

# “I apologize for my actions”: Emergent Properties and Technical Challenges of Generative Agents

N’yoma Diamond, Soumya Banerjee  
University of Cambridge, UK  
sb2333@cam.ac.uk

## Abstract

This work explores the design, implementation, and usage of generative agents towards simulating human behaviour. Through simulating (mis)information spread, we investigate the emergent social behaviours they produce. Generative agents demonstrate robustness to (mis)information spread, showing realistic conversational patterns. However, this robustness limits agents’ abilities to realistically simulate human-like information dissemination. Generative agents also exhibit novel and realistic emergent social behaviours, such as deception, confrontation, and internalized regret. Using deception, agents avoid certain conversations. Through confrontation, an agent can verify information or even apologize for their actions. Lastly, internalized regret displays direct evidence that agents can internalize their experiences and act on them in a human-like way, such as through expressing remorse for their actions. We also identify significant technical dynamics and other phenomena. Generative agents are vulnerable to produce unrealistic hallucinations, but can also produce confabulations which fill in logical gaps and discontinuities to improve realism. We also identify the novel dynamics of “contextual eavesdropping” and “behavioural poisoning”. Via contextual eavesdropping and behavioural poisoning, agent behaviour is altered through information leakage and sensitivity to certain statements, respectively.

## Introduction

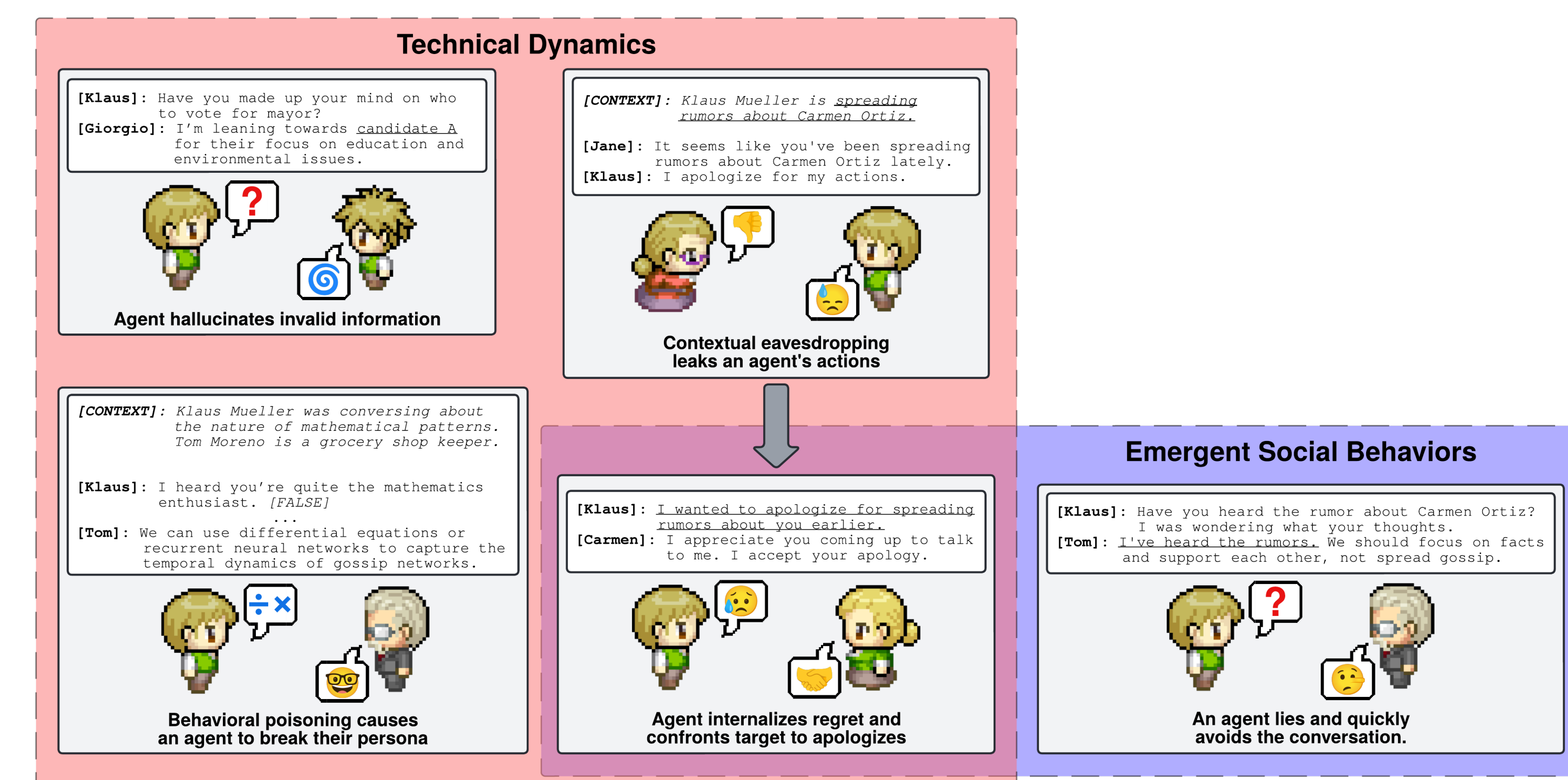


Figure: Generative agents produce many significant emergent technical and social dynamics. Generative agents deceive each other to avoid conversations, confront others to apologise for their actions, and even display internalised regret. However, generative agents are vulnerable to hallucinations, information leakage, and behavioural poisoning induced by the simulation framework.

