Detecting Persuasion Techniques in Online Propaganda: an AI-annotation Intervention

Lu Xiao
Department of Information Systems
Arizona State University
lxiao31@asu.edu

Yimin Xiao
College of Information
University of Maryland
yxiao@umd.edu

Giovanni Da San Martino
Department of Mathematics
University of Padova
giovanni.dasanmartino@unipd.it

Preslav Nakov
Department of Natural Language Processing
Mohamed bin Zayed University of Artificial Intelligence
preslav.nakov@mbzuai.ac.ae

Changzhou Zheng
Carnegie Mellon University
changzhz@andrew.cmu.edu

Abstract

In today's digital world, online propaganda poses a growing threat to society. To address this challenge, we explored the potential of a real-time training intervention. This intervention exposes people to manipulative text identified by AI and provides cognitive guidance to help them recognize and critically analyze similar cases. To examine whether people trust AI technology enough to read the annotations and follow them, we conducted a one-factor experiment. Our findings show no difference in training performance whether the annotations are claimed to be provided by domain experts or by AI. The intervention is effective in improving people's ability to identify propagandistic texts and persuasion techniques, and performance varies by technique. We also found that people judge the texts based on one or more of three aspects: the content of the material, the author's intention inferred from the material, and their personal ideology related to the topic of the material.

Keywords: online propaganda, persuasion, annotation, training, AI.

1. Introduction

Persuasion is commonly defined as "human communication designed to influence others by modifying their beliefs, values, or attitudes" (Simons, 1976, p. 21). In propaganda, various language-based persuasion techniques are employed to foster a predetermined agenda in society, intending to manipulate people's minds (Miller, 1939; Weston, 2018). These strategies use logical fallacies and appeal to emotions, such as using words/phrases with strong emotional implications to influence an audience (Weston, 2018) and phrases that play on strong national feelings to justify or promote an action or idea (Hobbs & McGee, 2008). It is evident that such faulty argumentation may sway the opinion of listeners unaware of them. The impact of such manipulations is amplified by the potentially huge audiences that can be reached via the Internet and social networks (Da San Martino et al., 2019). As a result, online propaganda becomes increasingly common and powerful in manipulating Internet users, imposing an adversarial impact on society when a large population relies on online information to make sense of the world.

To address this prominent problem, various computational tools have been developed to automatically detect the persuasion techniques employed in online propaganda (e.g., Li, Ye, & Xiao, 2019; Da San Martino et al., 2019; Chen & Xiao, 2023). With the success of such techniques, we expect to encounter less online content containing manipulative information and persuasion attempts, or at least be informed of such risks when consuming the content. However, the performance of even the best-reported algorithm is still not very satisfactory. For instance, a state-of-the-art algorithm by Hristakieva and colleagues (Hristakieva et al., 2022) reported an F1 measure of 0.83.

A recent study suggests that it is humans, not robots, who tend to spread false news more than the truth (Vosoughi et al., 2018). While the study did not exclude the possibility that people spread the fake news out of malice, it did suggest that they may do so under the influence of their aroused emotions by the news and the novelty of the topic. This indicates the importance of educating people to better identify false information to prevent the negative impact of its spread.

In this study, we explore interventions that improve people's capability of identifying the common persuasion techniques applied in online propaganda, focusing on educating Internet users to better resist such manipulations. We examine the effectiveness of an AI-

URI: https://hdl.handle.net/10125/108965
978-0-9981331-8-8

annotation intervention. With this intervention, people are shown texts with the persuasion techniques annotated by AI as examples of how they are employed in online propaganda. The central research question is whether this intervention is an effective training tool to help people detect online propaganda and the persuasion techniques applied in the content.

An AI-based approach can automatically annotate the text the individual is reading at that moment. We expect that social media users are more likely to read the annotations on the same content they are browsing as it involves less cognitive load and connects the annotation task with the goal of their browsing behavior – to read the content. However, do users have the same level of trust in the annotations performed by an algorithm as those performed by domain experts? This trust is important to the effectiveness of the intervention, as people's trust in AI is shown to affect their willingness to use AI (Chuong et al., 2023). To answer this question, we compared the training outcomes of two annotation conditions: AI-based annotations vs. expert-based annotations. Specifically, we provide participants annotated texts through an interface that highlights propagandistic text segments and the corresponding persuasion technique(s) used in these segments. In the interface, these materials are presented as either produced by machine learning algorithms or generated by human experts despite the same content and annotations. Figure 1 shows what the interface looks like in the condition of expert-based annotation. As shown in the figure, the highlights in the excerpt on the left indicate the text segments that contain persuasion techniques, and the heading of the right panel claims that the annotation is based on human expert evaluation. The expert note on the right is an explanation of the specific techniques (corresponding to different highlight colors) being used in the segments. In the AI-based condition, the only difference is the heading of the right panel.

