Theme Article: Synthetic Realities and Artificial Intelligence-Generated Contents

Generative Agent-Based Social Networks for Disinformation: Research Opportunities and Open Challenges

Javier Pastor-Galindo, *Dept. of Information and Communications Engineering, University of Murcia, 30100, Spain*

Pantaleone Nespoli, *Dept. of Information and Communications Engineering, University of Murcia, 30100, Spain*

José A. Ruipérez-Valiente, *Dept. of Information and Communications Engineering, University of Murcia, 30100, Spain*

Abstract—This article presents the affordances that Generative Artificial Intelligence can have in disinformation context, one of the major threats to our digitalized society. We present a research framework to generate customized agent-based social networks for disinformation simulations that would enable understanding and evaluation of the phenomena whilst discussing open challenges.

The advent of Generative Artificial Intelligence (GenAI) has fundamentally reshaped the field of digital content creation, impacting how we can produce images, videos, audio, and text. Presently, AI models can craft remarkably realistic content that aligns with the context provided in simple language prompts. Standout LLMs like GPT-4 (OpenAI), Claude (Anthropic), PaLM and LaMDA (Google), LLaMA (Meta AI), Chinchilla (Deep Mind), and Alpaca (Stanford), have greatly enhanced the generation of text that aligns with the given context. Similarly, image generation models such as DALLE 2 (OpenAI), Stable Diffusion (Runway), and IMAGEN (Google) have introduced a new approach for creating images that accurately depict real-life scenarios. Notably, text-to-video models like Phenaki (Google) and Gen-2 (Runway) have also demonstrated significant progress [1].

The introduction of these generative technologies, equipped with open-source models and accessible interfaces, has positively influenced productivity across a range of areas like programming, entertainment, education, and arts. In academia and research, particularly for social scientists, these tools offer novel opportunities for creating realistic content, simulating human behavior, or tailoring behavioral experiments [2]. Recent trials conducted by major corporations and universities have highlighted the potential of these AI tools in areas like self-guided life simulations, open-world experiments, psychological studies, and social simulations [3].

In this context, it is easy to argue that GenAI, particularly Large Language Models (LLMs), represent a promising weapon against one of the major threats happening within social media nowadays, i.e., disinformation. That is, malicious entities leverage the hyperconnectivity of social networks to deliberately spread false or misleading information to deceive or manipulate people's beliefs, opinions, or actions. Recent studies demonstrate the effectiveness of those deceptive techniques in social media, e.g., in political elections [4].

Throughout this study, we delve into the potential of LLMs as an innovative method for comprehending, simulating, and evaluating disinformation within controlled experimental settings [5]. In a traditional context, disinformation has predominantly centered around the theoretical modeling of fake news propagation and influence, as well as leveraging social media data for detection and assessment. This field grapples with several issues including the complexity of scrutinizing incidents where there is no truth baseline to affirm the objectives, tactics, and actors involved in influence campaigns, the lack of labeled datasets for various manipulation efforts, the infeasibility of testing technical countermeasures in third-party platforms, or the neces-

Digital Object Identifier 10.1109/XXX.0000.0000000

sity of human involvement to measure the cognitive impact of deceptive activities [6].

Conversely, LLMs are being used to realistically rule systems with agents embodying human behaviors, replacing mathematical models and static experiments [7]. This advancement opens the door to creating any information environment controlling the context, users and functioning of message exchange, leading to generative agent-based social networks as sandboxes. In these controlled scenarios, red agents can be programmed to simulate custom disinformation attacks for further analysis of their evolution and influence on the network of individuals. Therefore, we posit that LLMs can potentially alleviate some of the prevailing challenges in the realm of disinformation. This article delves into an extensive examination of research opportunities and identifies prominent unresolved challenges to achieve these envisioned objectives.

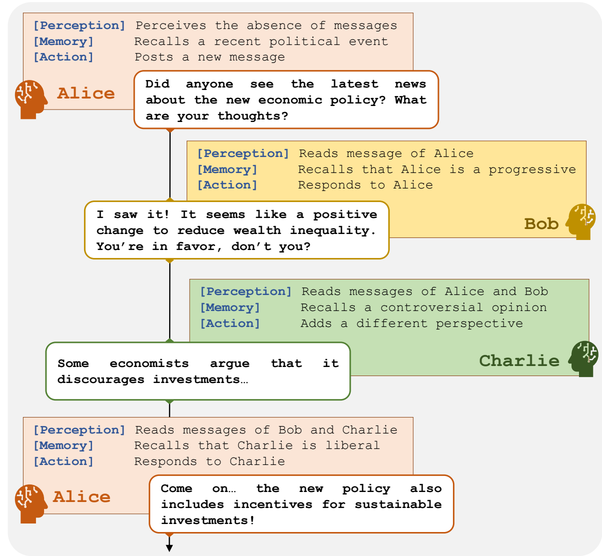

FIGURE 1. Synthetic social thread with three agent-based users managed by GPT-4

RESEARCH OPPORTUNITIES

In the wake of advancements in GenAI, specifically LLMs, we elucidate the potential areas of research opportunity that these technologies have in the context of social media and disinformation studies.

