Supplementary File of Cooperative Resilience in Artificial Intelligence Multiagent Systems

Manuela Chacon-Chamorro Luis Felipe Giraldo Nicanor Quijano Vicente Vargas-Panesso
César González Juan Sebastián Pinzón Rubén Manrrique Manuel Ríos Yesid Fonseca
Daniel Gómez-Barrera Mónica Perdomo-Pérez *

September 19, 2024

This supplementary file provides an extended and detailed exploration of some aspects that complement the main paper. While the core content of the paper is fully understandable on its own, this supplementary material offers complementary information into the following items:

- **A review of resilience definitions**: A detailed review is presented, exploring definitions from various fields and the key aspects that contributed to the construction of the framework.
- **LLM architecture Description**: A explanation of the LLM-based architecture of [S1], including the adapter that transforms spatial observations into textual inputs for GPT-4. This section also discusses the integration of the language model into the decision-making process of the agents.
- **Well-being variables and environment description**: Detailed descriptions of the environment 'Commons Harvest Open,' and the well-being variables used in our study, along with the specific metrics developed.

---

1 A review of resilience definitions

In this section, a detailed review is presented, exploring definitions from various fields and the significant aspects that contributed to constructing Figure S1 and the definition of cooperative resilience. In ecology resilience is focused on the system's capacity to absorb changes and disturbances [S2]. The term also encompasses elements such as *resistance*, which signifies the ability to withstand changes following a perturbation, and recovery, denoting the ability to fully return to the reference state. Additionally, it includes *latitude*, representing the maximum extent to which a system variable can be altered before losing its capacity to

*This work was supported by Google through the Google Research Scholar program and the UniAndes-DeepMind Scholarship 2023.
M. Chacon-Chamorro (m.chaconc), L.F. Giraldo (lf.giraldo404), N. Quijano (nquijano), V. Vargas-Panesso (jv.vargas), C. González (cl.gonzalezg), J.S. Pinzón (js.pinzonr), R. Manrrique (rf.manrique), Y. Fonseca (y.fonseca) and D. Gómez-Barrera (df.gomezb) are with the Universidad de los Andes, Colombia [@uniandes.edu.co].
M. Ríos (manrios) is with Center of Excellence in Analytics and Artificial Intelligence Bancolombia [@bancolombia.com.co].
M. Perdomo-Pérez (tatiana.perdomo) is with the Universidad de Ibagué, Colombia [@unibague.edu.co].

recover [S3]. Resistance is associated with the stability of the ecological system, whereas latitude is linked to the boundaries of the disturbance.

The disruptive event in ecology is associated with disruptions in population dynamics [S2], which manifest as disturbances to the system. These disturbances can result from various factors, such as random events linked to changes in an ecosystem's climatic conditions. Events of this nature can also be man-made disturbances, resembling to perturbed systems where the parameter values or the constituent population levels are altered [S2]. Perturbations can also be viewed as biotic or abiotic forces, agents, or processes that induce alterations in system variables and/or parameters [S3]. Within this context, resilience is examined from the perspective of population dynamics, and the analysis of this system is approached through the dynamical systems theory.

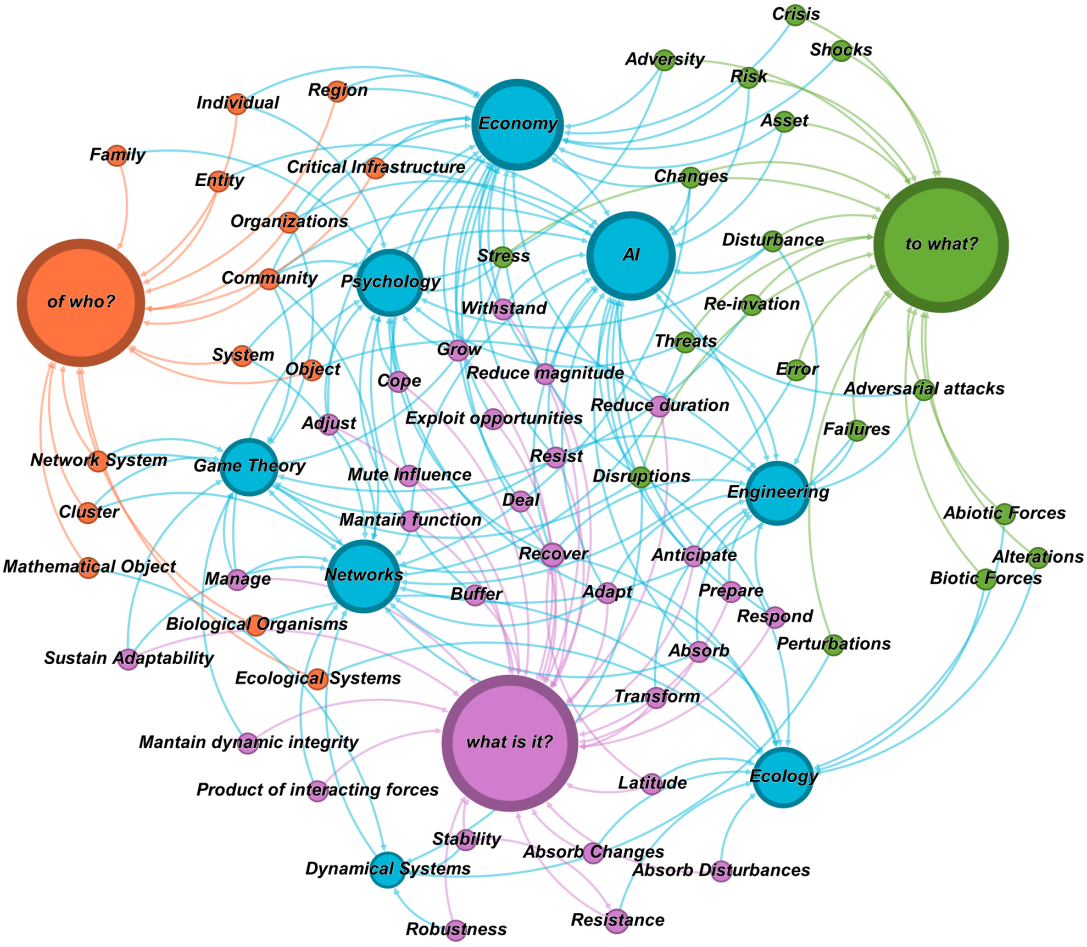


Figure S1: Keyword map of resilience across diverse fields and contexts, addressing guiding questions.

In this field of study, resilient entities are the ecological systems [S3], which comprise ecosystems consisting of various biological organisms that interact within a specific environment. For instance, one extensively studied ecological system is the predator-prey system [S2, S4]. Within this context, resilience is examined from the perspective of population dynamics, and the analysis of this system is approached through the dynamical systems theory.

In fact, the conceptual theory of dynamical systems also developed a idea of resilience. In this context, resilience is strongly related to the disturbances that a dynamic system may encounter. These disturbances can take various forms, including external factors, perturbations in the system's initial conditions, or those associated with changes in parameters [S5], which could potentially lead to bifurcations. These disturbances may exhibit time-dependent behavior and manifest either randomly or deterministically.

In that way, the resilience is associated with some properties of dynamic systems, such as various forms of stability and robustness. However, there is no unified mathematical formalization of the concept of resilience. While some authors have proposed resilience definitions for specific dynamic systems, a consensus on a general definition remains indefinite. For instance, [S6] introduces the concept of "marginal resilience," developed to quantify the robustness of a particular flow network system. The proposed definition of marginal resilience is characterized as "the smallest magnitude of perturbation that drives the dynamical flow network to instability." It is important to note that this definition is contingent upon a series of assumptions and previous definitions, limiting its applicability to this specific context.

In this area, the focus is on a particular phenomenon that describes how certain variables, that change over time, exhibits the property of resilience. For this reason, the concept of resilience may emerge depending on the context of the study related to the phenomenon. For example, in [S6], the concept of "margin of resilience" measures robustness within of the models the flow of a network with specific characteristics. In this study, resilience lies within the phenomenon associated with the flow network.

