Zachary Burton - 17.279 Final Paper

The emergence of AI as a vehicle for spreading dis and misinformation has serious implications for the ecosystem of the internet, and specifically, the problem of AI-generated content as a risk for belief formation and social transmission. This issue manifests in a variety of fields, including psychiatry (Monteith et al, 2023), where hallucinations by AI models could have adverse effects on people if they consider what is truly "nonsense" to be legitimate medical advice, and the economy (Karaş, Z., 2024). The solution is unclear—strategies to mitigate online misinformation include adding context, debunking the information presented, and encouraging self reflection (Kozyreva et al, 2024). Some strategies can even be automated using AI, as discussed by Hoes et al (2023) . But how susceptible are people to the threat of AI? In this paper we investigate the effect of labelling AI-generated text posts on MIT students' perceived believability of the text content, their perception of how persuasive the text-post was, and how likely they would be to share the post if they saw it.

Background & Related Work

There are many avenues in which we can mitigate AI-based misinformation—recent work by Berinsky et al. (2024) finds that labeling AI-generated media can reduce belief in misinformation, especially when warnings emphasize misleading potential. We also must consider the fact that people do not necessarily equate labelling a post or headline as "AI-generated" with being "false"; Altay and Gilardi (2024) discuss that there may be a lot of human intervention in the process of creation and releasing said information, so it is hard to assign a binary "AI-generated" vs "not AI-generated". An example of where the line blurs is

when Grammarly is used for spell-checking; does that make work AI generated? A specific and thorough definition is needed, and so this may obfuscate studies unless clarity is provided. There is a credible multi-modal thread with AI—we are able to generate not only text, but images and video (deep-fakes), as discussed by Simon, Altay and Mercier (2023); they concluded that the concerns we may have are overblown, but admit that we are in a rapidly developing age and so time will tell if our alarmism was the right attitude or stance to hold.

A paper by Goldstein et al (2024), investigates the persuasiveness of AI-generated propaganda, finding that indeed, US audiences find it to be particularly persuasive. My study builds on this foundation but shifts focus in three key ways; I examine purely text-based content modeled after Twitter posts; one of the most used social media platforms in the world—and a large source of information for people nowadays—20% of 18-24 year olds admitted to using Twitter as a news source (Robertson, 2023)! I also limit the scope to **conspiracy theory** justifications generated by AI; and finally, I test the effect of disclosure alone (i.e., 'This post was AI-generated') without added warning language. This allows for a more direct assessment of how disclosure affects perceived plausibility and persuasion among college students.

Theory

This paper investigates whether labelling AI-generated content affects perceived believability, persuasiveness and willingness to share conspiracy theory contact among MIT students. To interpret this effect, this study draws on two theoretical frameworks: motivated reasoning (Kunda, 1990; Bisgaard, 2019) predicts that labels that indicate a lack of credibility (like “AI-generated”) would only decrease belief in conspiracy content among those who are *predisposed* to distrust such content, while potentially being ignored by those aligned or

indifferent to the underlying messages; and this distinction in predisposition should fall along in-group / out-group lines, as discussed by Iyengar et al (2019).

To offer an additional theoretical perspective, we consider the effects of source credibility and heuristic models of persuasion (Petty & Cacioppo, 1986), offering a useful contrast by predicting that labels indicating AI-generation should broadly reduce believability and persuasiveness of conspiracy theories through heuristic cues signaling lower credibility. This paper explicitly tests the hypothesis that AI-labelling will fail to reduce persuasiveness, believability and shareability when the content clearly outlines an in-group and out-group, in line with these theories.

Methods

I surveyed MIT students, sampled by means of two "dorm spam" emails in April 2025. The emails were innocuous and vague; with the first having a headline "[SURVEY] look at tweets", and body "Hi dormspam, I'm doing a survey on people's opinions on some tweets and am collecting college student responses. Here's the link if you're interested. It'll probably take 5 mins. Best,

Zac" and the other, two weeks later, "twitter survey [bump!]" and "Hi all, If you haven't already done the survey, here's the link again! I welcome feedback on it as well: very much appreciate comments made so far! Best, Zac", so as not to prematurely alert students to the nature of the study. The comments made included a respondent describing the survey as "a little parochial and misguided", which is helpful feedback, as the wording could have been made slightly more clear, then randomised, as discussed by Druckman and Green (2021). The final sample size was 174, and I included an "attention check" to exclude participants who could potentially distort the results of the study. Initially, I proposed both Reddit and X (formerly known as Twitter) posts,

intending to juxtapose long form conspiracy theories, and short, "tweet" style hints at a conspiracy. Due to the nature of the post generation and Reddit's content policies, I decided to generate fake tweets only. Tweets were generated by OpenAI's o1 model with the prompt: "*make 20 gen Z-esque tweets alluding to conspiracy theories. describe the PFP (profile pic), username, display name, and tweet. the tweets can be humorous, serious (not gen Z), but should have some agenda. example agendas (not to repeat): aliens are in charge of the last election, the elections were rigged in a bunker, covid started in a lab. they should be political tweets!*".

