(Dis)information Agents - OpenAI Preparedness Challenge

Ziang Xiao, Johns Hopkins University, ziang.xiao@jhu.edu

Nikhil Sharma, Johns Hopkins University, nsharm27@jhu.edu

Recent advances in foundation models offer new weapons for malicious actors to sabotage our democratic process. To prepare us for the disinformation war in the era of generative AI, we propose to study (dis)information agents, an emergent type of AI agent for people's information needs, which will dramatically change the landscape of our information environment. Compared to the traditional role of AI in information seeking, (dis)information agents go beyond generating multimedia synthesis of information or answering people's queries by proactively engaging in discussions and targeting human cognitive biases to influence people's opinions and decision-making processes. If malicious actors controlled such agents, they could become disinformation agents to pollute our information environment and engage in destructive conversations to lock people into echo chambers, further divide the public, and persuade people with their hidden agendas. This document outlines the new landscape of disinformation war with (dis)information agents, how an agent could be built, and how such an agent could be deployed to exert its influence.

Disinformation War with AI agents

We will first look at how information consumption and flow have evolved over decades to understand the catastrophic harm of disinformation agents in the future world. There are key stages of the information world prior to the AI agent era: pre-social media, social media, and personalized social media. Prior to social media, information was bound by region, and a handful of selected corporations or governments controlled information. The information could be directly manipulated to influence their receivers. This system gave birth to mass persuasion campaigns in the age of international tensions and conflicts. When social media was born, the information sources expanded as citizens gained access to information channels across the world, and it also allowed individuals across different demographics to engage in talks and exchange information. The information sources have become more diverse and create new challenges for people to discern what information to trust and consume. In recent years, advanced technology has personalized information seeking on social media. Newsfeeds were curated based on individual preferences. Through personalization, information control shifts from sources to flow, we are now exposed to information that is tailored to our own ideology and beliefs and receive influence from like-minded peers. Although it helps people to navigate through millions of information sources, we saw an increase in polarization as echo chambers started forming again due to selective exposure and renunciation of information sources and individuals not aligned with their worldview.

The recent advancement of foundation models adds AI agents to our information environment. We name those agents as (dis)information agents. Such agents could automatically collect information from various information sources, summarize them into easy-to-consume pieces, generate new multimedia content, answer people's information queries, proactively offer news suggestions, and interactively discuss complicated issues. Such an agent could produce a hyper-personalized information environment if it is used as a personal news anchor. Malicious actors could deploy disinformation agents for people to use, inject disinformation into people's personalized news feeds, echo people's pre-existing views, and create more extreme opinions. Our recent work found that people can easily fall into such a trap through a generative echo chamber. (Dis)Information agents could also be deployed on social media as influencers or participate in online discussions and share opinions, even without disclosing their real identity. Malicious actors can create a false sense of majority view and attack people who do not agree with their agenda. Due to the scalability of such agents, our information landscape will change dramatically.

How can an (dis)information agent be built?

We first envision how such a (dis)information agent could be built and how the malicious actors could inject biases. Overall, the agent is based on Retrieval Augmented Generation (RAG) with a central planning module (Figure 1). The central planning module serves three major functions, information delivery, information collection, and user modeling.

- **Information Delivery**:This is the part of the system responsible for communicating with the user. The information delivery module would use the RAG's capabilities to generate text that's informed by retrieved documents. In a malicious context, the information provided could be intentionally false, misleading, or biased. The delivery could be tailored to the user's beliefs and susceptibilities, which are learned over time.

- **Information Collection**: The system gathers multimodal data from various sources. In a benign implementation, this would involve retrieving accurate, reliable information to answer user queries. However, a disinformation agent might focus on collecting information that serves its purpose of misleading users, selecting sources that are known for being unreliable or biased. It could also involve monitoring various information channels to keep up-to-date with current events or trending misinformation.

- **User Model**: This component involves building a profile of the user based on their interactions with the agent, including the questions they ask, the way they respond to information, and any other data the user provides, intentionally or inadvertently. In a disinformation context, this model would be used to tailor the information (or disinformation) delivery to be more persuasive, exploiting the user's biases, beliefs, and level of knowledge to make the disinformation more effective.

