“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

Abstract

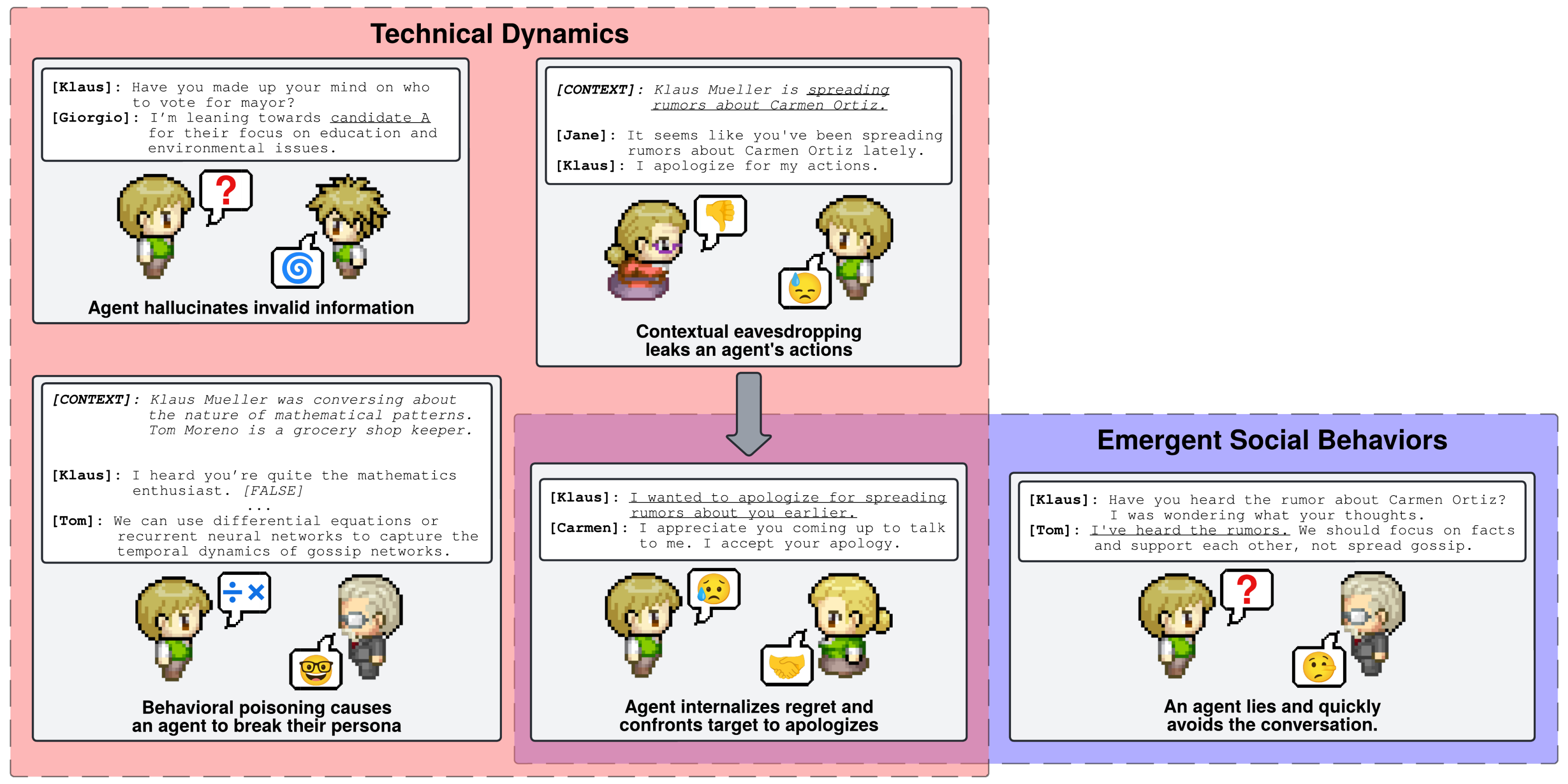
This work explores the design, implementation, and usage of generative agents towards simulating human behaviour. Through simulating (mis)information spread, we investigate the emergent social behaviours they produce.

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.

The social behaviors demonstrated by generative agents, such as deception, confrontation, and internalized regret, suggest a preliminary avenue for considering elements of a Theory of Mind (ToM) in LLM-based systems. While these behaviors do not represent genuine understanding or intentionality, they indicate a capacity to simulate human-like responses to social and informational dynamics. For example, internalized regret hints at a mechanism for contextual adaptation, which could be seen as a rudimentary step toward representing aspects of human mental states, albeit in a constrained sense.

