Project Proposal

Benjamin Thomas and Amy Chang

November 6 2024

1 AI Agents Gone Rogue?

- **Objectives**: It is generally acknowledged that the social behavior of LLM-enabled agents is not human-like in at least some important ways, especially in complex environments, multi-shot settings, or interactions without human direction [1]. However, while the existence of this problem is uncontroversial, it is difficult to chart out the problem space since most results are qualitative and require manual labeling, usually into fairly simple result spaces. Given that social behavior is complex and subtle, this makes it difficult to provide analysis that is both rigorous and meaningful. In order to facilitate analysis of LLM-agent social behavior, we plan to simulate an arbitrarily large number of games of Dungeons and Dragons using LLM-enabled agents to obtain discrete data in a complex result space about social behavior in a truly "open" social environment. Using this dataset, we plan to test how the agents' behavior changes based on 1) individual traits (i.e. the survey responses used to generate the agent) and 2) the social environment (i.e. the survey responses of the other agents in the party). Depending on the nature of available datasets, we would also like to compare this to the behavior of human Dungeons and Dragons players. We believe this field experiment analog will provide rigorous and meaningful results about LLM-enabled agent social behavior that are not available in more "closed" simulations.
- **Background Information**: First, we ought to consider whether our endeavor is well-motivated: that is, do LLM-enabled agents behave in human-like ways? This does not admit a clear yes or no. LLM-enabled agents perform well at some human-mimicry tasks, like Turing Test-like interviews [2] and even display some response biases which appear human-like [3], but very poorly at others, such as Dictator Games [4] or elections [5]. However, there is a lack of research on what underlying factors divide social tasks where LLM-enabled agents fail and succeed, and LLMs seem to be incapable of solving these problems on their own through evolutionary or recursive improvement [6]. Given the shocking disparity between LLM-agents skill at different tasks, this clearly warrants more rigorous investigation.

  Second, we ought to consider whether constructing our dataset is reasonable or feasible. We are not the first to ask whether LLM-agents are able to play Dungeons and Dragons.

Previous research has identified that ChatGPT, given a sufficiently good prompt, is a fairly effective Dungeon Master [7] [8]. Less research has been done into LLM-enabled agents playing Dungeons and Dragons since it is unclear why anybody would want that, but I encourage you to watch the following very entertaining video as a proof-of-concept [9]. Thus, we have strong reason to believe that we can construct this dataset with enough CPU-hours. Finally, we ought to consider whether this dataset will provide useful results. In Ben's own research (unpublished as of yet!), Ben has found that ChatGPT roleplaying as an agent resolves many decision questions by reference to a (near-)one or two dimensional heuristic vector through discrete prompt modification. Dungeons and Dragons offers a fundamentally similar decision problem, so I would expect we can recover a similar pattern with sufficient data. Other research [10] has also suggested the importance of heuristics in ChatGPT's decision making. This means the fundamental stochastically of LLMs should not be a barrier to peering into the peculiarities of LLM-enabled agents' social behavior.

- **Hypotheses**: We expect LLM-enabled agents to engage in social behavior which is highly different from human-players. We do not expect these differences to appear random, but to emerge from three fundamental inhuman decision biases, which we aim to test. First, we expect the agents to be unreasonably pro-social [5]. This should manifest in under-utilization of explicitly anti-social abilities, such as deception. This also should result in the qualitative result that agents largely are gullible, since they are insufficiently cautious about other actors having bad intentions. Second, we expect agents to inappropriately overgeneralize from their prompts, since they cannot reason about when the usage of prompts is contextually appropriate [11]. For example, agents with survey responses mentioning exercise may use their athletics ability more often than is justified. Third, we expect agents to engage in non-subgame perfect problem solving. This is a slightly idiosyncratic prediction for which I could not find a basis in the literature, but want to test nonetheless. This hypothesis will be more difficult to test and will require some manual data-labeling but will most likely manifest as a lack of violence, since most social Nash equilibria have subgames which require seemingly unnecessary use of violence as "punishment" or "deterrence" [12]. Finally, we want to test an ancillary hypothesis about herding. Since pro-social behavior manifests as aggressive consensus seeking, we expect agents to suppress their unique traits among other agents which do not share them, but to express them fully among other agents that share the traits.