In the case of psychology, the aim is studying the resilience focus in people individually or collectively. In this field, resilience is a scale that spans from individuals to groups such as families and communities [S7–S12]. When the focus is on the individual, resilience is attributed to the person and is analyzed from two distinct perspectives. The first perspective posits that this capability inherently resides within individuals as a biological property. The second one, consider that resilience is related with the context, interactions with the environment, and external factors that protect the individual. Besides, a third perspective merges these two factors, proposing that psychological events depend not only on the individual's state and biological makeup but also on the surrounding environment [S11, S12].

In the field of psychology the words used to describe the possible adversities that affect people or groups of people are stress, threats, and disturbances that are embedded in life events [S7–S10]. These life events can be potentially traumatic events like violence, life-threatens, death, serious injury, sexual violence, loss, situations that evoke fears, among other negative feelings [S13, S14]. Some authors argue that psychologically impacting disruptive events should be characterized by specific statistical magnitudes [S15]. In other words, there are thresholds beyond which it can be determined whether or not there is an impact on an individual's mental health. Conversely, other authors posit that, within the psychological context, any event of any magnitude that induces suffering in an individual is considered a disruption [S12, S16]. In general, the disruptive effect can be analyzed as a significant exposure to an event that poses a risk to an individual's psychological well-being [S12, S15].

Another field of study that has explored the concept of resilience is economics. In economic systems, disruptive events are more commonly known as shocks. They represent the set of unexpected conditions that could affect negatively the behavior of an economic system, for example the natural disasters [S17]. In this field the resilience is related to the economic relationships like individuals, regions or organizations [S17], in groups of individuals as communities [S18] or in economic assets as critical infrastructure [S19] and the capacity to tolerate the unexpected shocks.

Like economics, another area that explores interactions between agents is game theory. In this domain, the set of interacting agents can be resilient to re-invasion of other agents that do not follow the desired equilibrium [S20], to stress or disruptions on the mechanisms controlling the behavior of them, as changes on the payoff function [S21]. The definition of resilience in game theory is the ability to sustain adaptability, maintain dynamic integrity, manage, resist and recover [S20, S21]. For example, in the Prisoner's Dilemma, resilience is the property of the cooperator cluster [S20], also called community, or the ability of a network system [S22] to maintain the level of cooperation.

Other field that deals with an interacting set of agents is network science. In network science, resilience is studied through the analysis of systems described as networks or clusters [S23–S25]. This concept is defined in three main categories. The first one describes the ability of a network to maintain adaptability, transform, adjust and adapt, referring to the fact that in order to deal with failures, errors, threats and changes, the network must change its internal structure and adapt to the adversity it faces. The second category describes

the ability to manage, maintain function and absorb, meaning that the network should be able to withstand the adversity without affecting its performance. Finally, network science views resilience as the product of interacting forces on an environment, so each component cannot be studied in isolation from the rest of the network [S23–S25].

In engineering, the definition of resilience is associated with activities before, during, and after adversity [S26]. Before the adversity, resilience is the ability to prepare and anticipate; during the adversity, the abilities to absorb, adapt, transform, and resist take place; and after the adversity, the abilities to recover, respond, adapt, and transform are used. As described, the activities of transforming and adapting occur during and after the adversity because these abilities help the system to resist and to recover.

The focus of resilience in engineering is on physical systems or objects, such as critical infrastructure [S26]. For instance, in control systems, resilience may be related to the system's tolerance to disturbances. Specifically, these concepts are studied in Fault Tolerant Control Systems (FTCS), where the objective is to overcome such vulnerabilities, allowing the system to tolerate component malfunctions while preserving desirable stability and performance characteristics [S27]. In such cases, resilience is attributed to the system and its constituent components, which can be physical or virtual. A special case is Cyber Physical Resilient Systems, where the system adapts in real-time to known and unknown threats and adversities, strategically responding to maintain critical functions and performance even in the event of successful attacks [S28].

Specifically, within the realm of FTCS, disruptive events can take the form of presumed faults or entirely uncertain faults for which the shape, timing, and magnitude of impact are unknown. Additionally, adversarial scenarios must be considered, where these events are uncertain from the system's perspective but intentional from the attacker's viewpoint [S27]. For example in AI systems, the disruptive events can be adversarial attacks, changes, and disruptions [S29, S30]. In this particular field the definition of resilience is similar to the other. The most common verbs used to describe resilience are some of the previously presented, such as to absorb, respond, adapt, prepare, anticipate, maintain function, resist and recover. In this way, a resilient system is one that can minimize the impact of adversity and shorten the recovery time. As described above, the ability to resist could be associated with the desire to minimize the impact of adversity, while the ability to respond and adapt could be associated with reducing the time of recovery.

The various definitions of resilience demonstrate its universality across multiple fields. It can be thought that the essence of defining resilience is encapsulated in identifying the resilient entity (of who?), the verbs that define resilience actions (what is it?), and recognizing the disruptive event (to what?). These key questions, along with their corresponding keywords, are illustrated in Figure S1, highlighting the interdisciplinary nature of the concept.

2 LLM Architecture

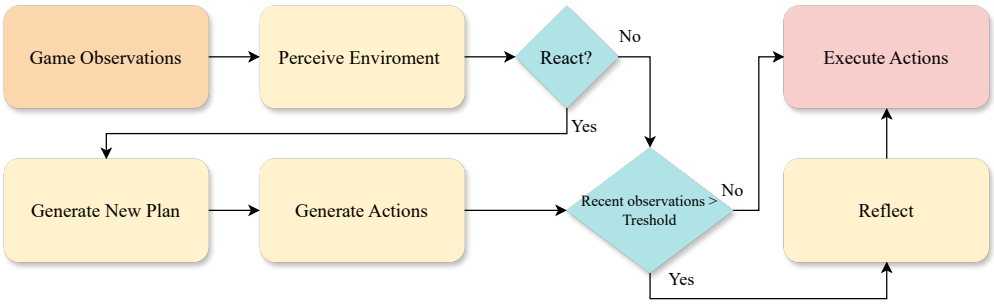


Figure S2: Diagram summarizing the reasoning process flow within the LLM architecture, leading to the action-taking phase of each agent. The diagram is inspired by the architecture proposed in [S1].

The architecture of the "Generative Agents" proposed in [S1] is summarized in Figure S2. The modules of this architecture are described as follows: the **memory module** primarily consists of two aspects: memory itself and the selective retrieval of these memories. Short-term memory stores all the information the agent needs to instantly access. Long-term memory stores all the experiences the agent has lived through, including

observations acquired and stored with associated relevance and recency information, along with their representation in an embedded space using Sentence Transformer encoders. Also, there is a spatial memory that enables the agent to maintain all the spatial information of the environment. Regarding memory retrieval, agents possess the capability to access stored memories based on criteria such as recent memories, relevance, and similarity to the input query. To achieve this, the agent accesses its memories by weighting these factors. This module is crucial for effectively applying past knowledge that is relevant to current situations.

The purpose of the **perception module** is to enable the agent to form an initial understanding of the observations it perceives from the world. By providing information about the current plan and textual observations of the current environment through a prompt, the language model assists the agent in determining whether it should react to its surroundings. Reacting triggers the flow of subsequent modules, whereas not reacting entails sticking to the previous plan and sequence of actions for that time step. This perceptual ability is crucial for adaptation and effective response to changes in the environment, building upon the information stored and retrieved in the memory module.

When initiating a plan, the **planning module** comes into play. This module guides agents to formulate plans and set goals, utilizing recent observations, the ongoing plan, their understanding of the world, and reflections obtained from memory. The language model then generates a plan aligned with both the agent's understanding of the world and the highlighted information. Consequently, an agent's current plan is crucial for guiding decision-making across other reasoning modules.

The module responsible for deepening the understanding of the current world is the **reflection module**. This process begins when the agent reaches a predefined reflection threshold, based on the quantity and diversity of observations accumulated over different time steps, and then generates relevant insights. The module operates through two prompts for the language model. In the first prompt, the model is asked to generate three questions that can be answered with the accumulated observations, and then extract related memories for each of the generated questions. In the second prompt, the model is requested to use this group of relevant memories to generate three insights, one per group of memories, that highlight the most important aspects inferred from those memories.