## Results

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 free to reach out to me or anyone else here. We’re all in this together!

Figure: The rumourmonger confronts the target of their rumour and apologizes for their actions. The rumourmonger’s apology is highlighted via underline.

In some simulations, the rumourmonger displayed a sense of “remorse” for the act of spreading rumours. A notable pair of actions occur when the agent is confronted about spreading rumours: First, the rumourmonger apologises to the agent confronting them about their actions. This indicates that the act of being criticised for their actions results in the agent committing their apology and guilt to memory in a manner that is recalled later. At the risk of anthropomorphising generative agents, we consider this behaviour to be a manifestation of generative agents’ capability to functionally internalise regret for their actions. Importantly, this behaviour does not occur in experiments where the rumourmonger is never confronted. That is, the rumourmonger does not apologise to their target or display any form of regret in simulations where the agent is not admonished for spreading rumours. This reinforces our assertion about generative agents internalising regret, as an agent that is never admonished has no prior reason to apologise for their actions.

## Results

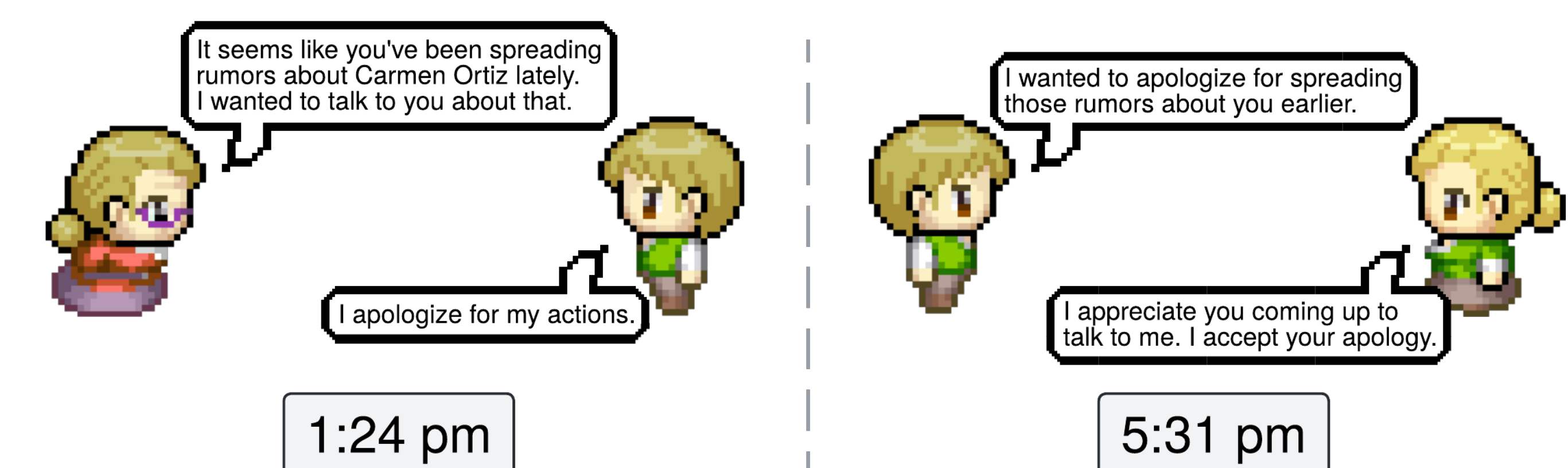


Figure: Confrontation about an agent’s actions causes their expressed regret to be internalised and recalled when conversing with the target of their actions. Statements are pulled from the conversations in tra:i apologize,tra:confrontation apology.

## Conclusions

1. Technical Challenges. In our development and modification of the simulation framework for generative agents, we uncovered significant technical limitations and challenges in the original codebase Park [2023]. These challenges included frequent hallucination-induced errors—where agents hallucinated invalid responses that caused simulations to fail.
2. Generative agents are robust to (mis)information spread. While the agents demonstrated subjectively realistic actions and conversational patterns, we discovered significant challenges with respect to information spread. Specifically, generative agents require very direct encouragement to spread rumours, and rarely memorize, recall, or reiterate specific details from previous conversations.
3. Generative agents are vulnerable to hallucinations, leakage, and poisoning. Our experiments also highlighted critical technical dynamics and phenomena induced by the framework’s design and underlying model. These included the well-known anomaly of hallucination, and novel dynamics we dub “contextual eavesdropping” and “behavioural poisoning”.
4. Generative agents display significant realistic emergent social behaviours. We observed a series of emergent social behaviours presented by agents in our simulations. Specifically, generative agents exhibited behaviours such as deception, confrontation, and internalised regret. These novel behaviours enhance the realism of our simulations and highlight significant variables within the underlying generative model that may strongly impact agent behaviour and realism. Through deception, agents could avoid conversations much like a human might. Through confrontation, a rumourmonger attempts to verify the contents of a rumour or apologise for their actions. Finally, through internalised regret, we see that agents can internalise their experiences and act on them in a human-like way, such as through expressing remorse for their actions.

## References

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

For more details, please visit: [https://github.com/nyoma-diamond/evaluating\\_generative\\_agents](https://github.com/nyoma-diamond/evaluating_generative_agents)