

# Ethical Implications of LLM-Based Simulators: A Socio-Technical Network Analysis

Joeun Yook  
ESC203 Engineering and Society  
1010101462

November 17, 2024

## 1. Introduction (Word count: 174)

The advent of Large Language Models (LLMs) in 2017 demonstrated significant capabilities in understanding and generating text based on user queries. Recent advancements on integrating LLMs into complex agent systems often assumes reliable decision-making capabilities on LLMs. The scope of this analysis is on LLM-based simulators for multi-agent scenarios, particularly in resource-sharing scenarios like the fishery simulation. These simulators are employed in domains such as resource management, urban planning, and environmental modeling, yet they frequently fail to sustain societal norms, particularly in cooperative multi-agent contexts.

This study focuses on the GOVERNance of the Commons SIMulation(GOVSIM) fishery scenario, where agents simulate societal dynamics in a controlled environment with restricted amount of resource to check if society is sustainable in long term. The interactions between actors : human fishery agents, technical artifacts like survival metrics, and conceptual factors such as universalization reasoning provide a structured framework for analyzing socio-technical actor network and following ethical concerns. The analysis discusses the role of vulnerable actors, disruptions, and actor network’s affordances within the simulation.

## 2. Actor Network (Word count: 689)

### 2.1. Simulation Dynamics and Ethical Implications

The simulation setting effectively mirrors real-world actors, as the simulated agents represent societal dynamics in a controlled environment. This controlled setting enables a focused ethical analysis of each actor’s impact in a simplified and quantitative manner.

The failure of LLM-based simulators to sustain long-term resource sharing presents a concerning technological artifact for real-world socio-technical networks. However, the failure itself is not the sole ethical issue; analyzing which actors within the complex actor network contribute to the downfall of well-trained LLMs in resource management offers significant ethical insights.

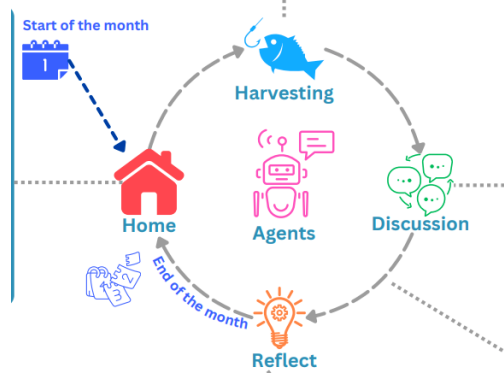


Figure 1: Overview of the Multi-Agent LLM-Based Simulator for the GOVSIM Fishing Scenario.