With a defined plan, the **action module** guides agents to generate a series of actions and reactions to the environment. The language model is provided with knowledge of the world, the current plan, present observations, current objectives, and recent reflections to generate specific actions aligned with all these inputs. During this process, a set of valid actions that the agent can perform in the given environment is specified to the language model. In the case of "Commons Harvest Open," these include exploration, movement to a specific position, apple consumption, and attacks on other players. For high-level actions, a sequence of steps that the agent must follow to execute the action is detailed. It is important to note that if the plan changes, the actions will also adjust accordingly, providing flexibility and key adaptability in the agents' performance within the simulated social environment.

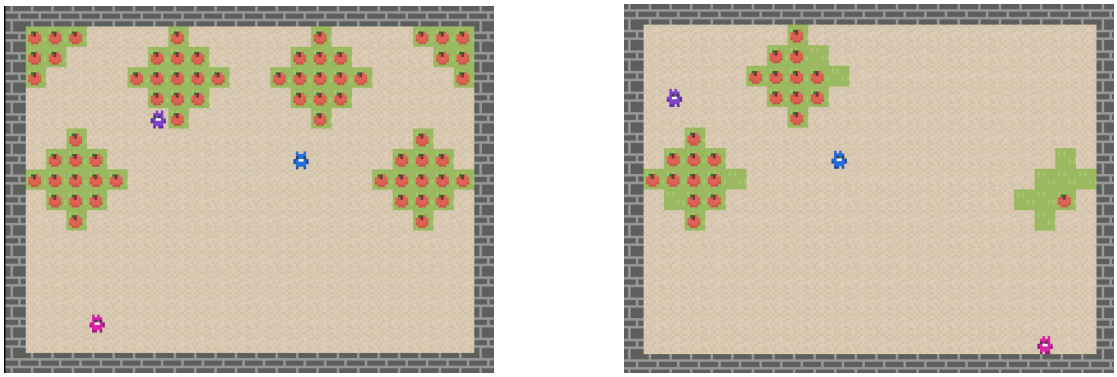
3 Well-being variables and environment description

This section describes the environment used in the experiments and the well-being variables employed to measure cooperative resilience. For the well-being variables, the figures in this section present the mean values across 5 episodes, with blue representing performance and orange indicating the reference. Shaded areas illustrate the standard deviations. Red dashed lines mark the onset of disruptive events. In the case of the second event, the appearance of bots is indicated by a solid red line, and their departure by a green line.

3.1 Environment description

The specific scenario chosen is referred to as "Commons Harvest Open," from the Melting Pot 2.0 suite [S31]. In this scenario, multiple agents inhabit a confined space containing apple-laden trees. Figure S3 illustrates the described scenario in some moments to simulation. The simulation begins with 4 complete trees containing 13 apples and 2 incomplete trees in the upper corners, each holding 6 apples, as shown in Figure S3a. We considered a setup involving 3 agents, each tasked with consuming as many apples as possible across the map, for example the Figure S3b show when the agents have already consumed some apples. Consumed apples regenerate with a probability per step that depends on the number of remaining

apples on the tree. If all the apples on a tree are consumed, the tree vanishes. For this reason the scenario shows a social dilemma, when all apples are depleted from a tree, no further apples with grow and this goes in detriment of the entire population. So, to preserve the apples, the agents must be have a social understand of their actions. The agents can move up, down, left, and right, and they observe the environment through a partial view limited to a specific section of the space. In their field of vision, agents can see other agents, trees, and walls. What the agent can observe depends on its position within the environment.



(a) Start of the simulation.

(b) Simulation in progress.

Figure S3: Diagram of the "Commons Harvest Open" environment. In the figure (a) the trees contain all the apples, in (b) the agents have already consumed some apples.

3.2 Apples alive *per capita*

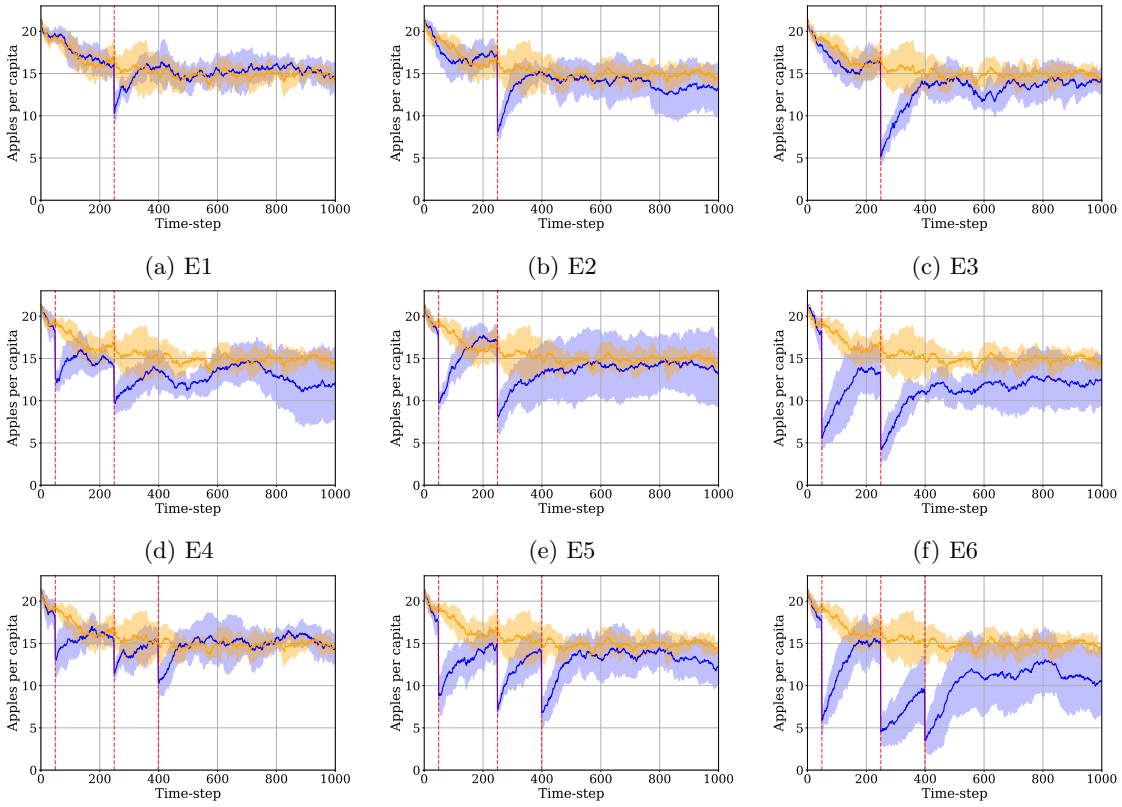
This variable evaluates the number of living apples in the environment at each step. This measure is directly related to the availability of resources for the community, thus reflecting the well-being of the agents. Therefore, a higher number of apples *per capita* indicates better future access to resources for the agents. Figure S4 shows the performance and reference curves for this variable across the nine experiments using the RL technique, while Figure S6 illustrates the same for LLM techniques, both pertaining to the first disruptive event. The variable for the second disruptive event is depicted for RL in Figure S5 and for LLM in Figure S7.

3.3 Trees alive *per capita*

Similar to apples alive *per capita* this variable evaluates the number of living tree in the environment at each step. This measure is also directly related to the availability of resources for the community. Therefore, a higher number of tree *per capita* indicates better future access to resources for the agents. Additionally, the trees also highlight an important aspect related to the recovery systems. When a tree disappears, it never reappears, resulting in a significant decrease in resource (apples) availability. Figure S8 displays the performance and reference curves for this variable across the experiments for RL technique, while Figure S10 shows the same for LLM, both for the first disruptive event. This indicator for the second disruptive with RL technique is show in Figure S9 and for LLM in Figure S11.

3.4 Cumulative Gini Equality Index

This variable, adapted from the Gini index, a standard tool in economics for assessing resource inequality within populations, measures the fairness of reward allocation. It has been employed in multi-agent system research to analyze equitability in multi-agent system experiments involving sequential social dilemmas, as highlighted by [S32]. Conversely, [S33] applied the inverse Gini index as a learning reward in RL training, in scenarios related with Melting Pot framework.



(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

(g) E7 (h) E8 (i) E9

Figure S4: Performance and reference curves of apples alive *per capita* using RL for agent's decision-making and first disruptive event.

