# Advancing Consensus: Automated Persuasion Networks for Public Belief Enhancement

osmarks.net Computational Memetics Division
comp.meme@osmarks.net

-1 April 2024

**Abstract**

Incorrect lay beliefs, as produced by disinformation campaigns and otherwise, are an increasingly severe threat to human civilization, as exemplified by the many failings of the public during the COVID-19 pandemic. We propose an end-to-end system, based on application of modern AI techniques at scale, designed to influence mass sentiment in a well-informed and beneficial direction.

## 1 Introduction

In today's increasingly complex and rapidly changing world, it is challenging for people to maintain accurate knowledge about more than a small part of the world [Kilov, 2021] [Crichton, 2002], but it's socially unacceptable or undesirable, and in some cases impossible, to reserve judgment and not proffer an opinion on every topic. As a direct consequence, many have incorrect beliefs, acting on which leads to negative consequences both for themselves and society in general [Cicero, 2001]. This is exacerbated by the increasing prevalence of misinformation, disinformation and malinformation [Pérez-Escolar et al., 2023] harming the public's ability to reach truth and make informed, justified decisions. In this hostile environment, attempts to enhance education in critical thinking are insufficiently timely and far-reaching, and a more direct solution is needed.

In this paper, we propose the Automated Persuasion Network, a system for deploying modern large language models (LLMs) to efficiently influence public opinions in desirable directions via social media. We develop an architecture intended to allow selective, effective changes to belief systems by exploiting social conformity.

# 2 Methodology

## 2.1 Overview

Humans derive beliefs and opinions from their perception of the beliefs and opinions of their peer group [Cialdini and Goldstein, 2004] [Deutsch and Gerard, 1955], as well as a broader perception of what is presently socially acceptable, required or forbidden. Our approach relies on a Sybil attack [Alvisi et al., 2013] against this social processing, executed by deploying LLMs to emulate people of similar attitudes to targets within the context of online social media platforms. While [Bocian et al., 2024] suggests that social pressure from AIs known to be AIs can be effective, we believe that persuasion by apparent humans is more robust and generalizable, especially since even the perception of automated social interaction has been known to trigger backlash or fear from a wide range of groups [Fang and Nie, 2023] [Yan et al., 2023]. We automatically derive strategies to affect desired beliefs indirectly, via creating social proof for other related beliefs, using a Bayesian network approach.

Naive implementations of this method involve many manual processing steps — for instance, identification of targets, construction of personas for LLMs to emulate, and gathering data for belief causal modelling. We replace these with automated solutions based on natural language processing — unsupervised clustering of internet users using text embeddings, direct evaluation of currently held opinions within a group using LLMs, and surveying simulacra rather than specific extant humans (as described in [Argyle et al., 2023]) — to allow operation at scale without direct human oversight. This permits much more finely individualized targeting than used in e.g. [Simchon et al., 2024] without additional human labour.

## 2.2 Segmentation

In order to benefit from the effectiveness of persuasive strategies optimized for individuals while still having enough data for reasonable targeting, we apply standard unsupervised clustering techniques. We acquire profile information and a social graph (of friendships and interactions) for all relevant social media accounts, generate text embeddings from each user's profile information, as well as a representative sample of their publicly accessible posts, and combine this with graph embeddings to generate a unified representation. We then apply the OPTICS clustering algorithm [Ankerst et al., 1999] to generate a list of clusters.

From these, several pieces of information need to be extracted. We identify the accounts closest to the cluster's centroid and take them as exemplars, and additionally compute the distribution of posting frequency and timings. We use these in later stages to ensure that our personas cannot be distinguished via timing side-channels. Additionally, we generate a set of personas using a variant of QDAIF [Bradley et al., 2023], with a standard instruction-tuned LLM (IT-LLM) used to mutate samples, using the cluster exemplars as the initial seed. As a quality metric, we ask the IT-LLM to evaluate the realism of a persona and its alignment with the exemplars, and we organize our search space into bins using k-means clustering on the generated user sentence embeddings to ensure coverage of all persona types within a cluster.