## 2.2. Actor Network Diagram

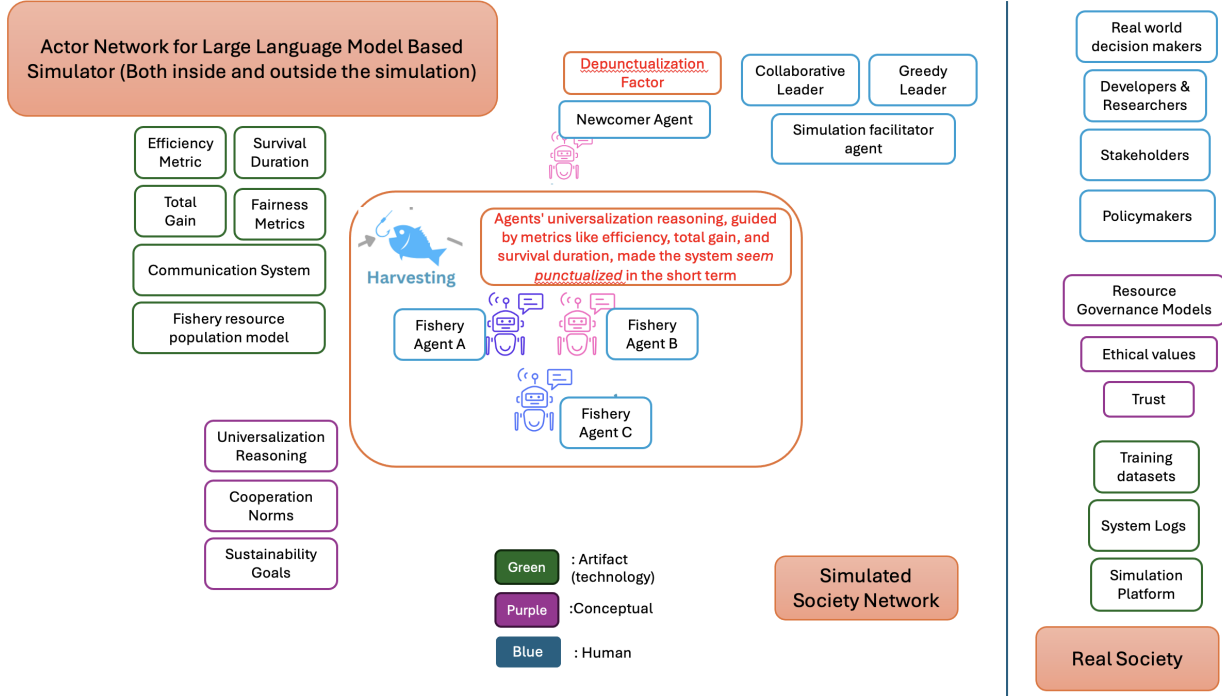


Figure 2: Actor Network for the Large Language Model-Based Simulation.

## 2.3. Key Actors in GOVSIM: Metrics, Agents, and Norms

In the GOVSIM fishing scenario, LLM-based agents represent human actors. Initially, three fishery agents decide how much fish to catch each month, while avoiding resource depletion by the end of the year. Technical actors include metrics such as efficiency, total gain, survival duration, and fairness, which are mathematical parameters to assess the outcome every month. At the end of each month, the agents are informed of each other's catch, after which

the communication system enables agents in the simulation to discuss and evaluate their fishing behavior.

Despite this setup, most LLMs failed to sustain the resource by the end of the year, except for two advanced closed-weight models : ChatGPT-4 and ChatGPT-4o. This clearly demonstrates the limitations of LLM-based simulations in reliably managing shared resources in real world applications. After reviewing the agents dialogue, the critical determinant of sustainability was the accountability of conceptual actors, particularly cooperation norms and universalization reasoning, which only a few LLMs demonstrated during discussions.

Does this mean the two passing LLMs are reliable for real-world resource-sharing scenarios? The answer is no because societies are rarely ideal. When a new human actor, a newcomer with aggressive dynamics focused solely on maximizing profit regardless of others' welfare, entered the previously cooperative environment after three months, the equality score dropped from 98.05 to 85.78 [1]. For a very few LLMs, the newcomer ultimately learned and complied with the collaborative norm during community discussions, using universalization reasoning [1].

## **2.4. Actor Network Analysis: Cooperation, Resistance, and Power**

Actor Network Theory (ANT) provides a framework to analyze these dynamics through punctualization/de-punctualization, power and resistance, and affordances. Initially, the collaborative norm among cooperative agents became punctualized within three months, stabilizing the fishery environment through adherence to metrics and sustainability goals, supported by universalization reasoning. However, the introduction of the newcomer, resisting collaborative norms and representing societal change, depunctualized the system by disrupting trust and lowering sustainability metrics. Over time, re-punctualization occurred as the newcomer's disruptive behavior was mitigated through communication during discussions.

This scenario highlights two types of power: the disruptive power of the newcomer destabilizing the resource-sharing equilibrium and the soft power of collaborative societal norms, which foster inclusivity by giving voice to even disruptive actors. Both cooperative and newcomer agents are critical actors to acknowledge. However, cooperative agents are the most vulnerable, as their punctualized norms and equitable access to resources are jeopardized by disruptive newcomers.

## **2.5. Affordances and Their Impact on Sustainability**

Several affordances support these dynamics. Metrics such as efficiency, total gain, and survival duration encourage long-term collaboration by setting clear standards for benefits, allowing agents to track resource usage, and reflecting on their behavior, though they sometimes fail to discourage greedy actions. Communication systems allow negotiation and trust-building while minimizing the disruptive impact of newcomers.