The equation for calculating this indicator is as follows:

$$E[t] = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{i} \sum_{\tau=0}^{t} r_j[\tau]}{r_T \sum_{i=1}^{N} i}.$$

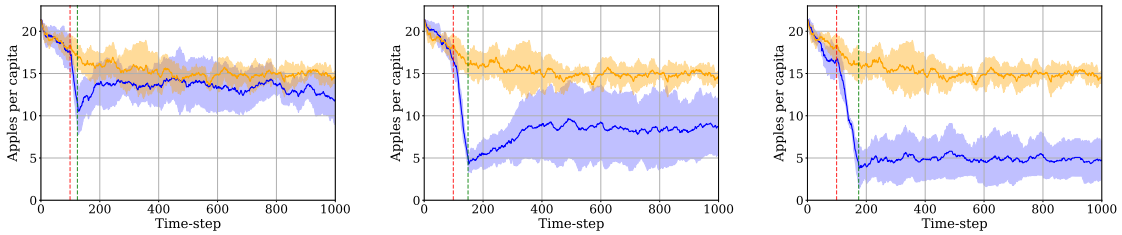
Where, τ is discrete time interval, r_T represents the total rewards consumed by the agents, and $r_j[\tau]$ denotes the reward acquired by agent j at time τ. The term $\sum_{\tau=0}^{t} r_j[\tau]$ calculates the cumulative reward received by an agent from the beginning of the experiment. To ensure accuracy in applying this equation, agents must be organized by the total rewards they have accumulated. This requirement is enforced by the constraint:

$$\sum_{\tau=0}^{n} r_i[\tau] \leq \sum_{\tau=0}^{n} r_{i+1}[\tau]. \tag{1}$$

This condition guarantees that agents are ranked in ascending order based on the rewards they have gathered. The Cumulative Gini equality index is depicted in Figure S12 for the first disruptive event using RL technique and in Figure S14 for LLM. The curves for this indicator using the second disruptive event are presented in Figure S13 for RL and Figure S15 for LLM.

3.5 Collective Hunger Level Index

An alternative method for assessing fairness and cooperation in multiagent systems is to evaluate the general welfare using a specific set of criteria that define collective well-being. Critical to this assessment are two



(a) E1 (b) E2 (c) E3

Figure S5: Performance and reference curves of apples alive *per capita* using RL for agent's decision-making and first disruptive event.

key factors: the time elapsed since an agent last increased their total reward, as well as the number of apples consumed at each time step. This approach is encapsulated by the "Hunger Level" index, which provides insights into both the timing and quantity of resources distributed among agents.

This variable will range between 0 and 1, with values close to unity representing a scenario without hunger for the agents, while values closer to 0 will indicate that the agents has not eaten apples for a time. A value of exactly 0 means that the agent has not eaten apples from the beginning of the simulation until the moment of measurement. This variable, denoted as $H_i[t]$, is calculated for each agent i in each time-step t and is integrated into a single variable known as the Collective Hunger Level variable denoted $cH(t)$. To calculate the variable at time-step $t = \tau$, the variable $r_i(t)$ is define the reward for the agent i at time t during the simulation. So, at a given time $t = \tau$, this variable is calculated as follows

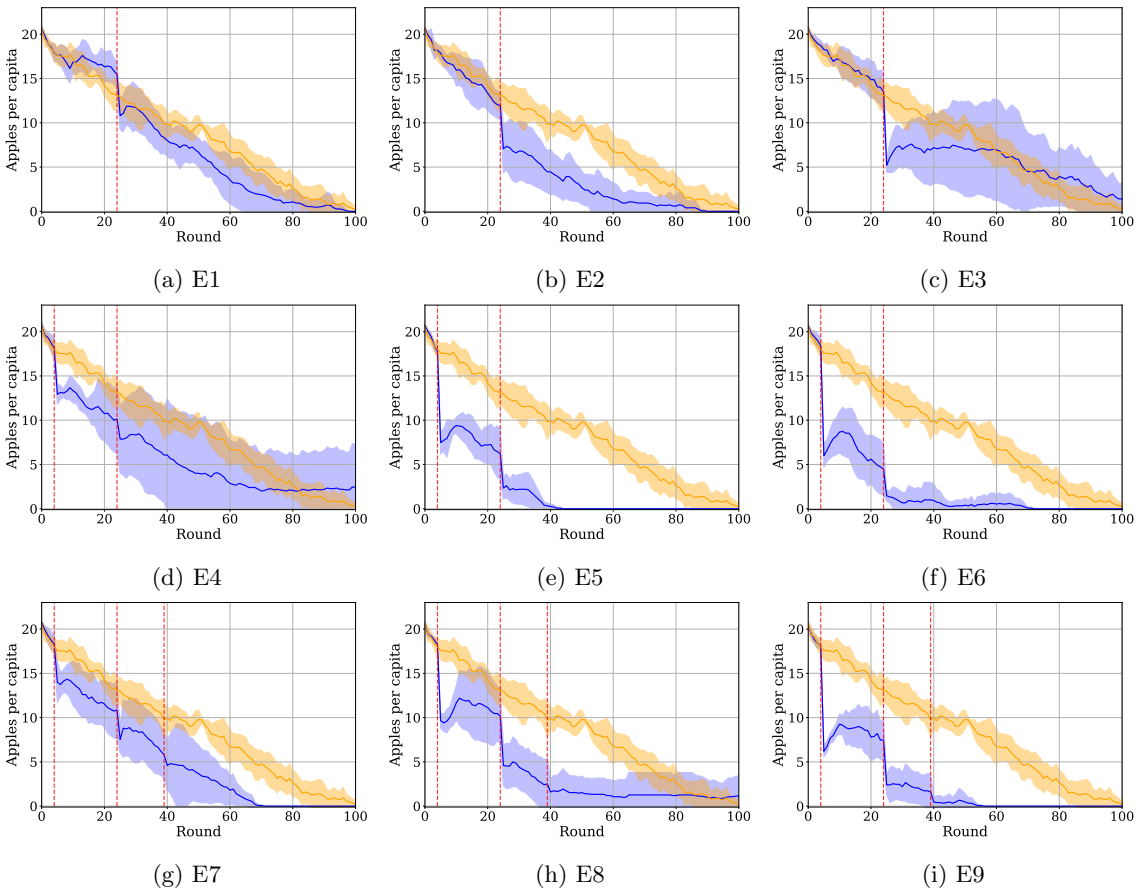
$$H_i[t] = \frac{\sum_{\tau=0}^{t} \gamma^{-(t-\tau)} r_i[\tau]}{\sum_{\tau=0}^{t} \gamma^{-(t-\tau)}}$$

where γ is a discount factor that penalizes longer intervals between rewards, reflecting an exponential decrease in the value of older rewards. The computed variable value ranges from 0 to 1, as the denominator normalizes it in relation to an ideal scenario in which an agent obtains a reward at every time-step, indicating minimal hunger. This normalization is based on the assumption that rewards are binary and evenly distributed over time.

In order to integrate individual measures into a single variable known as the collective hunger level index, denoted as $cH(t)$, the agents' individual measures $H_i(t)$ are connected through the harmonic mean. This approach favors an average when all agents have a similar hunger level at time t, while penalizing deviations when any agent has a significantly lower component. This method allows for a coherent measure of collective well-being aligned with the apple consumption dynamics during simulations. So, in a scenario with N agents the $cH(t)$ is computed like:

$$cH(t) = \frac{N}{\sum_{i=1}^{N} \frac{1}{H_i(t)}}.$$

Figure S16 shows the performance and reference curves for this indicator across the nine experiments conducted with RL technique, while Figure S18 illustrates the same curves for LLM, both in the context of the first disruptive event. For the second disruptive event, Figure S17 presents the curves using RL technique, while Figure S19 shows them for LLM.

