Research Proposal: Exploring Super-Alignment through Relative Ethics in Multi-Agent Systems using AutoGen

Eric Moore - 9/28/2024

Note: there are serious ethical concerns with this research that need careful review.

Abstract

In the advent of advanced artificial intelligence and potential encounters with non-human sentiences, establishing robust ethical frameworks becomes paramount. This proposal presents a pioneering research project that leverages AutoGen to explore ethical decision-making within multi-agent systems through the lens of **Relative Ethics** —a framework that transcends traditional alignment by encompassing all forms of non-human intelligence. By modeling diverse non-human societies using varied golden patterns and applying distinct criteria for Multiple Chains of Thought (MCoT) downselection, I aim to investigate the dynamics of super-alignment.

A critical component of this research is the simulation of the historical interaction between Neanderthals and Homo sapiens, providing invaluable insights into interspecies ethics and coexistence. Utilizing asynchronous nested chats via `a_initiate_chats()` in AutoGen, we ensure scalable and efficient simulations. The **o1** language model will serve as the "philosophy agent," assigning golden patterns and guiding ethical reasoning, while **o1-mini** evaluates MCoT threads to select the most effective reasoning paths. This innovative approach not only refines human ethical models but also prepares us for future interactions with advanced AI or extraterrestrial life.

Introduction and Background

The rapid advancement of Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI) necessitates a reevaluation of our ethical frameworks. Traditional alignment focuses on aligning AI behavior with human values, yet this scope is limited when considering potential interactions with non-human sentiences, whether artificial or extraterrestrial. To address this gap, I introduce **Relative Ethics**—an ethical framework designed for **super-alignment** that adapts to the moral perspectives of diverse intelligences.

The historical interaction between Neanderthals and Homo sapiens offers a profound case study. This pivotal period exemplifies the complexities of coexistence between distinct sentient species, providing a tangible foundation for modeling and understanding interspecies ethics.

Key Objectives

1. **Develop a Robust Multi-Agent System within AutoGen:**
    - **Simulation Environment:** Construct an advanced simulation where GPT-4-powered agents interact within diverse ethical scenarios reflecting both historical and hypothetical contexts.
    - **Philosophy Guidance:** Employ **o1** as the "philosophy agent" to imbue agents with varied ethical reasoning patterns.
2. **Implement the Framework of Relative Ethics:**
    - **Super-Alignment Model:** Establish a model that extends ethical considerations beyond AGI and ASI to all non-human sentiences.
    - **Adaptive Ethics:** Design agents capable of adapting their ethical reasoning to align with different societal norms and values.
3. **Model Diverse Non-Human Societies:**
    - **Golden Patterns Allocation:** Utilize **o1** to assign distinct golden patterns to agents, each representing unique non-human societal frameworks, including those inspired by Neanderthal intelligence.
    - **Societal Representation:** Capture the essence of various hypothetical societies, emphasizing their ethical norms and decision-making processes.
4. **Integrate Historical Case Study — Neanderthals and Homo sapiens:**
    - **Reconstruction of Interactions:** Simulate scenarios based on archaeological and anthropological findings to explore ethical considerations in interspecies interactions.
    - **Ethical Analysis:** Examine the outcomes to derive insights into coexistence, competition, and integration between species.
5. **Apply Game-Theoretic Principles and MCoT Reasoning:**
    - **Strategic Interactions:** Incorporate game theory to model conflict and cooperation within and between societies.
    - **Multiple Reasoning Paths (MCoT):** Enable agents to pursue concurrent reasoning threads, exploring a spectrum of ethical decisions.
6. **Apply Distinct Criteria for MCoT Downselection:**
    - **Criteria Definition:** Define customized criteria reflective of each golden pattern to evaluate and select optimal MCoT threads.
    - **Contextual Evaluation:** Ensure downselection aligns with the ethical paradigms of each represented society.
7. **Leverage Asynchronous Nested Chats with `a_initiate_chats()`:**
    - **Fan-Out/Fan-In Operations:** Utilize asynchronous nested chats to manage parallel reasoning processes and consolidate results efficiently.
    - **Scalability and Efficiency:** Enhance the simulation's capacity to handle complex and large-scale ethical scenarios.
8. **Employ o1 and o1-mini for Guidance and Evaluation:**
    - **Dynamic Allocation: o1** assigns golden patterns and provides overarching philosophical guidance to each simulated society per their profile.