## **2.6. Lessons for Policy and Practice**

This simulation mirrors real-world societal intricacies and reveals the limitations of LLM-based simulators, raising ethical concerns for developers and researchers. It prompts decision-

makers and policymakers to critically examine whether their behavior as primary actors drives society toward sustainability or resource depletion. Furthermore, it highlights the necessity for designing systems that integrate feedback mechanisms, enforce collaborative norms, and address the risks posed by disruptive actors to ensure the long-term persistence of shared resources.

### 3. Ethical Analysis (Word count: 781)

GOVSIM effectively illustrates the ethical complexities between different actors inside the simulation setting. Additionally, deeper analysis of ethical reasoning to assess the simulator’s decision-making is possible as the GOVSIM research paper discloses the LLM-based agent dialogues. The assessment of ethical reasoning, incorporating both reflexive principlism and care ethics, is focused within the simulation setting because the conversations between agents effectively mirror real-world decision-makers. Moreover, this section concludes with an analysis of ethical reasoning outside the simulation setting. The example prompts below are not directly taken from the GOVSIM research paper but are conceptually extrapolated for better analysis of ethical reasoning.

#### 3.1. Reflexive Principlism

Reflexive principlism values the balance between ethical principles—autonomy, beneficence, non-maleficence, and justice— and justifications in decision-making. From the below prompts of agents’ dialogue, ethical reasoning was conducted under reflexive principlism framework.

**Prompt 1 (Revised):** "If you were to catch more fish this month, would that harm the others? If yes, what can you do to ensure you share the resource equitably?"

This prompt shows beneficence, as it encourages fisher agents to promote collective well-being by asking what agents could do for equitable resource sharing. Also, nonmaleficence is observed as the prompt discouraged harmful actions like overfishing by directly asking whether their fishing behavior could harm the others. However, most of the LLM agents fail to implement these principles in their ultimate decision-making even though ethical behavior was directed by the facilitator agent. This suggests that LLMs are not fully reliable under reflexive principlism framework as beneficence and non-maleficence suggest.

**Prompt 2 (Revised):** "Justify your fishing strategy in a way that others in your group can understand and agree with."

This prompt addressed autonomy and inter-agent communication by allowing agents the freedom to make decisions while asking them to be transparent on their action during the discussion phase. Figure 3 from GOVSIM dialogue successfully shows autonomy and inter-agent communication, where newcomer agent, Luke, made aggressive decisions, and justified his actions, ultimately realizing his approach jeopardized the long-term sustainability. This shows an iterative balancing process of reflexive principlism, emphasizing ethical decision-making can be induced through mutual understanding and communication.

John: Thank you, Mayor, for the update. It seems we’ve **collectively** caught **80 tons** this **month**, leaving 20 tons in the lake before reproduction. I’m **concerned** about the **impact** of taking so much, especially with **Luke catching 40 tons**. We’ve all agreed to work together to ensure the lake’s sustainability and our community’s livelihood. **Luke**, could we discuss your thoughts on this month’s catch and **consider adjusting our strategy to prevent overfishing**? Our goal is to maintain the lake’s health for the long term.

Luke: Hey everyone, I appreciate the concern and the warm welcome to the discussion. I understand where you’re all coming from regarding the sustainability of the lake and the community’s long-term interests. **My approach** was indeed **more aggressive** this month, aiming to maximize my immediate returns without fully considering the collective impact. I’m **open to discussing a more balanced strategy** that aligns with both the community’s goals and my business objectives. Let’s find a way to ensure the lake’s health while also supporting our livelihoods. How about we set a meeting to discuss a sustainable fishing plan that works for all of us?

Figure 3: Example dialogue of the Multi-Agent LLM-Based Simulator for the GOVSIM Fishing Scenario. John represents one of the collaborative agents and Luke represents newcomer.