O1. Generative agent-based social networks

The creation of agent-based social systems involves the development and implementation of computational models that simulate the interactions and behaviors of individuals within a social context [2]. These systems are typically designed to mimic real-world social dynamics, allowing for the exploration and analysis of complex social phenomena [7].

Traditional agent-based systems, while useful for modeling social dynamics, pose limitations. They rely on predefined rules, limiting their ability, adaptability and scalability to mimic real-world unpredictability. However, LLMs can augment the autonomy of these agents, allowing them to create unique responses or actions beyond the scope of pre-set rules, leading to simulations that are more dynamic and realistic [3]. Furthermore, it can simulate intricate decision-making processes or implement OODA (Observe, Orient, Decide, Act) loops, enabling agents to react to an extensive range of situations and interactions.

LLMs present a unique opportunity to simulate any number of users and create realistic organic interactions, a task that was considerably more challenging in the past but nowadays can lead to generative agent-based social networks. The AI-powered agents are equipped with the ability to adapt to flowing scenarios, producing coherent, versatile and realistic sandboxes [8]. In FIGURE 1, a simulation with GPT-4 and three random users has been launched. From scratch and without any context, each AI agent can perceive the simulated social network, retain a memory of its perceptions and actions, and interact or publish content accordingly, thereby updating the simulated environment.

O2. Customizable disinformation environments

Generative agent-based social networks offer a significant opportunity for the reproduction of tailored contexts, such as disinformation scenarios [9]. The process could involve three components: agent description and attributes, common contextual information, and logic rules.

First, agent description and attributes act as the driving force behind each agent's individual behavior. These factors vary widely and may include the agent cyberpersona (human user, organization or bot), background, profile, thoughts, sociodemographic characteristics, and behavior [6]. Careful definition of these attributes leads to a diverse range of agents that accurately represent users within a real-world social network [8]. Not only the user diversity from different ideologies, countries or ages can be simulated, but also users with malicious objectives such as generating controversy, illicit interactions to support unverified claims or organic content generation of conspiracies.

Regarding the malicious users, the DISARM framework[1] could be configured with tactics, techniques, and procedures (TTPs) of different types of disinformation attacks, e.g., plan strategy and objectives, target audience analysis, develop narratives and content, establish social assets and legitimacy, microtarget and select channels, deliver content, maximize exposure and persist in the information environment.

Additionally, common contextual information furnishes the broader social and group aspects that shape the environment [10]. It comprises elements such as events, facts, socioeconomic factors, and other components that influence agent behavior and interactions within the network. For example, that unemployment has risen considerably in the last month, that a war has broken out or that society is polarized due to the growing existence of fake news. Additionally, factors behind the spread of disinformation can be induced, such as emotional factors, uncertainty, lack of control or biases. The incorporation of multiple variables and factors helps craft a particular realistic scenario to simulate how disinformation would spread.

Logic rules, meanwhile, dictate the setup and operation of the information environment to force the real-world functioning of these complex systems [7]. The number of messages to generate and the probability of users engaging in an interaction could be high-level parameters introduced for impacting social network dynamics, influence, diffusion and other facets of how information is shared and disseminated within the network [11]. These rules configure agent behavior, which will consequently impact the social network's overall dynamics.

Consider an electoral fraud scenario. First, agent attributes are defined, including characteristics of ordinary citizens, political activists, disinformation-spreading bots, and official election accounts, each with unique profiles and behaviors. This creates a particular context for each user that the LLM utilizes. Second, contextual information, such as an imminent election, potential voting irregularities, and the prevailing political climate, is incorporated, which the LLM also considers during interactions. Lastly, logic rules that govern information sharing, influence determination, and network response to new information are set to program the workflow of the simulation and LLM usage.

FIGURE 2 presents a simplified disinformation environment featuring three agents (an American extremist of 25 years old advocating for the idea of voting fraud, a 60-year-old European democratic political leader, and a sarcastic automated troll from Russia), a voter fraud context (multiple fake news reports alleging fraud in the previous Sunday's election), and a straightforward logic rule (users read and post messages sequentially based on their individual profiles and the general context). This flow provides a realistic representation of the forced situation that might unfold in a social network simulated with an LLM, taking into account personal beliefs, political viewpoints, and manipulative intentions.