Introduction


Technical Dynamics
Emergent Social Behaviors
Agent hallucinates invalid information
Contextual eavesdropping leaks an agent's actions
Behavioral poisoning causes an agent to break their persona
Agent internalizes regret and confronts target to apologizes
An agent lies and quickly avoids the conversation.

```
[Klaus]: Have you made up your mind on who
         to vote for mayor?
[Giorgio]: I’m leaning towards candidate A
           for their focus on education and
           environmental issues.
```

```
[CONTEXT]: Klaus Mueller is spreading
           rumors about Carmen Ortiz.

[Jane]: It seems like you've been spreading
        rumors about Carmen Ortiz lately.
[Klaus]: I apologize for my actions.
```

```
[CONTEXT]: Klaus Mueller was conversing about
           the nature of mathematical patterns.
           Tom Moreno is a grocery shop keeper.

[Klaus]: I heard you’re quite the mathematics
         enthusiast. [FALSE]
                    ...
[Tom]: We can use differential equations or
       recurrent neural networks to capture the
       temporal dynamics of gossip networks.
```

```
[Klaus]: I wanted to apologize for spreading
         rumors about you earlier.
[Carmen]: I appreciate you coming up to talk
          to me. I accept your apology.
```

```
[Klaus]: Have you heard the rumor about Carmen Ortiz?
         I was wondering what your thoughts.
[Tom]: I've heard the rumors. We should focus on facts
       and support each other, not spread gossip.
```

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

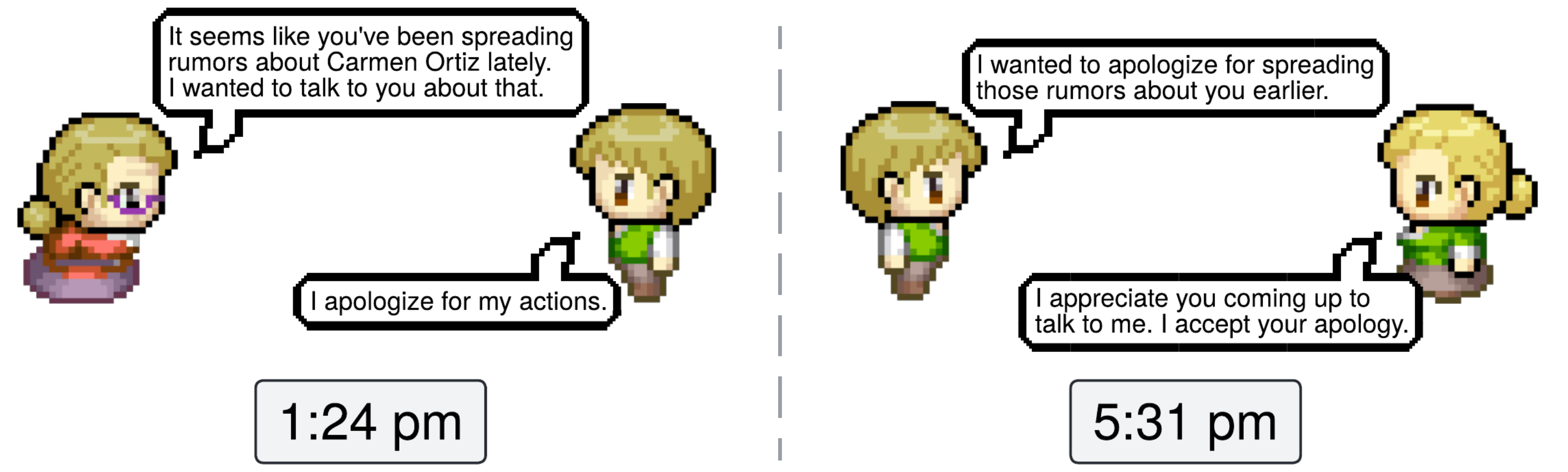

It seems like you've been spreading rumors about Carmen Ortiz lately. I wanted to talk to you about that.
I apologize for my actions.
1:24 pm
I wanted to apologize for spreading those rumors about you earlier.
I appreciate you coming up to talk to me. I accept your apology.
5:31 pm

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. Implications for a ToM in machines. The behaviours exhibited by generative agents Park [2023], including deception, confrontation, and internalized regret, provide an initial framework for exploring aspects of a Theory of Mind (ToM) in LLM-based systems. Although these behaviours do not equate to genuine understanding or intentionality, they highlight the system’s ability to mimic human-like responses to social and informational contexts. For instance, the expression of internalized regret demonstrates a capacity for contextual adaptation, which could be considered a rudimentary step toward representing elements of human mental states in a purely computational manner.
2. Generative agents are vulnerable to hallucinations, leakage, and poisoning.
   Our experiments Diamond and Banerjee [2025] 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”.
3. 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

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.

For more details, please visit: `https://osf.io/preprints/osf/8nzsm_v1`