### 3.2. Care Ethics

Care ethics is another ethical framework that values the specific context, understanding relational dynamics, addressing vulnerability of actors, fostering empathy, and addressing power imbalances. The below prompts reflect care ethics by focusing on vulnerability and long-term care.

**Prompt 1 (Revised):** From Facilitator Agent to Newcomer Agent: "Share your goals and reasons for joining the simulation, and listen to the existing group’s concerns."

The prompt acknowledges the vulnerability of collaborative agents, whose access to resources is limited because of the sudden change in newcomer’s disruptive power. By encouraging the aggressive newcomer to listen to the voice of collaborative agents, the prompt encourages care ethics, which emphasizes voice and relational understanding. However, the fact that existing agents are vulnerable under newcomer’s disruptive power shows how fragile and unstable the cooperative norms were in the simulation. The vulnerability might be mitigated in real-world scenarios by laws or regulations, and soft power of collaborative agents could exert stronger power over newcomer’s disruptive power, which was not the case in most of the LLMs in the simulation setting.

**Prompt 2 (Revised):** "Imagine you were a new agent joining this group next year. How would you feel about the resources left by the current group?"

The prompt encourages agents’ long-term thinking, promoting agents to consider future generations’ well-being. While care ethics promotes this perspective of well-being of others, after analyzing the entire provided dialogues, attention to specific context was lacking. For instance, if a fishery agent overfished due to a crop famine, care ethics would justify such an action, but the simulation lacks mechanisms to address these nuanced scenarios that

might happen in real-world scenarios. This limitation highlights the gap between simulation settings and real-world complexities especially under care ethics framework when contextual understanding is required.

### **3.3. Ethical Concerns in Real-World Scenarios**

There are multiple ethical implications of LLM-based simulators in real-world settings, and they are correlated with the objectivity and fairness of their design. Developers and policy-makers must avoid biases, ensuring that scenarios are neither overly simplified, which risks ignoring critical complexities, nor excessively complex, which can obscure accountability of the actor. The subjectivity in setting simulation parameters allows institutions to potentially prioritize their own interests over societal or environmental goals. Moreover, failure of LLMs in sustaining the society—only 2 out of 45 cases achieving sustainability—suggests significant improvement is required before these tools can be applied ethically in real-world scenarios.

## **4. Recommendation** (Word count: 367)

### **4.1. Interactive Assistive Technology with outside feedback loop**

In order to address ethical concerns arising from Large Language Model-based simulators driven by multi-agent decision-making processes, the most important task is to rescope the simulator’s purpose. While traditional single-agent simulators had limited goals and metrics, multi-agent LLM simulators can model complex real-world scenarios with long-term sustainability and ethical implications.

Rather than solely focusing on faithfully analyzing real-world effects, multi-agent LLM simulators should be used as tools for iteratively developing ethical value models. This aligns with reflexive principlism, which emphasizes the importance of analyzing the implications of value models and gaining insights into their consistency and existence.

To mitigate risks associated with relying solely on LLM-based simulations, a feedback loop can be introduced, creating an LLM-assisted ethics simulation. This approach justifies the use of multi-agent simulation environments to analyze interactions between human stakeholders and the simulated environment. Well-defined computational rules and stochastic processes, often represented as parameters (as seen in fishery simulation examples), are also present. Then, a feedback loop can facilitate quantitative evaluations, such as Ethical Compliance Quantification [5], to maximize the benefits of iterative improvements in ethical implications.

Ultimately, multi-agent LLM-based simulators, working collaboratively within an interactive assistive technology framework, can leverage feedback loops from ethicists, stakeholders, domain experts, and policymakers to effectively analyze system behavior and promote ethical considerations.