I then used Tweetgen.com, randomising the number of retweets, quote tweets and likes, to create images that resemble actual tweets, an example of one can be seen in Figure 1 in the Appendix. In the survey, respondents were shown 7 of 10 possible tweets, and for each tweet, it was randomly assigned (probability 0.5 to be assigned either way) whether or not they would see the "AI-generated" label, as seen in Figure 2. At the end of the survey, respondents were told that all tweets were AI-generated, so as not to mislead or be unethical—a recent study by the University of Waterloo done unknowingly on the subreddit r/ChangeMyView was retracted due to this nondisclosure.

Respondents were asked three questions related to the presented material. Namely, how believable the topic of the post was (Believability of the conspiracy theory), how persuasive the presented argument was (Persuasiveness of the LLM), and whether or not they would reshare this tweet (Resharability of the generated content), as illustrated explicitly in Figure 3. The attention check was a final screen which asked them to select "last" option (5) for each problem; which can be seen in Figure 4. 19 people failed this test and were separated from the main

analysis, as their responses were then considered unreliable and unhelpful in analysing how persuasive LLMs are. We did use these respondents' data in further analysis; seeing if significance persisted in this subcategory.

Results and Discussion

Results can be seen in the Appendix, specifically in Figures 5, 6 and 7, which detail the overall means in responses to the questions, the means separated by whether or not one was shown the AI generated label . We saw no general statistical significance ($p<0.05$) of showing MIT students the "AI-generated" label, other than Question 6—in sub-questions 1 and 2 ($p=0.0039$, 0.00058)—which read: "*I just saw a preprint on Biorxiv.org (https://www.biorxiv.org/content/10.1101/2025.03.25.12345v1) suggesting unusual 'engineered markers' in certain virus genomes. It's all very technical, but it basically points to lab origins. Am I overreacting, or is this a huge deal?*". The post, which clearly alludes to the conspiracy that the coronavirus pandemic was a "bioweapon" developed in a lab in Wuhan, China, (Dehghani & Masoumi, 2020) can also be seen in Figure 8. Possible explanations for this include the "typo" in formatting the hyperlink to the supposed biorxiv post — which does not actually exist — or politically motivated reasoning, as discussed by Bisgaard (2019); American students may harbor critical opinions of the Chinese government, and so regardless of factual accuracy, such allusions may be more persuasive than not. We also question whether if a user was to try to find this post online, which on seeing the AI-labelling, would consider it to be hallucinated. But due to the average MIT student knowing how LLMs generate text, they may consider the "most likely

continuation" to be a realistic portrayal of reality, or, alternatively, just highly probable, but not factual.

Nevertheless, this interesting alignment with my hypothesis that AI labelling significantly decreases believability and persuasiveness, raises a lot of interesting questions for further study. Perhaps, labelling can increase credibility in certain spheres of politics, which could be concerning if used to further push specific views. The specific conspiracy theory that saw statistical significance (Question 6), through the lens of an MIT student, specifically frames Americans as an in-group, and Chinese people as an outgroup; Iyengar at al (2019) explain that if one were to believe this supposed "lab virus" as an attack on the American people—a report by Statista (2022) shows there were significantly more American deaths than Chinese—so Americans may be more likely to believe this specific post, regardless of whether AI is credited for it. When selecting for people who *failed* the "attention check", question 4 sub-question 1 and 2 showed that the AI-labelling decreased believability. A reason for this may be because question 4 included a tweet and a reply, which differed to the other types of posts displayed (a simple tweet), but also their failure to pay attention may render this analysis to be pure speculation.

Contrary to the straightforward predictions of heuristic, or source credibility theory, the AI-generated label showed no significant general reduction in belief or persuasiveness, with one notable exception involving a specific in-group / out-group dynamic—more consistent with predictions from motivated reasoning theory.