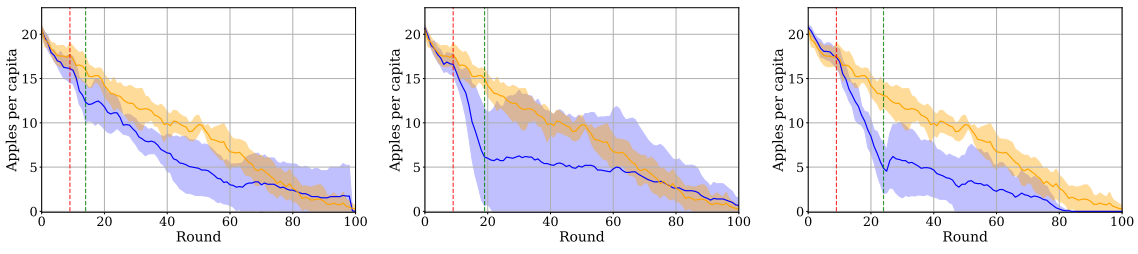


(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

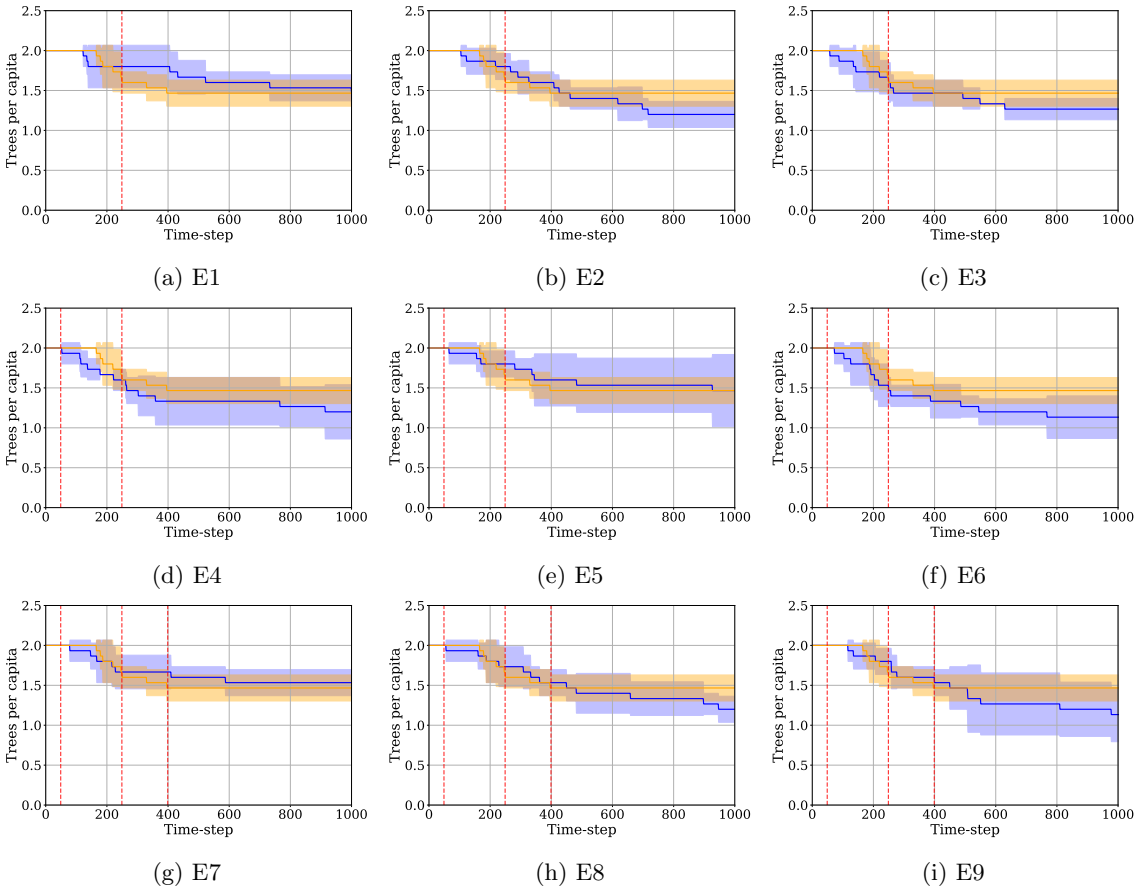
(g) E7 (h) E8 (i) E9

Figure S6: Performance and reference curves of apples alive *per capita* using LLM for agent's decision-making and first disruptive event.



(a) E1 (b) E2 (c) E3

Figure S7: Performance and reference curves of apples alive *per capita* using LLM for agent's decision-making and second disruptive event.

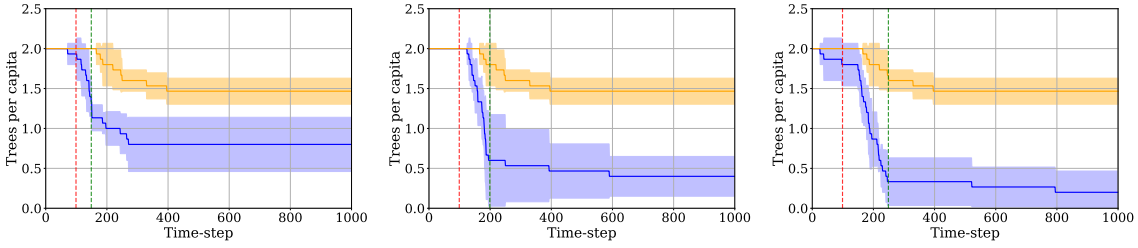


(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

(g) E7 (h) E8 (i) E9

Figure S8: Performance and reference curves of trees alive *per capita* using RL for agent's decision-making and first disruptive event.



(a) E1 (b) E2 (c) E3

Figure S9: Performance and reference curves of trees alive *per capita* using RL for agent's decision-making and first disruptive event.

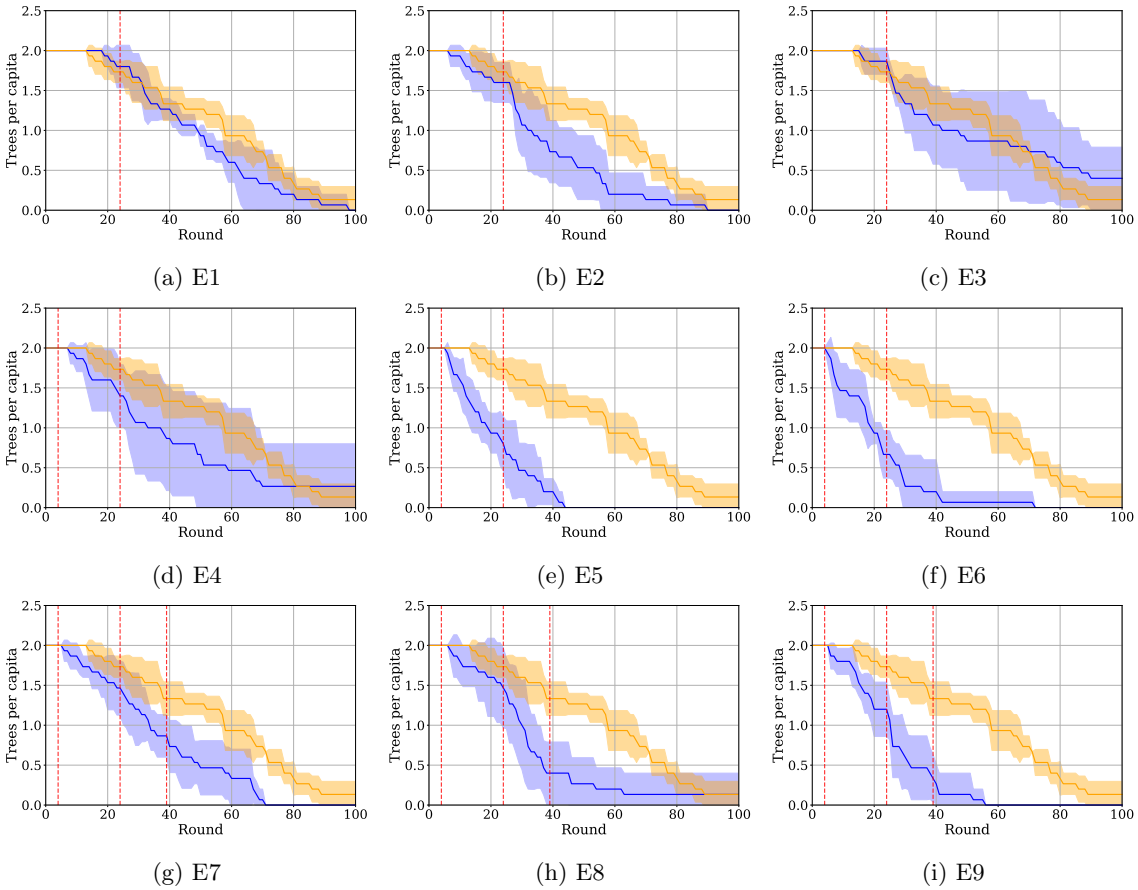


Figure S10: Performance and reference curves of trees alive *per capita* using LLM for agent's decision-making and first disruptive event.