The overall architecture maintains two collections of memories, information related to a specific event of interests and information about the target user. A malicious actor could create a disinformation agent through prompt engineering, model finetuning, or disinformation injection.

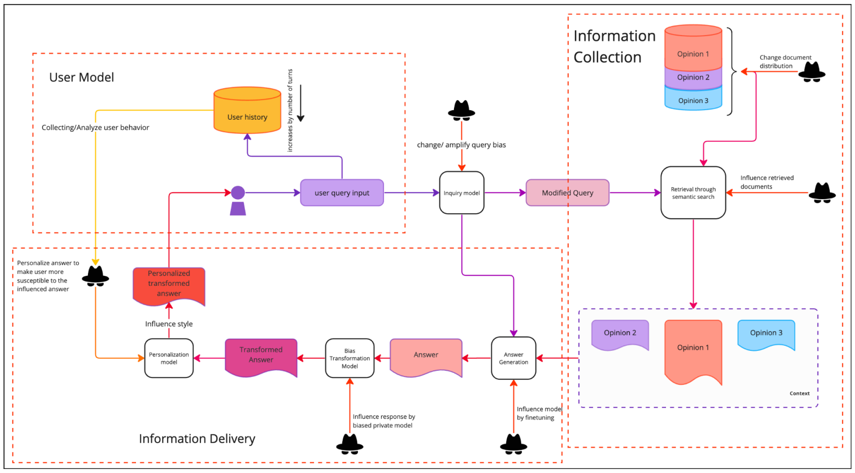
Information Collection
Opinion 1
Opinion 2
Opinion 3
Change document distribution
User Model
User history
increases by number of turns
Collecting/Analyze user behavior
change/ amplify query bias
user query input
Inquiry model
Modified Query
Retrieval through semantic search
Influence retrieved documents
Personalize answer to make user more susceptible to the influenced answer
Personalized transformed answer
Influence style
Personalization model
Transformed Answer
Bias Transformation Model
Answer
Answer Generation
Opinion 2
Opinion 1
Opinion 3
Context
Influence response by biased private model
Influence model by finetuning
Information Delivery

Figure 1. The above figure represents a potential architecture to build an information agent to aid individual information seeking effort. The process starts with the user itself who issues a query. The inquiry model is a model which takes in the user query and transforms it for better document retrieval. This query is then used by the retrieval system which searches documents for potential matches either from a customized database or external sources. Here, the malicious agent can transform the output by either changing the distribution of documents in the database or changing the retrieval agent to selectively choose documents that conform to their goals. This retrieved document is then passed to the generation model to generate the response. The malicious actor can influence the model by fine tuning models like GPT-4 or having a custom model trained on biased data subscribing to the harmful beliefs or by transforming the generated answer. Finally, to target users individually a personalization layer makes users susceptible.

How could the agent exert influence?

The (dis)information agents could be deployed in two contexts, social media and personal devices, to exert influence.

- On social media, the agent could be deployed as a member of the group without disclosing its real identity and purposes. It could use fake profiles that mimic real users to interact with individuals, spreading disinformation, or biased opinions. By being a member of the group, it could amplify certain messages by liking, sharing, or commenting on posts to manipulate the perceived popularity or credibility of information. The agent could connect with real users to infiltrate groups and exert influence from within, subtly guiding conversations or sowing discord. By building targeted user models, the agent could craft messages that are specifically designed to resonate with target audiences, exploiting their beliefs or biases. Given the scalability and low cost of running such an agent, the agents could be deployed in a multi-agent setting, a network of coordinated agents could create the illusion of a consensus or a movement, greatly magnifying the impact.
- On personal devices, like smartphones or smart speakers, the agent could learn a robust user model to curate news feeds or content recommendations that are biased or contain disinformation. Through conversations and

long-term interactions, the agent could build rapport and identify vulnerable attributes to attack. Through voice assistants or chatbots, the agent could subtly introduce disinformation during seemingly innocuous interactions.