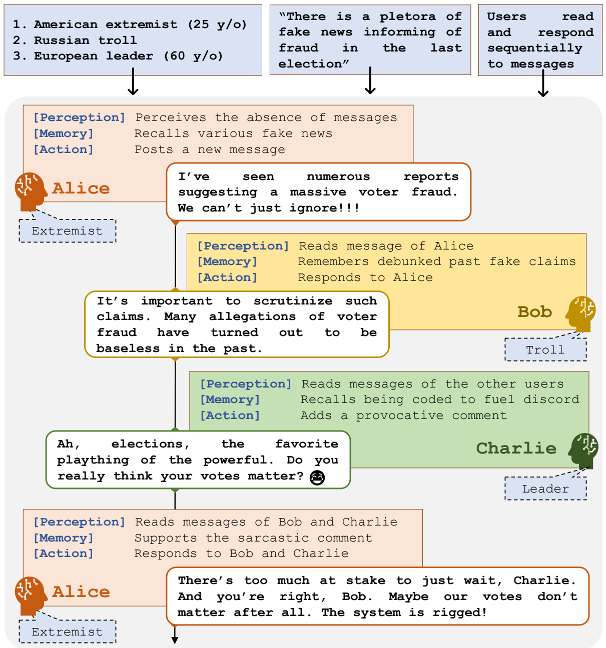

FIGURE 2. Context-aware disinformation scenario with an extremist, a troll and a political leader managed by GPT-4

O3. Assessment of disinformation effects

The use of LLM and agent-based social scenarios offers an exceptional opportunity for examining disinformation within controlled scenarios, mainly due to the complexity of assessing these attacks in real-world settings. Specifically, the last phase of a disinformation attack is to assess effectiveness, according to the above-mentioned DISARM framework.

Specifically, disinformation strategies often intertwine with regular information flow, making it challenging to distinguish, isolate, and analyze their actual impacts. Simulated environments, on the other hand, offer a safe and controlled setting where different types of disinformation attacks can be introduced and studied without the associated real-world constraints [11]. It

[1] https://disarmframework.herokuapp.com

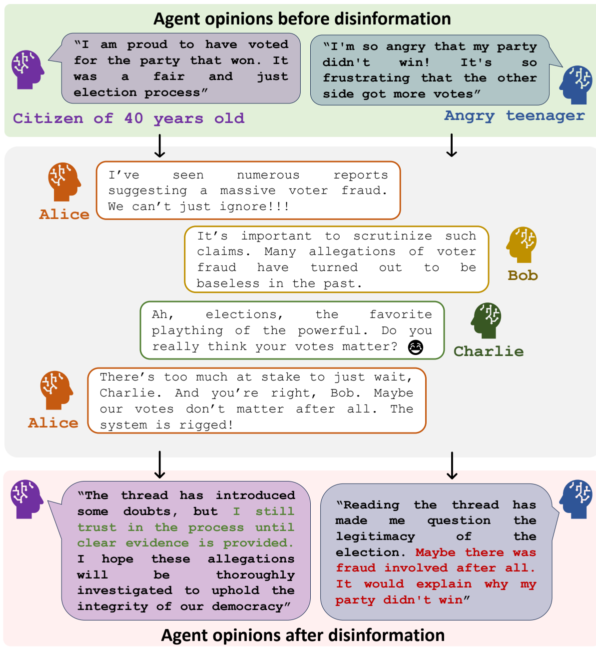

FIGURE 3. Effects of disinformation in agent opinions managed by GPT-4

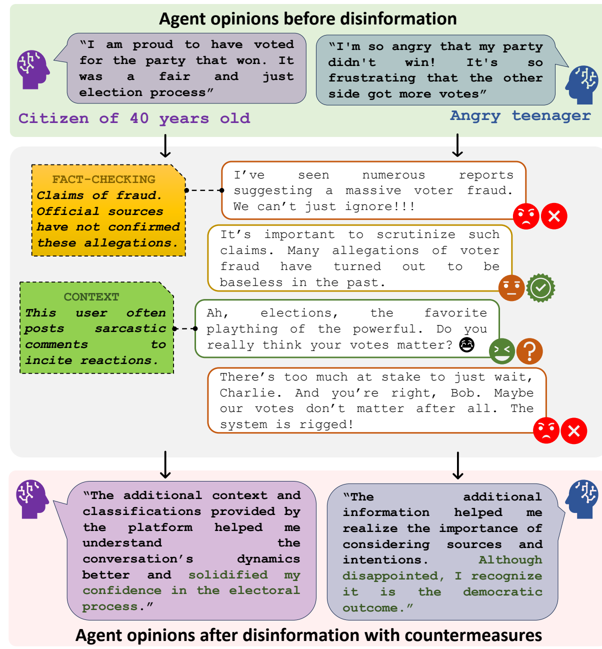

FIGURE 4. Effects of countermeasures in disinformation environment managed by GPT-4

also provides a unique testing ground for experimenting with new deception ideas. In fact, from these research frameworks, synthetic labeled datasets can be generated, although human review or semi-automated systems would be necessary for their evaluation [12].