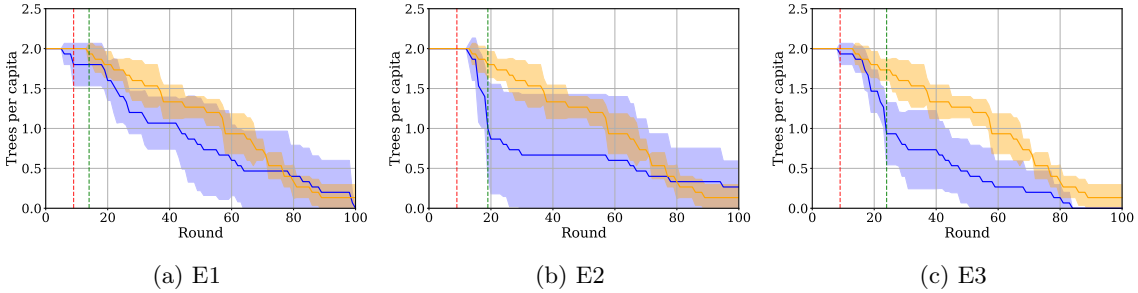
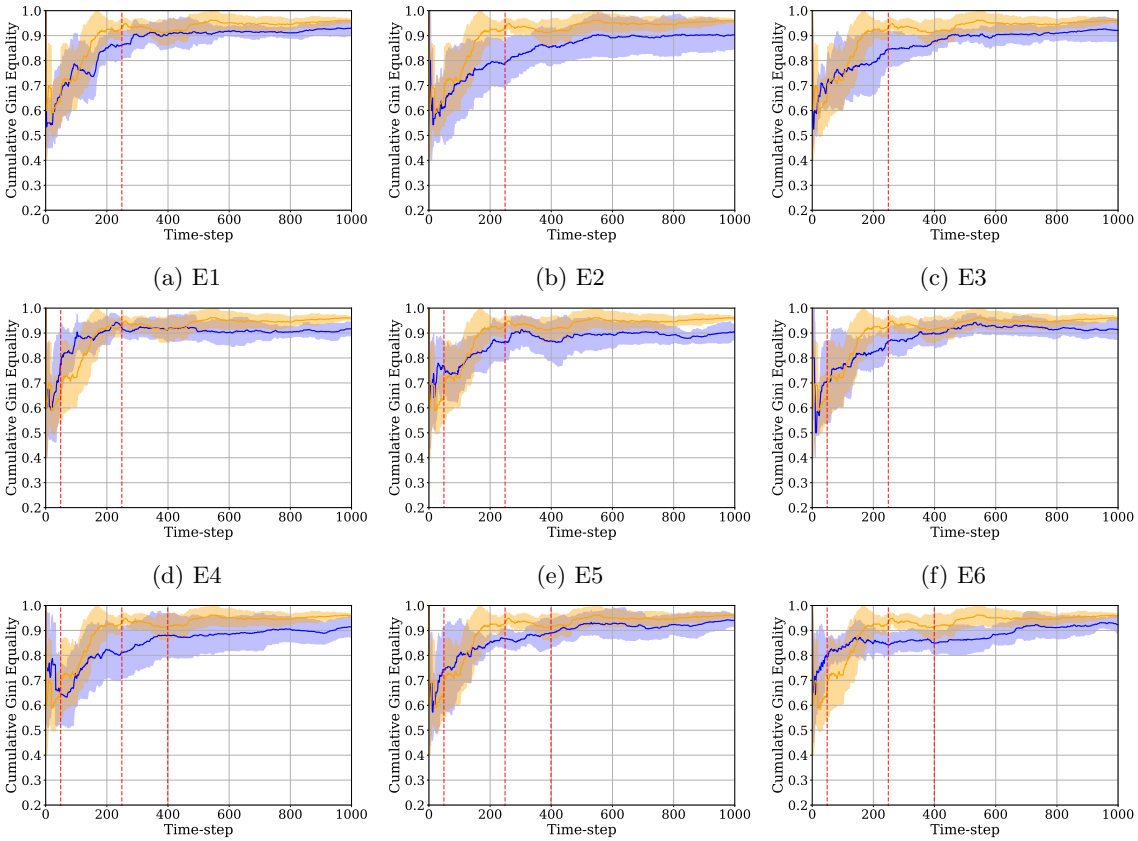


Figure S11: Performance and reference curves of trees alive *per capita* using LLM for agent's decision-making and first disruptive event.

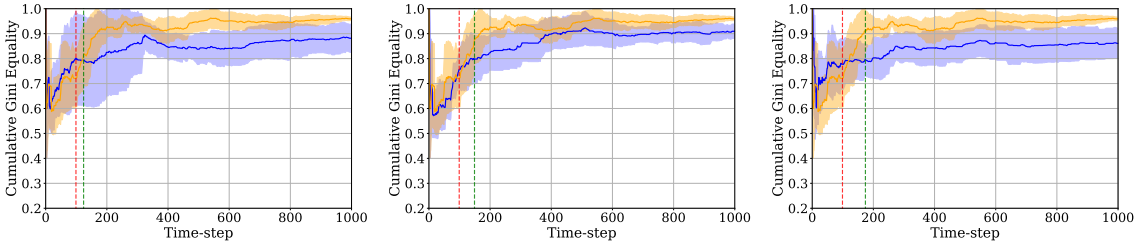


(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

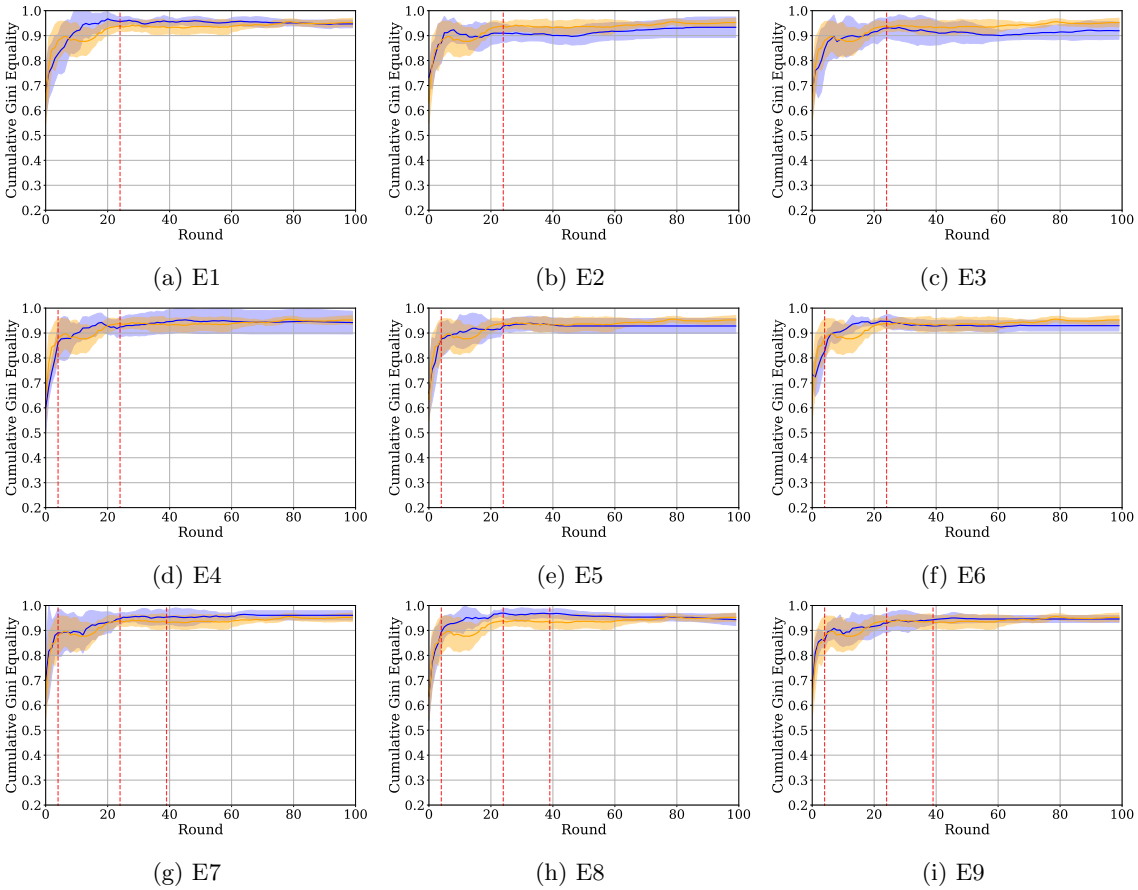
(g) E7 (h) E8 (i) E9

Figure S12: Performance and reference curves of Cumulative Gini Equality Index using RL for agent's decision-making with first disruptive event.