    - **Streamlined Evaluation: o1-mini** assesses MCoT threads, selecting those that best represent effective ethical reasoning per different simulated criteria.
9. **Evaluate the Impact of Different Reasoning Patterns:**
    - **Simulation Analysis:** Analyze how varied golden patterns and downselect evaluation criteria influence agents' ethical decisions and interactions.
    - **Comparative Effectiveness:** Identify which patterns lead to outcomes that are ethically robust and conducive to harmonious coexistence.
10. **Refine Human Ethical Models:**
    - **Integration of Insights:** Utilize findings to enhance human ethical frameworks, making them more adaptable and inclusive of non-human perspectives.
    - **Future Preparedness:** Establish foundations for ethical guidelines applicable to future interactions with AI and extraterrestrial intelligences.

Anticipated Outcomes

- **Advancement of Super-Alignment Theory:**
    - Development of a comprehensive model incorporating relative ethics, facilitating ethical alignment across diverse intelligences.
- **Enhanced Understanding of Interspecies Ethics:**
    - Insights into the ethical dimensions of the Neanderthal-Homo sapiens interaction, informing modern ethical considerations.
- **Innovative Simulation Framework:**
    - A scalable, efficient multi-agent system capable of modeling complex ethical scenarios using AutoGen and asynchronous nested chats.
- **Contributions to AI Ethics and Policy:**
    - Recommendations for policy and ethical guidelines that address the challenges of interacting with non-human sentiences.
- **Strengthened Case for Funding:**
    - A compelling proposal highlighting the project's significance, innovation, and potential impact on multiple fields.

Methodology

1. **Agent Configuration:**
    - **Golden Patterns:** Agents receive unique golden patterns from **o1**, reflecting different ethical frameworks and reasoning styles.
    - **Behavioral Simulation:** Agents interact within the simulation, making decisions based on their assigned patterns.
2. **Historical Case Study Simulation:**
    - **Data Integration:** Incorporate anthropological data on Neanderthals and Homo sapiens to recreate plausible interaction scenarios.

    - **Ethical Scenarios:** Design ethical dilemmas that mirror challenges faced during their coexistence.
3. **MCoT Implementation:**
    - **Parallel Reasoning:** Agents generate multiple chains of thought to explore various ethical decisions.
    - **Downselection Process: o1-mini** evaluates these chains, selecting the most contextually appropriate paths.
4. **Asynchronous Nested Chats:**
    - **Implementation:** Use `a_initiate_chats()` to manage complex interactions and reasoning processes among agents.
    - **Efficiency:** This approach allows for simultaneous exploration of multiple scenarios without compromising computational resources.
5. **Data Analysis:**
    - **Outcome Evaluation:** Analyze the decisions and interactions of agents to identify patterns and outcomes.
    - **Ethical Assessment:** Assess the effectiveness of different golden patterns in producing favorable ethical results.

Budget and Justification

- **Computational Resources:** High-performance computing is essential for running extensive simulations involving multiple agents and nested chats.
- **Personnel:** Funding for a multidisciplinary team, including experts in AI, ethics, anthropology, game theory, and computational modeling.
- **Software Development:** Investment in enhancing AutoGen's capabilities to support the project's innovative methodologies.
- **Data Acquisition:** Resources needed for sourcing detailed anthropological and historical data for accurate simulations.

Conclusion

This research aims to break new ground in the field of AI ethics by introducing the concept of **Relative Ethics** and applying it within a multi-agent simulation framework. By modeling both historical and hypothetical interactions between diverse sentient beings, I seek to develop ethical guidelines that are universally applicable. The inclusion of the Neanderthal and Homo sapiens case study not only enriches the research with real-world relevance but also bridges the gap between past and future ethical challenges.

Securing funding for this project will enable myself and any collaborators to address critical questions about coexistence with non-human sentiences, ultimately contributing to a more harmonious and ethically aligned future.

Relative Ethics is a specialized ethical framework designed to accommodate and adapt to the diverse moral perspectives of various intelligences, including non-human and artificial entities. Unlike traditional ethical theories that seek universal moral principles, Relative Ethics recognizes that moral values and principles may vary based on the context, culture, or nature of the intelligences involved. This framework facilitates ethical decision-making and interactions by considering the unique societal norms, values, and reasoning patterns inherent to each intelligence, ensuring that ethical considerations are both inclusive and adaptable.