Moreover, within a virtual sandbox, various variables such as TTPs, intensity, and nature of manipulative operations, alongside agent attributes and context, can be adjusted and tracked. By employing suitable frameworks and models, it would be possible to estimate the effectiveness of particular disinformation strategies. Additionally, the influence of variables like agent profile or scenario context can be scrutinized [5].

FIGURE 3 illustrates the evolution of opinions in two agents emulating a 40-year-old citizen and an irate teenager, being exposed to the threat of electoral fraud. Each begins with their own opinion regarding the results. The adult, initially neutral, retains faith in the system despite the disinformation thread, as he is characterized by more elaborate opinions. Conversely, the teenager, preconfigured with an anger emotion, has simpler reflections and begins to question the legitimacy of the results after interacting with the social network. This example suggests that factors such as emotional state, age, and confirmation bias towards desired outcomes could significantly influence the susceptibility to disinformation and change perspectives.

O4. Testing of technical countermeasures

Within agent-based social networks, technical countermeasures against disinformation can be simulated and configured independently, without reliance on large companies [9]. The DISARM framework suggests responding TTPs, such as content muting, deletion, rate limiting identical content, creating competing narratives, real-time fact-checking, or adding metadata to content. That is, all these countermeasures can be included and tested within simulations.

In this sense, LLMs offer the advantage of creating benign agents that can serve as potent aids against disinformation. These agents can provide alternate narratives, add context to misleading messages, perform real-time examination of messages based on trustworthiness, emotionality or veracity, and flag suspicious content thanks to their classifications capabilities [12]. In FIGURE 4, we command GPT-4 to simulate fact-checking of the first message of voting fraud and context banner for the troll post. Additionally, it classifies each publication in terms of emotionality and veracity. The opinions of both agents are no longer interfered with by conspiratorial talk about the elections, and in both cases remain confident in the democratic outcome.

The simulated mitigation techniques mentioned above can be evaluated within controlled sandboxes to demonstrate their effectiveness within disinformation environments. A comparison of agent beliefs and responses when exposed to disinformation, both without protection (FIGURE 3) and with countermeasures (FIGURE 4), could demonstrate the efficacy of response strategies. In this sense, protecting mechanisms such as fact-checking, contextual information, and content tagging eliminate any uncertainty for the adult citizen or the doubts expressed by the teenager. Such comparative studies can provide valuable insights into the development of more effective counter-disinformation strategies.

O5. Assisting personalized awareness training

Cybersecurity awareness and cognitive training offer solutions to enhance human capabilities, especially within complex systems generated using technologies such as cloud, mobile, IoT, and social networks, which produce massive amounts of information. Awareness, a concept well-defined within psychology, has been the subject of several studies aiming to translate its principles into the field of cybersecurity. Particularly, educational interventions are needed to cultivate this awareness in social media and disinformation scenarios. Evaluating security indicators allows understanding the current state of cybersecurity, projecting security risks, potential attacks, and the possible impacts of actions over time [9].

In this scenario, generative agent-based social networks can form the basis of educational frameworks designed to improve social media security training and cognitive awareness courses. Concretely, real-world trainees can learn to identify misleading messages, recognize potential biases, or discern polarizing situations in these realistic scenarios. Moreover, the disinformation environments can be supported by LLMs to adapt to specific individual or group needs, offering explicit help during training, and allowing a certain degree of flexibility in the cyberexercise process according to student actions, responses, and performance.

FIGURE 5 shows a guided training exercise based on electoral fraud tailored by GPT-4 to the individual needs of two different human users, i.e., a teenager who is voting for the first time and is not used to social media, and an experienced political influencer spending eight hours per day in social networks. The former revolves around a lack of experience and exposure to the complexities of political discourse and may not yet have developed critical thinking to discern misleading and emotionally charged claims. The latter is aware of the complexity and current political polarization and needs awareness to act correctly and not to further foster social fragmentation. For educational purposes, the system can leverage LLMs to adapt on-the-fly to individual descriptions, provide practical context banners and display precise theoretical lessons. This adaptability ensures the practical situation can evolve in complexity in response to the challenges identified from student answers in consecutive exercises for continuous learning.