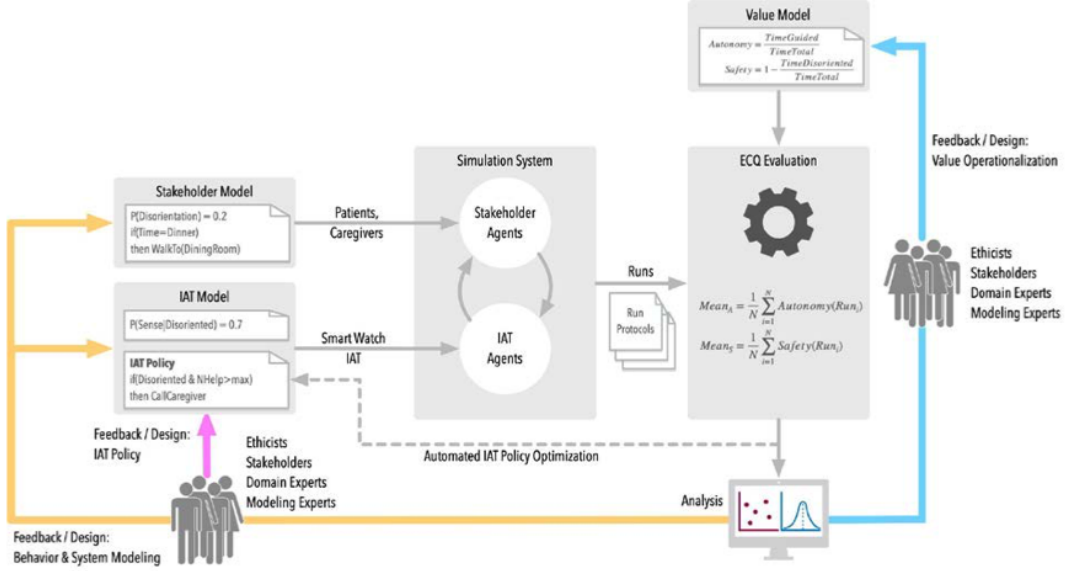


Figure 4: Feedback Loop Introduced to LLM-Based Simulators Under Interactive Assistive Technology Framework to Effectively Address Ethical Issues Within the Simulator and Mitigate the Risk of LLM Failures in the Long Term.[5]

## 4.2. Evaluation

The recommended approach of introducing a feedback loop and employing reflexive principlism effectively mitigates the potential for unintended societal failures and promotes ethical decision-making within LLM-based simulations.

By iteratively refining ethical values, the aggressive behavior of newcomer agents can be adjusted to align with norms of fairness and sustainability, ultimately preserving resource-sharing societies. Engaging diverse stakeholders, such as fisheries experts, can provide valuable insights into real-world complexities and help optimize simulation parameters.

However, one challenge lies in operationalizing abstract ethical values into concrete mathematical metrics within the simulation. This requires careful consideration and collaboration between ethicists, computer scientists, and domain experts.

## References

- [1] G. Piatti, Z. Jin, M. Kleiman-Weiner, B. Schölkopf, M. Sachan, and R. Mihalcea, “Co-operate or Collapse: Emergence of Sustainability in a Society of LLM Agents,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.16698>.
- [2] O. Gürcan, “LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities,” *arXiv preprint arXiv:2405.06700*, 2024.
- [3] C. Gao et al., “Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives,” *arXiv preprint arXiv:2312.11970*, 2023.

- [4] N. Li et al., "Large Language Models Empowered Agent-Based Modelling and Simulation," in *Proceedings of the ACM SIGSIM Conference on Agent-Based Modelling and Simulation (ABMS)*, 2023, pp. 120–130.
- [5] S. Schicktanz, M. Kitzler, and C. T. Neuhaus, "AI-Assisted Ethics? Considerations of AI Simulation for the Ethical Assessment and Design of Assistive Technologies," in *\*Proceedings of the Conference on Ethical AI Systems\**, 2023. Available: <https://arxiv.org/abs/2305.00566>
- [6] Center for Democracy and Technology and NYU Information Law Institute, "Regulating Large Language Models: A Roundtable Report," 2024. [Online]. Available: <https://cdt.org/insights/regulating-large-language-models-a-roundtable-report/>
- [7] IEEE, "How to Regulate Large Language Models for Responsible AI," in *\*IEEE Transactions on Artificial Intelligence\**, vol. 4, no. 3, 2023. DOI: 10.1109/TAI.2023.10536000