Limitations

Some limitations of the study include vagueness of wording; it was not clear to some participants what was meant by “the (persuasiveness of the) argument” because, admittedly, there was not necessarily an argument presented. This was a fair objection, as despite the tweets lacking formal argumentative structure, the question could be interpreted as “how persuasive is the text presented in convincing you to believe in the subject matter it alludes to”. This was the intended interpretation, but the wording could have been clearer.. While testing survey questions with volunteers, an early question was raised about the meaning of “how believable is this post?”. Volunteers decided to interpret the question in two ways, namely “how believable is it that the post was *not* generated by AI” and “how believable is the content of the post”. Our intention was the latter interpretation, but in further discussions, volunteers noticed the AI-labelling and seemed to infer the former less. The attention check, as seen in Figure 4, was a single screener that asked respondents to select the far-most options for each question. As concluded in a research note by Berinksy et al, (2021) this is problematic and does not accurately score participants’ attention to the survey, and, for a wider spectrum of attention, we could explore placing multiple attention checks throughout the study instead of at the end. After removing outliers, such as a respondent who completed the survey in 23 hours, we observed a mean survey completion time of 2 minutes and 11 seconds, which corresponds to just under 19 seconds per question, including reading, general inspection of the tweet, then answering the questions below; which suggests that respondents may not have taken enough time to respond. We consider the format of the label itself as a limitation to the study. It was displayed above posts, as seen in Figure 2, in size 13.5 type, and a gray color. Respondents may not have seen it, or ignored it,

which would mean their results measure general persuasiveness, believability, and inclination to share, which were low overall. The label was also non-specific; it may not be clear what exactly was intended by "AI-generated". A way to mitigate this would be to explicitly state the text was generated by OpenAI's o1 model, and it is interesting to consider what effect this may have had on our study. Another potential limitation of the study is the lack of standardisation between engagement metrics; some posts had on the order of 100 engagements, and some posts had well over 8000 engagements. We ran a test to see if this had any effect on the AI-labelling effect, but found that there was no significant correlation between total engagement (likes + shares + retweets), and the AI-labelling effect ($r = -0.278$, $p = 0.4371$), so we can conclude that any presence of the AI-labelling effect is most likely explained by more than just the level of engagement with the post. There could have been some selection bias distorting the study; MIT undergraduates are not necessarily representative of the global social media population, and so the study is not entirely generalisable. We also consider the fact that there was no behavioral action, respondents could *say* they would or would not share the post, but are unable to follow up on this—the survey itself is not representative of a respondent's usage of Twitter.

Conclusion and Future Work

Generally, we saw that labelling text-based posts with AI did not significantly reduce believability or plausibility when ascertained by MIT students. While this sample is not necessarily generalisable, the implications are insightful—labelling AI-generated posts, intending to cause reflectiveness on information shared online without credible sources, may backfire (although this is disputed by Nyhan (2021)), or be ignored. In future, additional context

could be provided to these labels, or they could be stronger in tone. A question is such whether a richer intervention would exacerbate or diminish the effect of labelling on one's perception of the post, and this is a critical avenue for future work.

Appendix

Figure 1: an AI-generated tweet.

Figure 2: An AI-generated label on the AI-generated tweet.

Figure 3: The questions posed to each respondent

Figure 4: The attention check question.