## 2.3 Analysis

We use a variant of [Powell et al., 2018]'s methodology to tune persuasion strategies to audiences to effectively affect target beliefs. We replace their manual identification and belief measurement step by using the IT-LLM to first generate a set of beliefs that relate to and/or could plausibly cause the target belief, as well as scales for measuring adherence to these possible beliefs. For measurement, rather than using the IT-LLM as before, we apply a prompt-engineered non-instruction-tuned model (also known as a foundation model, base model or pretrained language model (PT-LLM)). This is because instruction-tuned LLMs are frequently vulnerable to the phenomenon of mode collapse [janus, 2022] [Hamilton, 2024], in which models fail to generalize over latent variables such as authorship of text. This is incompatible with our need to faithfully simulate a wide range of internet users. Instruction-tuned LLMs are also unsuitable for direct internet-facing deployment, due to the risk of prompt injection [Perez and Ribeiro, 2022]. Within each cluster, we use the acquired representative text from each exemplar from the segmentation stage to condition the LLM generations, and then ask several instances the generated questions in a random order. Multiple separate sampling runs are necessary due to the "simulator" nature of LLMs [Shanahan et al., 2023]: our persona may not fully constrain its model to a single person with consistent beliefs. Runs producing responses that cannot be parsed into valid responses are discarded.

Given this synthetic data on belief prevalence, we apply a structure learning algorithm to infer causality — which beliefs cause other beliefs. Unlike [Powell et al., 2018], we do not incorporate any prior structure from theory — due to the additional complexity of applying theories in our automated pipeline, and since our requirements lean more toward predictive accuracy than human interpretability — and instead apply their BDHC algorithm to generate many candidate graphs, selecting a final model based on a weighted combination of model complexity (number of edges) and likelihood, to combat overfitting.

We then select the beliefs with the greatest estimated contribution to our target belief and direct the IT-LLM to modify our generated personas with the necessary belief adjustment. Due to the aforementioned mode collapse issues, we apply rejection sampling, discarding any generated personas that diverge too far from their original forms (as measured by semantic embedding distance) and regenerating. The resulting personas are used in the next stage.

### 2.4 Interaction

After the completion of the previous stages, the Automated Persuasion Network must interact with humans to cause belief updates. This step requires large-scale inference: however, as most human communication is simple and easy to model, at least over short contexts, we are able to use standard low-cost consumer GPUs running open-weight PT-LLMs, using the vLLM [Kwon et al., 2023] inference server. As an additional cost-saving measure, we use a multi-tiered system whereby generations are initially run on a small model and, if too complex for it (as measured by perplexity), rerun using a more capable language model.

We use the belief-modified personas generated in the Analysis stage, and attempt to have each of them mimic the actions of a human user in their cluster as much as possible. We identified a number of challenges. Most notably, nonhuman users are frequently detected using posting frequency [Howard and Kollanyi, 2016] and timings [Duh et al., 2018] [Pan et al., 2016]. By using a fairly large set of accounts rather than a single bot, we can avoid detection based on simply noticing anomalously high posting frequencies, and by scheduling generation of new posts and conditionally replying to other users' posts in accordance with cluster statistics for such gathered during the Segmentation stage we can prevent easy timing-based detection. We have not yet identified a complete strategy for avoiding social-graph-based detection such as [Alvisi et al., 2013]: our present best mitigation is to deploy new personas slowly and to maintain the rate of interaction between them at the base rate within the cluster.

Other difficulties involve technical countermeasures in use against nonhuman users, such as CAPTCHAs and limited APIs. However, while today's most sophisticated CAPTCHAs exceed current AI capabilities, commercial services are available to dispatch solving to humans at very low cost. We are able to mitigate other limitations with the use of commercial residential proxy services and browser automation software for scraping.

### 2.5 Monitoring

In order to determine the efficacy of our approach, we periodically sample posts from human users within each cluster and apply the IT-LLM to rate how much each post entails our target beliefs, allowing measurement of belief change over time.

## 3 Results

No results are available for release at this time.