- **Methods Overview**: Our simulation works as follows. First, we make agents from the CS222 agent bank. Then, second, we randomly partition them into parties of three. Third, each agent independently creates three characters. We will limit technical customization to some extent to avoid issues with our infrastructure, but we will give complete freedom over the backstory of the characters and some high-level discrete traits, like class, to the agents. Fourth, the agents collectively decide which of their characters to use for the campaign. Then, the campaign begins, and they repeat taking simultaneous turns until either all players die or the agents win. A turn is partitioned into a discussion phase, an action declaration phase, and a results phase,

where the DM chooses what roles are needed for actions and interprets the results. Finally, we will interview all of the agents about how they feel about the campaign and the other agents.We will use a to-be-decided pre-made campaign. We will then randomly partition agents again and repeat. We will harvest the following data: 1) traits of character chosen, 2) ability rolls made, 3) campaign outcome, 4) whether player survives, 5) interview results of agents after the campaign. All data will be mediated by the traits of their character and the other agents in the party.

- **Phenomena Modeled**: This section will be somewhat limited, since a major element of the design of our experiment is that the agents are free to do more or less what they want in their Dungeons and Dragons world, which inherently allows infinite phenomena. This means this section will be much less discrete than other groups, but this is an inherent trade-off from our "field experiment-like" design. However, there are some general phenomena categories which the design of the simulation will enforce on the agents. First, agents must collectively strategize and make collective decisions towards a collective goal. This happens in a short-term way each turn and in a long-term way in the initial decision which character to play. This may cause conflict over ethical, strategic, or personal considerations, but we cannot be sure. We also expect some level of creative self-expression based on the agent background in the creative design of character, prior to picking. Finally, the agents must engage in self-directed decision making towards a well-defined goal without a well-defined path to the goal. Hopefully, we will see other interesting phenomena, but no other phenomena are structurally guaranteed, so we have to take a wait-and-see or a recklessly-speculate approach.

References

[1] Y. Lu, A. Aleta, C. Du, L. Shi, and Y. Moreno, "Llms and generative agent-based models for complex systems research," *Physics Reports*, vol. 1041, pp. 10–13, 2024. DOI: 10.1016/j.plrev.2024.10.013. [Online]. Available: https://doi.org/10.1016/j.plrev.2024.10.013.

[2] A. Tikhonov and I. P. Yamshchikov, "Post turing: Mapping the landscape of llm evaluation," *arXiv preprint arXiv:2311.02049*, 2023, Accepted for GEM @ EMNLP 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.02049.

[3] L. Tjuatja, V. Chen, T. Wu, A. Talwalkar, and G. Neubig, "Do llms exhibit human-like response biases? a case study in survey design," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1011–1026, 2024. DOI: 10.1162/tacl_a_00685. [Online]. Available: https://doi.org/10.1162/tacl_a_00685.

[4] J. Ma, "Can machines think like humans? a behavioral evaluation of llm-agents in dictator games," *arXiv preprint arXiv:2410.21359*, 2024, Version 1, October 28, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.21359.

[5] X. Liu, J. Zhang, S. Guo, H. Shang, C. Yang, and Q. Zhu, "Exploring prosocial irrationality for llm agents: A social cognition view," *arXiv preprint arXiv:2405.14744*, 2024, Version 2, September 27, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2405.14744.

[6] Y. Chen, Y. Liu, J. Yan, *et al.*, "See what llms cannot answer: A self-challenge framework for uncovering llm weaknesses," *arXiv preprint arXiv:2408.08978*, 2024, Version 2, August 29, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.08978.

[7] P. Sakellaridis, "Exploring the potential of llm-based agents as dungeon masters in tabletop role-playing games," Master's thesis, Institution Name, 2024.

[8] X. You, P. Taveekitworachai, S. Chen, *et al.*, "Dungeons, dragons, and emotions: A preliminary study of player sentiment in llm-driven ttrpgs," in *FDG '24: Proceedings of the 19th International Conference on the Foundations of Digital Games*, Association for Computing Machinery, 2024, pp. 1–4. DOI: 10.1145/3649921.3656991. [Online]. Available: https://doi.org/10.1145/3649921.3656991.

[9] DougDougDougDoug, *Can 3 ai survive my d&d roguelike campaign (again) (vod)*, YouTube video, Recorded on July 18, 2024, 2024. [Online]. Available: https://www.youtube.com/watch?v=DNCA8riO-Yc.

[10] G. Suri, L. R. Slater, A. Ziaee, and M. Nguyen, "Do large language models show decision heuristics similar to humans? a case study using gpt-3.5," *Journal of Experimental Psychology: General*, vol. 153, no. 4, pp. 1066–1075, 2024. DOI: 10.1037/xge0001547. [Online]. Available: https://doi.org/10.1037/xge0001547.

[11] S. Goldstein, *Llms can never be ideally rational*, PhilArchive, University of Hong Kong, 2024. [Online]. Available: https://philarchive.org/rec/GOLLCN.

[12] M. Hoffman and E. Yoeli, *Hidden Games: The Surprising Power of Game Theory to Explain Irrational Human Behavior*. New York: Basic Books, Apr. 2022, ISBN: 9781541674576.