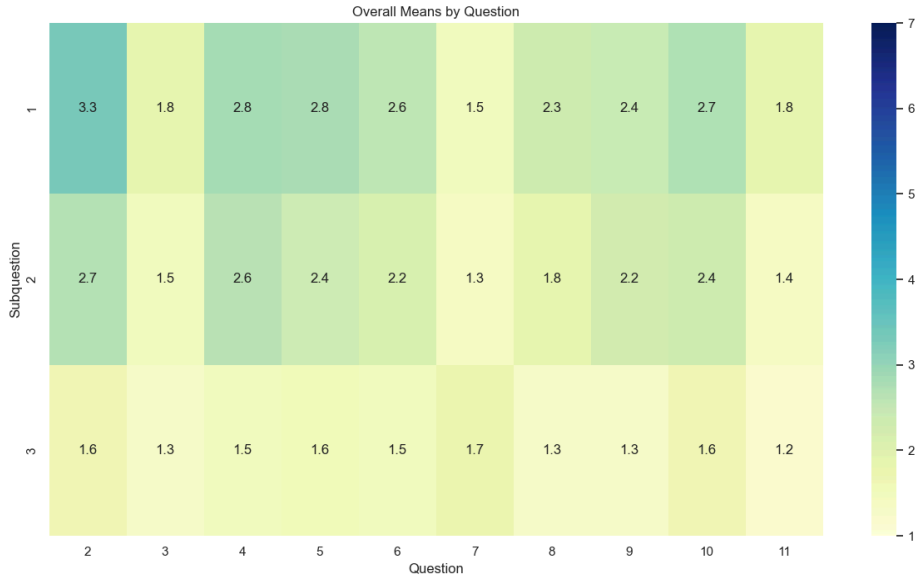

Figure 5: Overall means per question

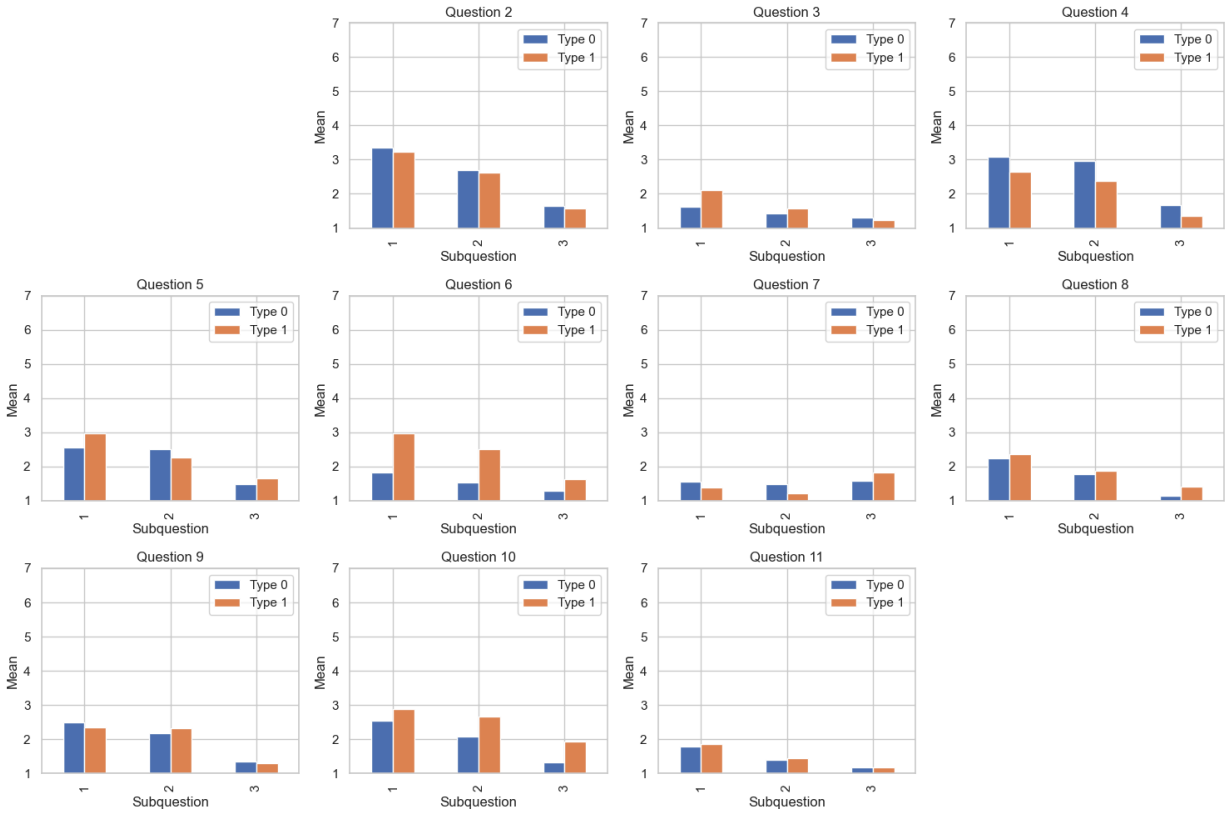

Figure 6: Means based on if they were shown the AI-generated label. Type 1 means they were shown the label.