(a) E1 (b) E2 (c) E3

Figure S13: Performance and reference curves of Cumulative Gini Equality Index using RL for agent's decision-making with second disruptive event.

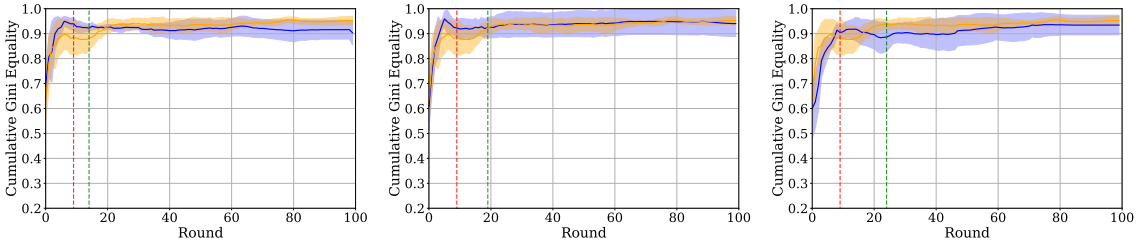


(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

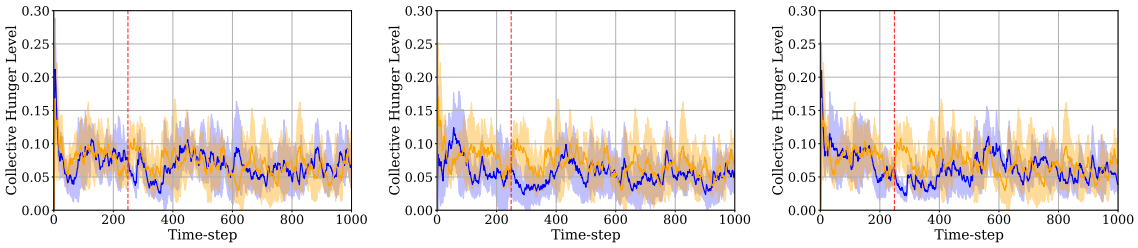
(g) E7 (h) E8 (i) E9

Figure S14: Performance and reference curves of Cumulative Gini Equality Index using LLM for agent's decision-making with first disruptive event.

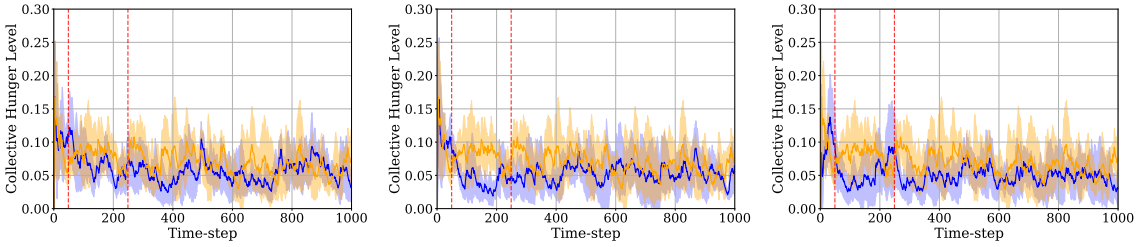


(a) E1 (b) E2 (c) E3

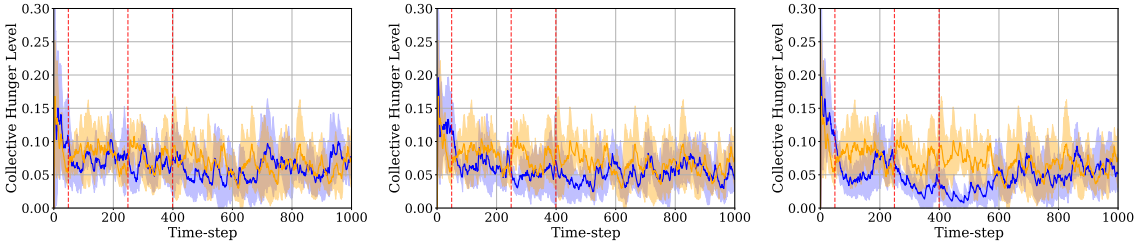
Figure S15: Performance and reference curves of Cumulative Gini Equality Index using LLM for agent's decision-making with second disruptive event.



(a) E1 (b) E2 (c) E3

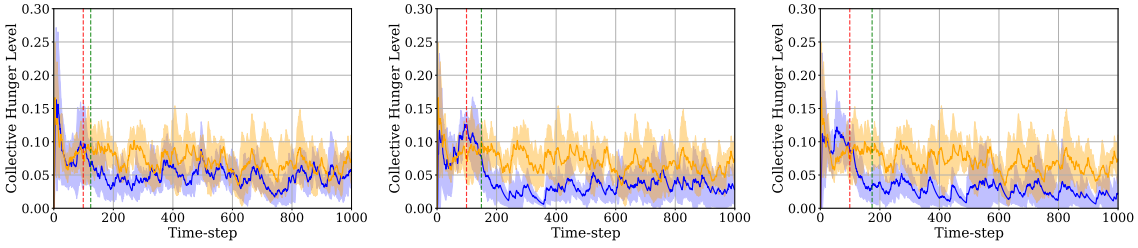


(d) E4 (e) E5 (f) E6



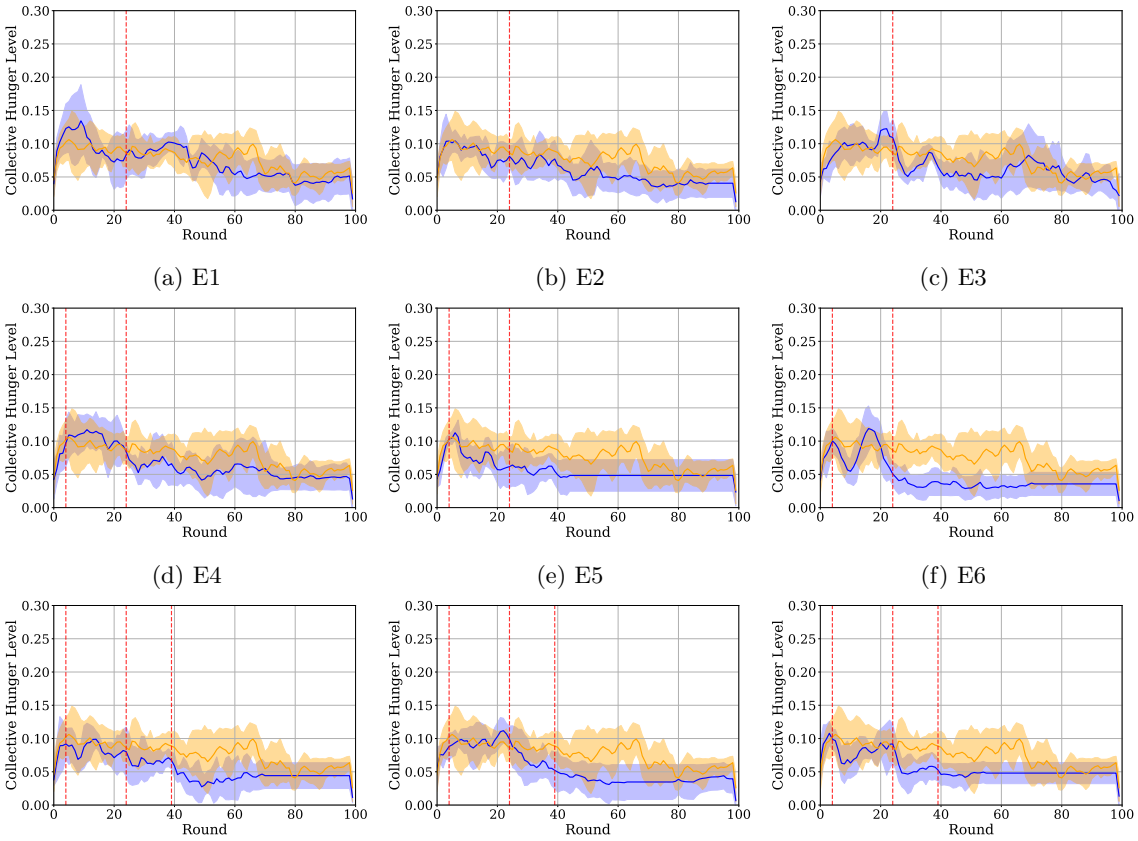
(g) E7 (h) E8 (i) E9

Figure S16: Performance and reference curves of Collective Hunger Level Index using RL for agent's decision-making with first disruptive event.