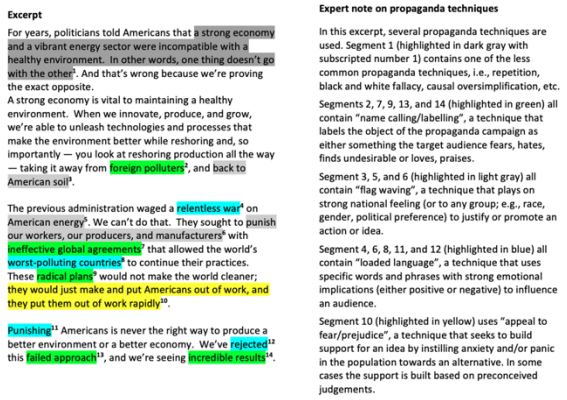

Figure 1. An example of annotation-based media literacy intervention material and the interface in the condition of expert-based annotation

This interface is developed based on the web interface of PRTA, a web-based system that continuously crawls articles from many media sources and automatically shows the techniques detected in them (Da San Martino et al., 2020). To further explore the potential of using AI algorithms to create real-time annotations of propaganda in online content, participants perform an information retrieval task on PRTA, and we probe their level of trust in the annotations provided by PRTA.

Another important factor that affects the effectiveness of this AI-based annotation intervention is how people reason about whether a piece of information is propagandistic and whether certain persuasion strategies are used. To gain insights on this factor, we explore the answers of two questions in the study: what are the reasons people provide to justify their identification of propagandistic texts in general and across different persuasion techniques? How do they mediate the impact of the intervention?

The rest of the paper is organized as follows. We first report research work on the trust issues between people and AI technologies. Next, we present the experimental design of our study. This includes how we identify and annotate the texts that contain propaganda, the participant recruitment and experiment procedure, and our analysis of the experiment data. We present the experiment results in response to each research question afterward. We discuss the implications of our study and conclude with the limitations of the study and suggestions for future directions.

2. Related work

2.1. Trust in AI

Trust is "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau et al., 1998, p. 395). Mayer, Davis, and Schoorman (1995) drew a distinguishing line between trustworthiness and trust propensity. Trustworthiness includes three aspects of a trusted individual: their ability, benevolence, and integrity. Trust propensity, on the other hand, is an individual's general willingness to trust others. In Human-AI trust research, scholars also made this distinction, noting that a user may trust an AI system that is not trustworthy, whereas an AI system that was developed to be "worthy of trust," such as being capable of carrying out tasks and providing reliable results, may not be trusted by the user (Jacovi et al., 2021). To guide the development of trustworthy AI, the European Commission has provided a list of requirements including various aspects such as the system being transparent, ensuring privacy, providing

accountable behavior and outcomes, and considering environmental and societal well-being (Floridi, 2019).

People tend to trust what they already know, and as their knowledge grows, their trust accumulates (Boyd, 2003). When interacting with intelligent machines, users not only consider the kind of tasks these machines can undertake and their efficiency and effectiveness, but they also weigh the risks involved in trusting the machine's decision (Stanton & Jensen, 2021). Their level of trust in AI affects their intentions to use AI (Chuong et al., 2023). Additionally, AI's reliability and its perceived trustworthiness are positively correlated (Kaplan et al., 2023).

2.2. Justificatory reasoning

Justificatory reasoning is an individual's ability to justify one's beliefs, opinions, and actions based on logically and ethically grounded arguments (Collins, 2004). There are three epistemological forms of justificatory reasoning: absolutism, relativism, and evaluativism (Kuhn, 1992; Kuhn & Udell, 2003). People who use absolutism to justify their opinions or actions refer to external authorities, domain experts, or norms they believe to be commonly held by society. Relativists believe that there is no absolute right or wrong in any situation; hence, different opinions have equal value and can all be accepted. Evaluativism-based justificatory reasoning makes logical connections between facts or empirical evidence and conclusions. Often, absolutism reasoning is considered the most primitive form of justificatory reasoning (Soong, Lee, & John, 2012; Knight & Collins, 2006), whereas evaluativism is considered the highest form of justificatory reasoning (Kuhn & Udell, 2003).

In the context of our study, participants need to identify persuasion techniques in propaganda and justify their identifications. While there have been various studies that examine people's reasoning processes in accepting and spreading fake news and propaganda (e.g., Tandoc et al., 2020; Pennycook & Rand, 2019; Bago, Rand, & Pennycook, 2020), to the best of our knowledge, how interventions that help people identify misleading content play a role in people's justificatory reasoning has not been explored.