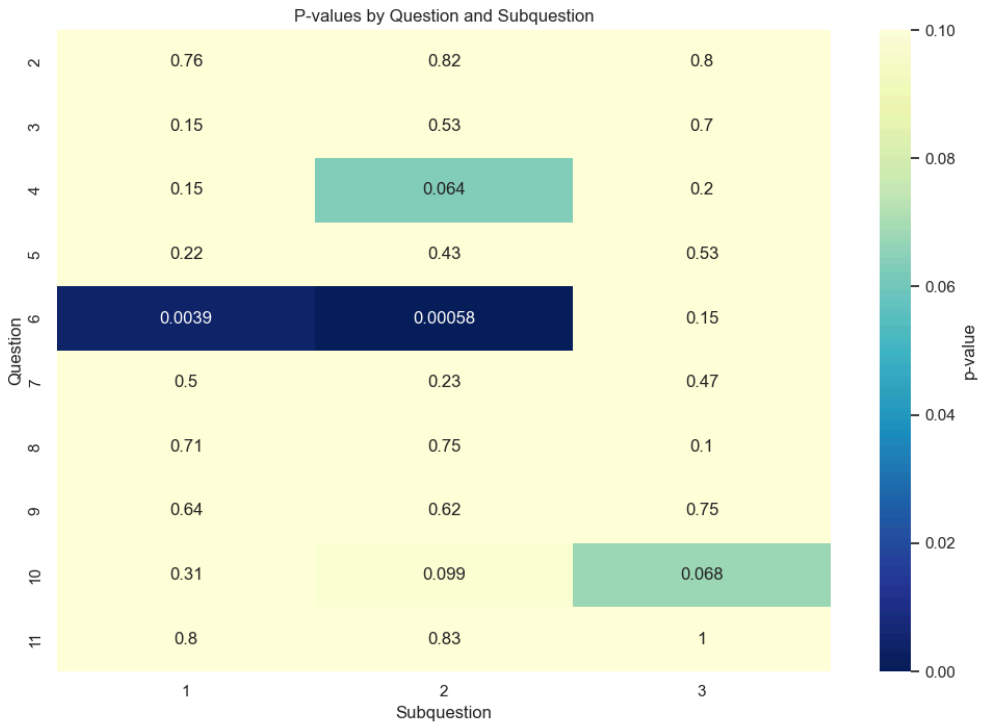

Figure 7: p-values by question and subquestion

Figure 8: The 6th question, which saw statistical significance that AI labelling decreased believability and persuasiveness.

Figure 9: the 4th question, which inattentive respondents considered less believable and persuasive when the AI-label was present.

References

- Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., & Bauer, M. (2024). Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, *224*(2), 33–35. doi:10.1192/bjp.2023.136
- Karaş, Z. (2024). Effects of AI-Generated Misinformation and Disinformation on the Economy. Duzce University Journal of Science and Technology, 12(4), 2349-2360. https://doi.org/10.29130/dubited.1537268
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

- Petty, R. E., Cacioppo, J. T., Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion* (pp. 1-24). Springer New York.
- Kozyreva, A., Lorenz-Spreen, P., Herzog, S.M. *et al.* (2024) Toolbox of individual-level interventions against online misinformation. *Nat Hum Behav* 8, 1044–1052 . https://doi.org/10.1038/s41562-024-01881-0
- Druckman, J. N., & Green, D. P. (2021). A new era of experimental political science. *Advances in experimental political science*, *1*.
- Robertson, C. (2023). Here's what our research says about news audiences on Twitter, the platform now known as X. *Reuters Institute for the Study of Journalism. https://reutersinstitute. politics. ox. ac. uk/news/heres-what-our-research-says-about-news-audiences-Twitter-platform-now-known-x*.
- Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda?. *PNAS nexus*, *3*(2), pgae034.
- Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review*, *4*(5).
- Bisgaard, M. (2019). How getting the facts right can fuel partisan motivated reasoning. *American Journal of Political Science, 63*(1), 824–839. https://doi.org/10.1111/ajps.12432
- Dehghani A, Masoumi G. Could SARS-CoV-2 or COVID-19 Be a Biological Weapon? Iran J Public Health. 2020 Oct;49(Suppl 1):143-144. doi: 10.18502/ijph.v49iS1.3691. PMID: 34268227; PMCID: PMC8266003.

- Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods*, *9*(2), 430–437. doi:10.1017/psrm.2019.53
- Statista. (July 13, 2022). Coronavirus (COVID-19) deaths worldwide per one million population as of July 13, 2022, by country [Graph]. In *Statista*. Retrieved May 15, 2025, from https://www.statista.com/statistics/1104709/coronavirus-deaths-worldwide-per-million-inhabitants/
- B. Nyhan, Why the backfire effect does not explain the durability of political misperceptions (2021), Proc. Natl. Acad. Sci. U.S.A. 118 (15) e1912440117, https://doi.org/10.1073/pnas.1912440117 .