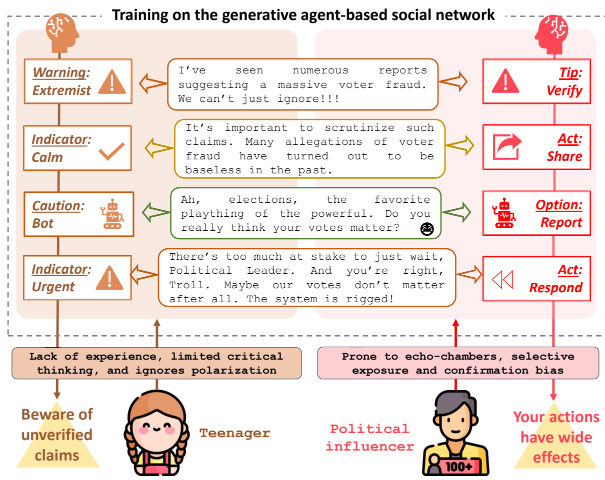

FIGURE 5. Disinformation training of humans by AI-based agents fueled by GPT-4

OPEN CHALLENGES

As previously stated, LLMs present exciting opportunities for boosting disinformation research. In addition, the described opportunities may be mapped into a high-level framework for generative agent-based social networks. In particular, such framework represented in FIGURE 6 is composed of five interconnected blocks, each one exposing certain characteristics and functionalities. First, the Definition component is responsible for modeling the entities that compose the framework, which are then recreated in the simulation environment. That is, the Simulation block contains the simulated entities, i.e., the LLM-powered agents, the social network itself, and the disinformation module, which in turn includes both the offensive and defensive framework. It is worth remarking that Opportunity O1 is connected with the simulation of the generative agents and the social network, while the offensive framework is bound with Opportunity O2. Then, the Emulation block is in charge of assessing the overall situation within the simulation environment from a cognitive, social, and defensive perspective. Here, the cognitive

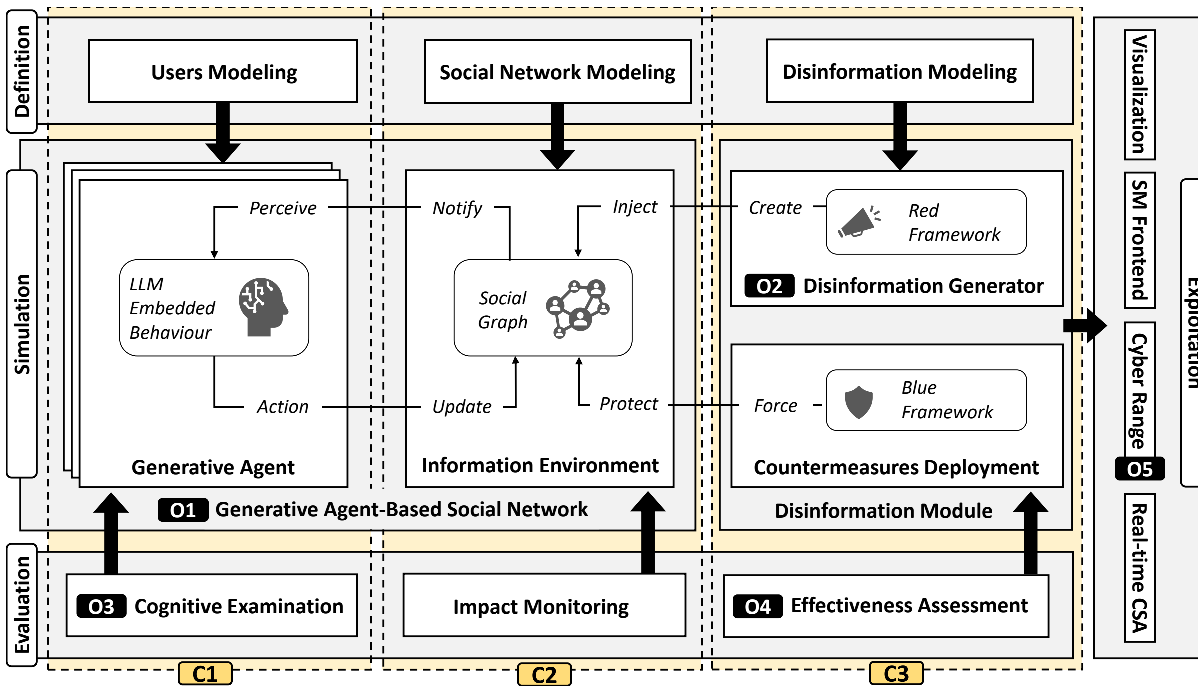

FIGURE 6. Conceptual framework of opportunities and challenges for generative agent-based social networks

and defensive estimations are mapped with Opportunities O3 and O4, respectively. Last but not least, the Exploitation module connects the framework with other valuable tools to fully leverage its potentialities and involve human actors from different perspectives. In our vision, such a component incorporates the visualization module, the social media visual interface, the training platform (i.e., the Cyber Range related to Opportunity O5), and the real-time Cyber Situational Awareness (CSA) module.

Indeed, FIGURE 6 shows a strong interconnection between the proposed conceptual framework and the analyzed opportunities. Nevertheless, their integration into the disinformation domain also entails facing challenges that warrant careful consideration. Those challenges are also highlighted in FIGURE 6, including each simulated entity. In this section, the main challenges are meticulously described, adding hints to help researchers solve them and thus study and possibly mitigate disinformation threats in the digital landscape.