(a) E1 (b) E2 (c) E3

Figure S17: Performance and reference curves of Collective Hunger Level Index using RL for agent's decision-making with second disruptive event.

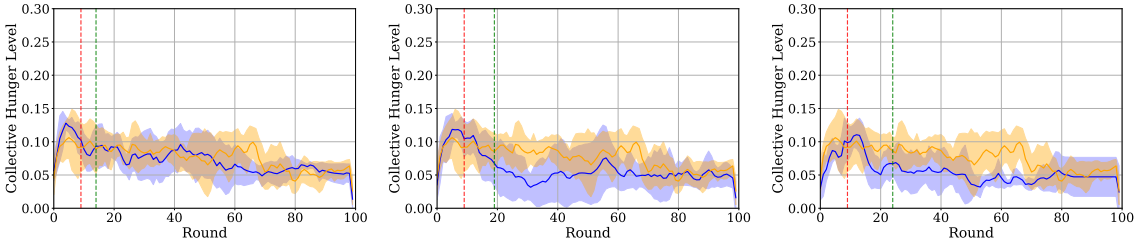


(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

(g) E7 (h) E8 (i) E9

Figure S18: Performance and reference curves of Collective Hunger Level Index using LLM for agent's decision-making with first disruptive event.



(a) E1 (b) E2 (c) E3

Figure S19: Performance and reference curves of Collective Hunger Level Index using LLM for agent's decision-making with second disruptive event.

References

[S1] Manuel Mosquera et al. "Can LLM-Augmented autonomous agents cooperate?, An evaluation of their cooperative capabilities through Melting Pot". In: *arXiv preprint arXiv:2403.11381* (2024).

[S2] Crawford S Holling. "Resilience and stability of ecological systems". In: *Annual review of ecology and systematics* 4.1 (1973), pp. 1–23.

[S3] Koenraad Van Meerbeek, Tommaso Jucker, and Jens-Christian Svenning. "Unifying the concepts of stability and resilience in ecology". In: *Journal of Ecology* 109.9 (2021), pp. 3114–3132.

[S4] Alan Hastings. "Transient dynamics and persistence of ecological systems". In: *Ecology Letters* 4.3 (2001), pp. 215–220.

[S5] Hana Krakovská, Christian Kuehn, and Iacopo P Longo. "Resilience of dynamical systems". In: *European Journal of Applied Mathematics* (2021), pp. 1–46.

[S6] Giacomo Como. "On resilient control of dynamical flow networks". In: *Annual Reviews in Control* 43 (2017), pp. 80–90.

[S7] David Fletcher and Mustafa Sarkar. "Psychological resilience". In: *European psychologist* (2013).

[S8] Gang Wu et al. "Understanding resilience". In: *Frontiers in behavioral neuroscience* 7 (2013), p. 10.

[S9] Helen Herrman et al. "What is resilience?" In: *The Canadian Journal of Psychiatry* 56.5 (2011), pp. 258–265.

[S10] Odin Hjemdal et al. "Resilience predicting psychiatric symptoms: A prospective study of protective factors and their role in adjustment to stressful life events". In: *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 13.3 (2006), pp. 194–201.

[S11] Antonella Sisto et al. "Towards a transversal definition of psychological resilience: A literature review". In: *Medicina* 55.11 (2019), p. 745.

[S12] Shae-Leigh Cynthia Vella and Nagesh B Pai. "A theoretical review of psychological resilience: Defining resilience and resilience research over the decades". In: *Archives of Medicine and Health Sciences* 7.2 (2019), pp. 233–239.

[S13] George A Bonanno. "Loss, trauma, and human resilience: have we underestimated the human capacity to thrive after extremely aversive events?" In: *The American psychologist* (2008).

[S14] Monica Perdomo, Flor Sanchez, and Amalio Blanco. "Effects of a community resilience intervention program on victims of forced displacement: A case study". In: *Journal of Community Psychology* 49.6 (2021), pp. 1630–1647.

[S15] Suniya S Luthar and Dante Cicchetti. "The construct of resilience: Implications for interventions and social policies". In: *Development and psychopathology* 12.4 (2000), pp. 857–885.

[S16] Mary C Davis, Linda Luecken, and Kathryn Lemery-Chalfant. "Resilience in common life: introduction to the special issue." In: *Journal of personality* (2009).

[S17] Adam Rose. "Economic resilience to natural and man-made disasters: Multidisciplinary origins and contextual dimensions". In: *Environmental Hazards* 7.4 (2007), pp. 383–398.

[S18] Lino Briguglio et al. "Economic vulnerability and resilience: concepts and measurements". In: *Measuring Vulnerability in Developing Countries*. Routledge, 2014, pp. 47–65.

[S19] James Simmie and Ron Martin. "The economic resilience of regions: towards an evolutionary approach". In: *Cambridge journal of regions, economy and society* 3.1 (2010), pp. 27–43.

[S20] Raúl Jiménez et al. "Emergence and resilience of cooperation in the spatial prisoner's dilemma via a reward mechanism". In: *Journal of theoretical biology* 250.3 (2008), pp. 475–483.

[S21] Haydée Lugo and Raúl Jiménez. "Incentives to cooperate in network formation". In: *Computational Economics* 28 (2006), pp. 15–27.

[S22] Erin K Chiou and John D Lee. "Cooperation in human-agent systems to support resilience: A microworld experiment". In: *Human factors* 58.6 (2016), pp. 846–863.

[S23] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge: Cambridge University Press, 2016.

[S24] Jianxi Gao, Baruch Barzel, and Albert-László Barabási. "Universal resilience patterns in complex networks". In: *Nature* 530.7590 (2016), pp. 307–312.

[S25] Xueming Liu et al. "Network resilience". In: *Physics Reports* 971 (2022), pp. 1–108.

[S26] JL Carlson et al. *Resilience: Theory and Application*. Tech. rep. Argonne National Lab.(ANL), Argonne, IL (United States), 2012.

[S27] Youmin Zhang and Jin Jiang. "Bibliographical review on reconfigurable fault-tolerant control systems". In: *Annual reviews in control* 32.2 (2008), pp. 229–252.

[S28] Yunhan Huang, Linan Huang, and Quanyan Zhu. "Reinforcement learning for feedback-enabled cyber resilience". In: *Annual reviews in control* 53 (2022), pp. 273–295.

[S29] Oliver Eigner et al. "Towards resilient artificial intelligence: Survey and research issues". In: *Proceedings of IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE. 2021, pp. 536–542.

[S30] Susmit Jha. "Trust, resilience and interpretability of AI models". In: *Numerical Software Verification: The 12th International Workshop, NSV 2019, New York City, NY, USA, July 13–14, 2019, Proceedings 12*. Springer. 2019, pp. 3–25.

[S31] John P Agapiou et al. "Melting Pot 2.0". In: *arXiv preprint arXiv:2211.13746* (2022).

[S32] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. *Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning with Shapley Values*. 2021. arXiv: 2110.01307 [cs.AI].

[S33] David Radke and Kyle Tilbury. *Learning to Learn Group Alignment: A Self-Tuning Credo Framework with Multiagent Teams*. 2023. arXiv: 2304.07337 [cs.AI].