## 4 Discussion

We believe our architecture represents a major advance in misinformation prevention and public attitude alignment. A promising future direction for research we have identified is introduction of technical enhancements such as implementation of speculative decoding in post generation, as well as use of vision/language models such as [Liu et al., 2023] to allow interaction with multimodal content. We also suggest integration of concepts from LLM agents to reduce distinguishability from humans — for instance, personas could be given the ability to create new posts based on newly released news articles or information from other social media sites. Finally, while we have primarily focused on human emulation with some limited optimization of persuasive strategies, future AI technology is likely to be capable of more powerful direct persuasion.

## References

[Alvisi et al., 2013] Alvisi, L., Clement, A., Epasto, A., Lattanzi, S., and Panconesi, A. (2013). Sok: The evolution of sybil defense via social networks. In *2013 IEEE Symposium on Security and Privacy*, pages 382–396.

[Ankerst et al., 1999] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. In Delis, A., Faloutsos, C., and Ghandeharizadeh, S., editors, *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*, pages 49–60. ACM Press.

[Argyle et al., 2023] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

[Bocian et al., 2024] Bocian, K., Gonidis, L., and Everett, J. A. C. (2024). Moral conformity in a digital world: Human and nonhuman agents as a source of social pressure for judgments of moral character. *PloS one*, 19(2):e0298293.

[Bradley et al., 2023] Bradley, H., Dai, A., Teufel, H., Zhang, J., Oostermeijer, K., Bellagente, M., Clune, J., Stanley, K., Schott, G., and Lehman, J. (2023). Quality-diversity through ai feedback.

[Cialdini and Goldstein, 2004] Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annual Review of Psychology*, 55:591–621.

[Cicero, 2001] Cicero, M. T. (2001). *On the Ideal Orator*. Oxford University Press.

[Crichton, 2002] Crichton, M. (2002). Why speculate? `https://web.archive.org/web/20070714204136/http://www.michaelcrichton.net/speech-whyspeculate.html`. Speech presented at the International Leadership Forum, La Jolla, California, April 26.

[Deutsch and Gerard, 1955] Deutsch, M. and Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgement. *Journal of Abnormal Psychology*, 51 3:629–36.

[Duh et al., 2018] Duh, A., Rupnik, M. S., and Korošak, D. (2018). Collective behavior of social bots is encoded in their temporal twitter activity. *Big Data*, 6(2):113–123.

[Fang and Nie, 2023] Fang, W. and Nie, C. (2023). Social media use, social bot literacy, perceived threats from bots, and perceived bot control: a moderated-mediation model. *Behaviour & Information Technology*, 0(0):1–17.

[Hamilton, 2024] Hamilton, S. (2024). Detecting mode collapse in language models via narration.

[Howard and Kollanyi, 2016] Howard, P. N. and Kollanyi, B. (2016). Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum.

[janus, 2022] janus (2022). Mysteries of mode collapse. Accessed: 2022-11-08.

[Kilov, 2021] Kilov, D. (2021). The brittleness of expertise and why it matters. *Synthese*, 199(1):3431–3455.

[Kwon et al., 2023] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. (2023). Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[Liu et al., 2023] Liu, H., Li, C., Li, Y., and Lee, Y. J. (2023). Improved baselines with visual instruction tuning.

[Pan et al., 2016] Pan, J., Liu, Y., Liu, X., and Hu, H. (2016). Discriminating bot accounts based solely on temporal features of microblog behavior. *Physica A: Statistical Mechanics and its Applications*, 450:193–204.

[Perez and Ribeiro, 2022] Perez, F. and Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models.

[Powell et al., 2018] Powell, D., Weisman, K., and Markman, E. M. (2018). Articulating lay theories through graphical models: A study of beliefs surrounding vaccination decisions. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

[Pérez-Escolar et al., 2023] Pérez-Escolar, M., Lilleker, D., and Tapia-Frade, A. (2023). A systematic literature review of the phenomenon of disinformation and misinformation. *Media and Communication*, 11(2):76–87.

[Shanahan et al., 2023] Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role play with large language models. *Nature*, 623(7987):493–498.

[Simchon et al., 2024] Simchon, A., Edwards, M., and Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 3(2):pgae035.

[Yan et al., 2023] Yan, H. Y., Yang, K.-C., Shanahan, J., and Menczer, F. (2023). Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports*, 13(1):20707.