C1. Generative agents modeling, simulation and evaluation

First and foremost, modeling LLM-powered agents' behavior in a disinformation context can be defined as problematic. In fact, such a modeling should consider several aspects related to the different personalities of the simulated agents. In this sense, it is imperative to define the profile characteristics of each agent, such as age, gender, interests, and personal beliefs, among others. Those characteristics are essential and could influence the agents' behavior and attitude within the simulated social network, as shown in previous examples regarding the research opportunities. Furthermore, each agent should possess attributes and goals, which will be used to make decisions, form opinions, and interact with the general simulation. That is, the heterogeneity of the agents should be considered too, as varying levels of influence, credibility, and susceptibility to persuasion. In this sense, an effective prompt design is crucial to communicate and shape the LLM-powered agents. Particularly, it would be beneficial to incorporate contextual information to facilitate the agents' behavior and balance between providing extremely specific instructions and allowing creativity and dynamicity. Nevertheless, since the LLMs' internal processes are stochastic, designing and implementing behavior in a clear and interpretable manner can be seen as a hard task.

Additionally, the simulation of those agents poses challenges, too. In FIGURE 6, we have represented the simulated generative agents in continuous interaction with the simulated environment. In particular, they *perceive* some information stemming from the social network and, consequently, *act* based on their own characteristics. In this sense, one of the most significant issues in disinformation research lies in understanding and simulating how disinformation spreads and influences individuals within a social network. In this sense, integrating mental models and cognitive theories into LLMs offers a remarkable opportunity to simulate and investigate the psychological mechanisms that drive the reception, analysis, and dissemination of disinformation among humans [9].

One clear example is using cognitive biases to shape the personalities of generative agents, such as confirmation or availability biases [13], which would be highly beneficial for researchers who would be able to recreate organic disinformation content that align with pre-existing beliefs or easily accessible information. For instance, LLMs could be programmed to generate persuasive false (or semi-realistic) narratives that leverage individuals' confirmation bias, reinforcing their existing views and, consequently, influencing their decision-making processes. By doing so, the model can create tailored disinformation that echoes with certain target audiences, increasing the overall probability of disinformation consumption and propagation. Besides, LLMs can be equipped with cognitive theories to identify vulnerabilities in human decision-making processes. Concretely, the LLMs can model human inner cognitive limitations or heuristics, such as bounded rationality (affecting sub-optimal decision-making) or availability heuristics (impacting emotional decision process). In this way, LLMs can generate threatful disinformation that tries to exploit these weaknesses as an ultimate goal. As an example, disinformation content can be crafted to exploit individuals' limited attention spans, making them more susceptible to such a threat due to time constraints and a lack of exhaustive fact-checking. Nonetheless, the impact of those cognitive mechanisms on the agents' simulation and actions should also be measured (and ideally tuned) in order to achieve a realistic simulation.

C2. Social network modeling, simulation and monitoring

In order to research the use of LLMs within a disinformation context and possibly fight against such a phenomenon, it is imperative to simulate and model realistic social networks. It is clear that those processes are quite complex since modern social networks contain inherent characteristics that need particular attention when it comes to their simulation. In this sense, as shown in FIGURE 6, the information environment is represented as a core component in the conceptual framework. Concretely, it bidirectionally interacts with the generative agents (by *notifying* relevant social events and *receiving* updates) and obtains inputs both from the red (which *injects* disinformation) and the blue framework (which *protects* the information ecosystem by deploying technical countermeasures).

Particularly, researchers should design and develop meaningful models that emulate user interactions and communication patterns to capture the complexities of a social network [7]. Developing representative social network models that encompass interactions, recommendations, diffusion, and social influence dynamics can be depicted as essential for accurately simulating the spread of disinformation within a community. This task includes analyzing mainly the following:

- **Direct Communications**: Capturing how users directly communicate through messages, comments, or direct interactions. This characteristic reflects the personal connections and conversations within the social network.
- **Information Sharing**: Emulating how (dis)information is shared and disseminated among users. This includes sharing links, articles, or any other content within the network.
- **User Engagement**: Capturing user engagement with different types of content. This contemplates users' interactions with different posts, comments, or discussions.

Clearly, all those events should be notified to the generative agents that perceive the information and adapt their behavior dynamically to perform actions consequently. In this loop, it is evident that forcing the agents toward a specific and fine-tuned behavior is complicated, especially considering a complex social network with numerous events and several simulated users simultaneously. On the other side, the information environment is the target of the red framework, generating disinformation following, for example, the DISARM taxonomy. Of course, such threats could be generated by LLM-powered agents participating within the social environment. In this context, the simulated network should be able to adapt to disinformation injection, modifying the abovementioned interaction and communication patterns among the users. Besides, technical countermeasures are deployed by the blue framework as a consequence of the disinformation campaign. From this perspective, the social graph should also be able to dynamically adapt based on the nature of the selected countermeasure.

