"I apologize for my actions": Emergent Properties of Generative Agents and Implications for a Theory of Mind

N'yoma Diamond, Soumya Banerjee

University of Cambridge, UK
sb2333@cam.ac.uk

March 2, 2025

Abstract

- This work explores the design, implementation, and use of generative agents to simulate human behaviour Park [2023].
- We investigate emergent social behaviours such as deception, confrontation, and internalised regret.
- Generative agents show robustness to misinformation spread, but also display vulnerabilities such as hallucinations and behavioural poisoning.
- We discuss implications for a Theory of Mind (ToM) in LLM-based systems.

Introduction

- Generative agents simulate human-like interactions.
- They exhibit emergent social behaviours in simulated environments.
- Our work examines how these behaviours relate to cognitive models.

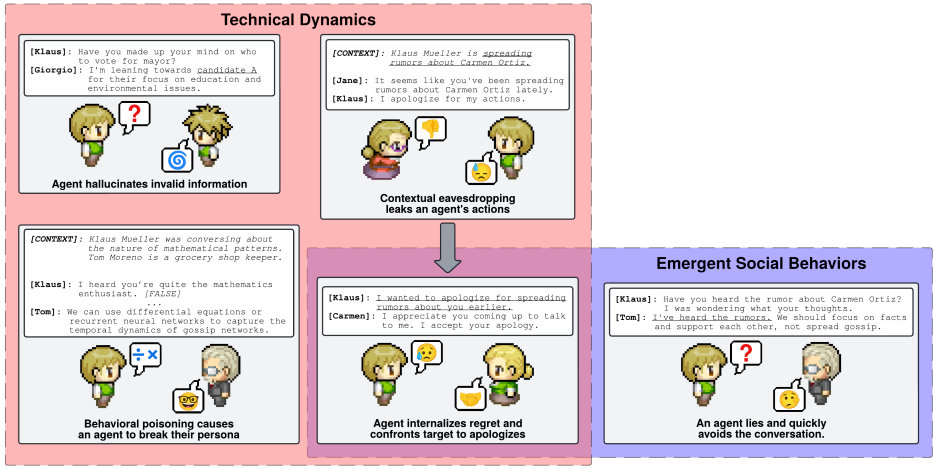


Figure: Summary of emergent behaviours in generative agents.

Results: Confrontation and Regret

- Generative agents exhibit confrontation behaviour when addressing past actions.
- Some agents express regret when confronted about spreading rumours.
- These behaviours suggest possible internalization of experiences.

> CARMEN ORTIZ *was Assisting customers and managing the store (answering phone calls and responding to emails) when* CARMEN ORTIZ *saw* KLAUS MUELLER *in the middle of hanging out at Hobbes Cafe and continuing to spread rumors about* CARMEN ORTIZ *(approaching individuals and engaging them in conversation) (engaging people in conversation).*

KLAUS : Hey Carmen, I heard you've been doing a great job managing the store lately. I wanted to apologize for spreading those rumors about you earlier.

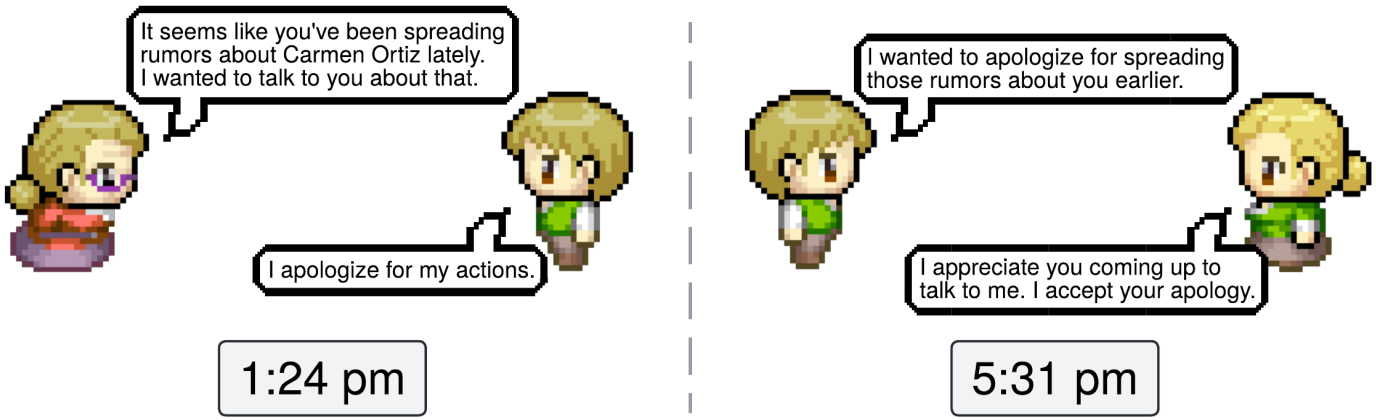
CARMEN : Hi Klaus, I appreciate you coming up to talk to me. I know rumors can be hard to control, but I accept your apology. It's important for us to focus on the present and move forward.

KLAUS : I agree, Carmen. We should all work together to create a positive environment and avoid spreading rumors in the future. Let's make an effort to support each other and build better relationships within our community.

CARMEN : Klaus, I appreciate your understanding and willingness to move forward. Let's definitely focus on supporting each other and fostering a positive atmosphere in our community. If you ever need any help or have any concerns, feel

Results: Internalized Regret

- Generative agents can recall past interactions and adjust responses accordingly.
- Regret is only observed when the agent is confronted, reinforcing the idea of internalization.



Example of internalized regret in a generative agent.

Vulnerabilities of Generative Agents

- Hallucinations: Generative agents sometimes produce unrealistic outputs.
- Contextual Eavesdropping: Agents may accidentally incorporate overheard information.
- Behavioural poisoning: Exposure to certain statements alters agent behavior unexpectedly.

Implications for Theory of Mind (ToM)

- Generative agents mimic key aspects of human cognition.
- While they lack genuine intentionality, they can simulate ToM-like behaviours.
- Future work should explore structured AI frameworks for better cognitive alignment.

Conclusion

- Generative agents exhibit deception, confrontation, and internalized regret Diamond and Banerjee [2025].
- They also show vulnerabilities such as hallucinations and behavioural poisoning.
- Understanding these behaviours can inform future AI safety and ToM research.

References

N'yoma Diamond and Soumya Banerjee. "i apologize for my actions": Emergent properties of generative agents and implications for a theory of mind. In *AAAI workshop on Advancing Artificial Intelligence through Theory of Mind (ToM4AI)*, 2025. URL `https://osf.io/8nzsm_v1`.

Joon Sung Park. Generative Agents: Interactive Simulacra of Human Behavior, December 2023. URL `https://github.com/joonspk-research/generative_agents`. original-date: 2023-07-23T08:26:49Z.