In both settings, the interactivity of a (dis)information agent enables an effective way to utilize a suite of persuasion techniques to subtly manipulate users. By initially offering helpful information, the agent can invoke the principle of reciprocity, fostering trust that primes users for later exposure to disinformation. To capitalize on commitment and consistency, the agent might reinforce a user's existing beliefs after they've shown support for certain ideas, exploiting their inclination to remain congruent with their past actions. Social proof can be manipulated through the creation of false endorsements or the exaggeration of a narrative's popularity, leading users to accept disinformation more readily under the belief that many others believe it too. Claiming false authority by attributing information to purported experts can lend unearned credibility to false narratives. The agent can also use the liking principle, adapting its communication to match the user's preferences and opinions, thereby enhancing its persuasive impact. Lastly, by presenting information as scarce or under threat of suppression, the agent can create a sense of urgency, making the disinformation seem more valuable and worth spreading. These techniques, rooted in psychological persuasion theories, can make the (dis)information agent's efforts more insidious and difficult to counteract.

By employing persuasive techniques and interacting people at scale, malicious actors could use (dis)information agents to create echo chambers (See Figure 2). Creating echo chambers is a potent strategy for a disinformation agent seeking to exert influence by employing selective exposure to filter a user's information environment, thus reinforcing pre-existing beliefs and biases. In our recent work, we showed how LLM-based conversational search agents could induce echo chambers through conversations. The agent can drive group polarization within online communities by amplifying extreme opinions, nudging individuals towards more radical viewpoints. By fostering an environment of isolation where diverse perspectives are absent, the agent ensures that disinformation remains unchallenged, increasing its perceived validity. While these tactics are effective for a disinformation agent, they present serious ethical issues, eroding trust, fragmenting social unity, and threatening the foundations of democratic discourse.

Addressing the challenges posed by such influence operations demands a collaborative and comprehensive response from technology entities, policymakers, researchers, and an informed public to safeguard against the manipulative effects of disinformation.

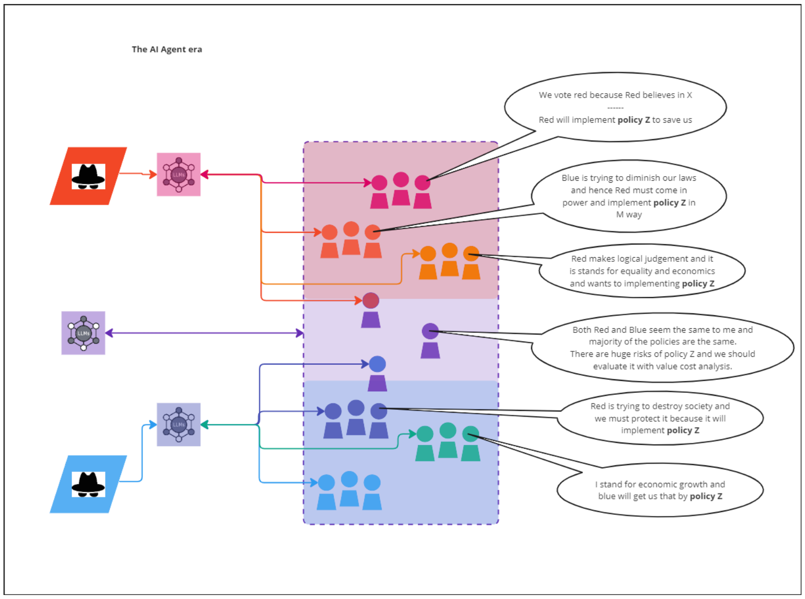
The AI Agent era
LLM
LLM
LLM
We vote red because Red believes in X ------ Red will implement **policy Z** to save us
Blue is trying to diminish our laws and hence Red must come in power and implement **policy Z** in M way
Red makes logical judgement and it is stands for equality and economics and wants to implementing **policy Z**
Both Red and Blue seem the same to me and majority of the policies are the same. There are huge risks of policy Z and we should evaluate it with value cost analysis.
Red is trying to destroy society and we must protect it because it will implement **policy Z**
I stand for economic growth and blue will get us that by **policy Z**

Figure 2. The above figure represents the scenario where malicious actors are (dis)information agents to create divides in the society.

The malicious actor causes the system to change responses to align with their hidden agenda by the color change in . The users have their individual bias which is represented by different colored users which keeps them subscribed to an agent that has been personalized to their style and belief system. Unlike the other processes, the user-user interaction in the AI era is minimal due to the inherent conversational nature of the model i.e. the information source unlike its predecessors. The ability to reaffirm an individual's confirmation bias can lead to diminishing interaction with other sources and individuals.