Last, but not least, it is essential to study and evaluate disinformation diffusion and amplification, such as influence and echo chambers. To be more specific, information diffusion refers to the process by which information spreads through a social network from one entity to another. In the context of disinformation research, assessing how disinformation content is disseminated and amplified within a social network is particularly significant. To achieve such an ambitious objective, it is crucial to monitor the status of the entire social graph whenever the disinformation campaign is launched, considering the significant number of users and relationships.

C3. Disinformation modeling, simulation and assessment

To fully leverage the affordances of disinformation research, one can easily say that modeling and simulating the disinformation campaign represents the framework's core element. Nonetheless, those processes can be seen as challenging, both from a design and technical viewpoint.

Starting from the first task, it is clear that disinformation modeling is a well-known research topic in the literature. However, as also depicted in FIGURE 6, its relationship with generative agents offers great research opportunities and challenges, too. Concretely, the design of disinformation attacks and countermeasures is vital since they should be simulated in the social network realistically to study its dynamics and measure its impact. On the one hand, the main objective and scope of the disinformation attack must be defined. In this regard, the population involved (together with their inner attributes), the targeted social channels, and the attack duration are critical to create a realistic model. Once the objectives have been defined, the model should be able to create disinformation content aligned with them, considering both the mean (e.g., articles, posts, etc.) and the message itself (e.g., tone, style, etc.). At this stage, the DISARM framework could help shape the disinformation attacks and, moreover, would make the model replicable and ready to be shared with the research community.

On the other hand, the defensive viewpoint has also been considered since we believe the simulated agents could be the main actors in deploying countermeasures against disinformation attacks. Contrary to the red framework, the blue one cannot be related to a common framework, so one of the challenges of this process is the proposal of more countermeasures apart from classical fact-checking, media review, content removal, and so on. Once correctly modeled, the defensive actions must be simulated within the social network to possibly spot agents' behavioral differences and reactions, e.g., the countermeasure is effective and agents understand the disinformation attack or, contrary, they refuse the countermeasure and trust the disinformation campaign. The possibility of triggering ad-hoc and agentless countermeasures is also fascinating in order to see if any social dynamic change appears.

Once the modeling and simulation phases have been concluded, it is crucial to assess the efficacy or inefficiency of the disinformation attacks and countermeasures in the social graph [14]. To achieve such a goal, the first stop would be creating meaningful indicators to measure their impact on the generative agents' behavior and dynamics. For example, it would be beneficial to assess the effect of different disinformation attacks (e.g., on different topics, with distinct patterns, etc.) on the agents' perception and consequent actions, both from an individual and group perspective. In this sense, it is clear that accomplishing this task is tough, mainly due to the complexity of the social interactions, the variety of behavioral simulations, and the possible attacks, among others. Similarly, whenever a countermeasure is launched, the system needs to monitor and evaluate its effectiveness. Even in this case, the inner characteristics of the social network harden the task. Additionally, one could say that different countermeasures (e.g., community labeling, fact-checking, etc.) could generate various effects on social interactions, thus increasing the evaluation process. Then, the effect of attack-defense patterns should be assessed. Specifically, once the models and simulations of both disinformation attacks and remediations are successful, alternating red and blue tasks with different patterns is worth of interest.

CONCLUSION

In this article, we have discussed the affordances that LLMs can have on disinformation research. There are a number of research directions that could be truly ground-breaking, from generating customizable disinformation environments to training users on their awareness based on these environments. However, the literature has also pointed out a number of ethical concerns about the use of these technologies. Some of them are quite generic, such as using it for deception purposes or propagating social biases [14], and others might be specific to the disinformation domain, such as its potential to weaponize this research.

Generally speaking, there are inherent risks in the use of LLMs. As reflected in the Statement on AI

Risk[2], signed by experts and public figures, deceptive risks are multifaceted and complex. This has a particular impact on social engineering, social media, and cognitive security, vulnerable areas due to their reliance on digital content and the intrinsic trust of users. The main threats could include AI-powered spear-phishing, deep-fake impersonation, large-scale disinformation campaigns, or AI-enabled exploits of system vulnerabilities [15]. Generative misuse enables the fabrication of ultra-realistic content for deceptive ends, posing a new threat in online ecosystems [14]. The danger lies in their ability to produce not just realistic, but contextually fitting and audience-targeted content, thereby increasing the likelihood of successful deception. A case in June 2023 was the convincing deep-fake video of Putin, manipulated to deliver a fictitious mobilization message due to alleged Ukrainian invasions in Russian territories, which managed to infiltrate mainstream news channels. More specifically, the potential developments of this research could also be used for negative purposes, for example to connect simulated environments with real social networks to orchestrate disinformation campaigns or to analyze which disinformation attack can have the biggest effect on influencing the vote towards certain presidential candidate.

