas del mundo en inteligencia artificial. Hasta donde sabemos, somos el único equipo latinoamericano dentro de este proyecto global, compartiendo autoría con investigadores líderes en IA.



¿Qué tal la experiencia?

El inicio ...



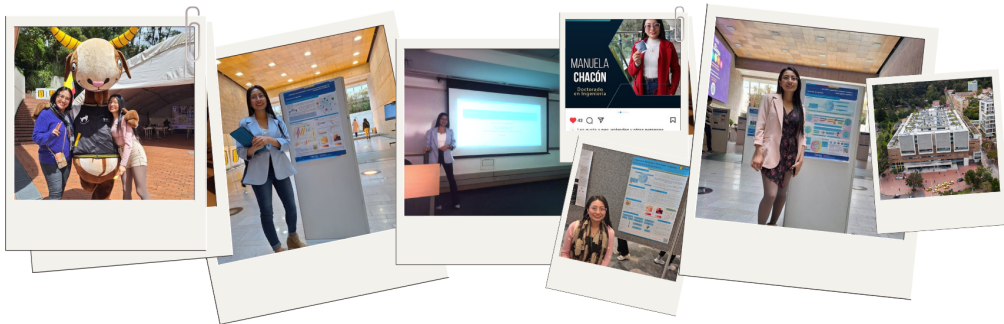
El inicio ...



El inicio ...

- **Financiación:** Pregunta todo desde el inicio; entiende bien becas, tiempos y obligaciones.
- **Asesores:** Compatibilidad de trabajo y comunicación, no solo renombre.
- **Cursos:** buscar cursos que alimenten tu tema, no que te desgasten.
- **Tema:** Eligir algo que te apasione y sea viable en tiempo y recursos.
- **Exámenes y plan de trabajo:** Conocer los hitos clave del programa y tener un plan para lograr estos hitos.

Lo bonito ... pequeñas (grandes) victorias.



Lo bonito ... pequeñas (grandes) victorias.



Lo bonito ... pequeñas (grandes) victorias.

- Oportunidades académicas: posters, presentaciones y conferencias.
- Experiencias en la universidad: campus y laboratorios.
- *Summer schools* y estancias de investigación.
- Viajes y nuevas experiencias.
- Disfrutar de la agencia del tiempo.

Lo no tan bonito ...

- Soledad académica y sensación de no avanzar.
- Rechazos, revisiones duras y frustración constante.
- Incertidumbre: ¿voy bien?, ¿es suficiente?, ¿cuándo termina?
- Carga emocional por la autoexigencia y la comparación.
- Experimentos fallidos y semanas completas sin resultados.
- Dificultad para equilibrar vida personal y trabajo.

Lo que me ha funcionado ...

- Recordar que el PhD es solo una parte de la vida, no toda mi identidad.
- Tomarme las cosas con más calma: nada es tan urgente como parece.
- Mantener hobbies y actividades fuera de la academia.
- Buscar ambientes que te hagan feliz.

Agradecimientos

A la Facultad de Ciencias Exactas y Naturales de la Universidad Nacional de Colombia, Sede Manizales, por la invitación al Workshop en Ciencias de la Computación; en particular, al profesor Fabián Fernando Serrano por la gestión de mi participación y, de manera muy especial, al profesor Juan Carlos Riaño por la invitación. Agradezco asimismo a mis asesores de doctorado, Luis Felipe Giraldo y Nicanor Quijano, por su guía y apoyo continuo.

Gracias

¿Preguntas?