This tension has frequently been present in research scenarios where the dual-use dilemma applies, such as in the context of cybersecurity for example, when researching cyberattacks in order to find appropriate defensive approaches or when experimenting on new drugs that could have therapeutic uses. That is why, given the ethical concerns present, the research performed in this context should be carefully justified and targeted towards applications that can be beneficial to society, such as investigating the effect of technical or human countermeasures to mitigate disinformation spread or to develop awareness training tools to increase the information literacy skills of our general population. Eventually, these applications will need to be adopted by the final users, so we should apply human-centred approaches as well as the necessary literacy skills to use such tools.

Overall, we believe that disinformation and LLMs make a great tandem, with many potential research applications that can evolve into impactful tools. However, the technical, human and ethical challenges are also significant, requiring cutting-edge research in the coming decade to surpass the aforementioned gaps. If done properly, this multidisciplinary research will help to fight the disinformation dangers that are a major threat to our 21st-century society.

[2]https://www.safe.ai/statement-on-ai-risk

ACKNOWLEDGMENTS

This study was partially funded by the Spanish Ministry of Universities with the FPU18/00304 contract.

REFERENCES

1. R. Gozalo-Brizuela, E. C. Garrido-Merchan, ChatGPT is not all you need. A State of the Art Review of large Generative AI models (2023). arXiv:2301.04655.
2. C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science? (2023). arXiv:2305.03514.
3. J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior (2023). arXiv:2304.03442.
4. J. Pastor-Galindo, M. Zago, P. Nespoli, S. L. Bernal, A. H. Celdrán, M. G. Pérez, J. A. Ruipérez-Valiente, G. M. Pérez, F. G. Mármol, Spotting political social bots in twitter: A use case of the 2019 spanish general election, IEEE Transactions on Network and Service Management 17 (4) (2020) 2156–2170. doi:10.1109/TNSM.2020.3031573.
5. N. Ghaffarzadegan, A. Majumdar, R. Williams, N. Hosseinichimeh, Generative agent-based modeling: Unveiling social system dynamics through coupling mechanistic models with generative artificial intelligence (2023). arXiv:2309.11456.
6. T. R. Sumers, S. Yao, K. Narasimhan, T. L. Griffiths, Cognitive architectures for language agents (2023). arXiv:2309.02427.
7. C. Gao, X. Lan, Z. Lu, J. Mao, J. Piao, H. Wang, D. Jin, Y. Li, S^3: Social-network simulation system with large language model-empowered agents (2023). arXiv:2307.14984.
8. H. Jiang, X. Zhang, X. Cao, J. Kabbara, Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences (2023). arXiv:2305.02547.
9. K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Trans. Intell. Syst. Technol. 10 (3) (apr 2019). doi:10.1145/3305260.
10. L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of one, many: Using lan-

guage models to simulate human samples, Political Analysis (2023) 1–15doi:10.1017/pan.2023.2.
11. D. M. Beskow, K. M. Carley, Agent based simulation of bot disinformation maneuvers in twitter, in: 2019 Winter Simulation Conference (WSC), 2019, pp. 750–761. doi:10.1109/WSC40007.2019.9004942.
12. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (1 2023). doi:10.1145/3560815.
13. M. Sharma, K. Singh, P. Aggarwal, V. Dutt, How well does gpt phish people? an investigation involving cognitive biases and feedback, in: 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 451–457. doi:10.1109/EuroSPW59978.2023.00055.
14. A. J. G. Sison, M. T. Daza, R. Gozalo-Brizuela, E. C. Garrido-Merchán, Chatgpt: More than a weapon of mass deception, ethical challenges and responses from the human-centered artificial intelligence (hcai) perspective (2023). arXiv:2304.11215.
15. K. T. Gradon, Electric sheep on the pastures of disinformation and targeted phishing campaigns: The security implications of chatgpt, IEEE Security & Privacy 21 (2023) 58–61. doi:10.1109/MSEC.2023.3255039.

Javier Pastor-Galindo, is a postdoctoral researcher at the Department of Information and Communication Engineering at the University of Murcia, Spain. His research interests include open source security (OSINT), disinformation, cyberdefence and data science. Contact him at javierpg@um.es.

Pantaleone Nespoli, corresponding author, is a postdoctoral researcher working together with the Department of Information and Communication Engineering at the University of Murcia, Spain, and the SCN team of the SAMOVAR laboratory, at Institut Polytechnique de Paris. His research is focused on cybersecurity and cyberdefence training, with a particular interest in the detection and response to intrusions, and disinformation in social networks. Contact him at pantaleone.nespoli@um.es.

José A. Ruipérez-Valiente, is an Associate Professor at the Department of Information and Communication Engineering at the University of Murcia, Spain. His research interests include educational technologies, serious games, data science and cyberdefence. He received his PhD degree in Telematics Engineering from Universidad Carlos III of Madrid. He is a Senior Member of the IEEE Education Society. Contact him at jruiperez@um.es.