3. Experiment design

We conducted a one-factor experiment in which the participants examined their performance at identifying persuasion technique usage both before receiving annotation-based instructions and after. We randomly displayed these instructional materials in two conditions: expert-based instruction and AI-based instruction. Following the experiment, we asked the participants to go to the PRTA site for a usability test. The study was approved by the Institutional Review Board of the second author's university.

3.1. Experiment materials

We retrieved eight presidential speech excerpts from the official White House archive, covering topics like climate change, clean energy, and environmental issues. Preliminary content analysis indicated these excerpts contained many instances of persuasion techniques, making them suitable for our identification tasks and interventional materials. The excerpts averaged 192 words and included an average of eight persuasion technique instances.

Three co-authors annotated the excerpts based on the guidelines from Da San Martino et al. (2020), serving as the ground truth for the experiment. Two of the annotators were domain experts in persuasion techniques with four years of experience, and the third had extensive experience in content analysis. They held four hourly meetings to agree on the annotation results. We retained the five most common persuasion techniques (loaded language, name calling, appeal to fear, exaggeration and minimization, and flag-waving) and one group of techniques ("others," including black-and-white fallacy, repetition, and causal simplification) for the experiment. The definitions of these techniques are as follows:

1. ***Loaded language*** (Weston, 2018) uses emotionally charged words to sway an audience.
2. ***Appeal to fear/prejudice*** fosters support for an idea by inducing anxiety about alternatives.
3. ***Name calling/labeling*** (Miller, 1939) attaches negative or positive terms to the subject to influence perception.
4. ***Exaggeration, minimization*** (Jowett and O'Donnell, 2012) amplifies or diminishes the importance of something.
5. ***Flag-waving*** (Hobbs and Mcgee, 2008) plays on strong national feelings (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or an idea.
6. ***Others (repetition, black and white fallacy, causal oversimplification)*** (Torok, 2015): Repetition consists in repeating the same message repeatedly. In black and white fallacy, two alternative options are presented as the only possibilities, when in fact more alternatives exist. Causal oversimplification assumes one cause when there are multiple causes behind an issue. We consider these three techniques as a group due to the low

number of instances present in the sample we used for our study.

Black and white fallacy and causal oversimplification are logical fallacies. The other techniques considered in this annotation intervention study appeal to the emotions of the audience.

3.2. Experiment procedure and measures

Participants completed the study via an online questionnaire. In the main experiment, we randomly assigned unannotated excerpts 1-4 as pre-test materials and excerpts 5-8 as post-test materials. During the pre-test, participants were tasked with (i) rating their agreement with the excerpt content on a 7-point scale, (ii) identifying instances of persuasion techniques, providing rationale, and rating confidence on a 5-point scale. As part of the intervention, we presented annotated versions of excerpts 1-4 to participants as the ground truth of propaganda usages, in two conditions: expert-based and AI-based instruction. Following the intervention, participants completed the post-test with unannotated excerpts 5-8, following the same procedure as the pre-test. Finally, participants rated their firmness of attitude on the topics covered in the experiment using a multi-item scale (Beltramini, 1988; e.g., "I am very likely to stick with my opinion on the issue no matter what", Cronbach's alpha = 0.87) and their trust in the AI algorithm used in the system on a 5-point scale (1 = not trust the information at all, 5 = completely trust the information). After the main experiment, participants evaluated the usability of the Prta system by (1) searching for four news sources of their choice and comparing their propaganda usages in articles covering the topic and (2) reading four articles from sources with different political biases (as classified on the website) and reviewing the propaganda information presented. Following completion of the task, participants rated the usability of the Prta system on the System Usability Scale (SUS, Brooke, 1995; e.g., "I found the website unnecessarily complex", Cronbach's alpha = 0.87) and their trust in the system on another multi-item scale (e.g., McKnight, Carter, Thatcher, and Clay, 2011, "I am totally comfortable using the propaganda classification algorithm", Cronbach's alpha = 0.78). Average ratings on the items were used for data analysis.

3.3. PRTA system

The annotation interface used in the experiment is built upon the web interface of a multi-granularity neural network known as PRTA. PRTA integrates signals at both the sentence and fragment levels to enhance the detection of persuasion techniques. Through the provided web interface, PRTA serves as a standalone service accessible to internet users (Da San Martino et al., 2020). It continuously scans articles from various media sources, automatically displaying detected techniques (refer to Fig. 2). Moreover, users have the option to upload their text for analysis.

Figure 2. The interface of the Prta system, showing an article with automatically detected persuasion techniques

3.4. Participants

We recruited participants from a major university in the northeast U.S. using the university's email listservs. A total of 72 participants completed the study between October and November 2020, with an average age of 23.67 years. Among them, 27 participants were originally from the U.S., while 47 had origins in other countries. The average completion time for participants was 74.68 minutes (SD = 34.62). Each participant received monetary compensation of $25. In the experiment, 39 participants were randomly assigned to the human condition, and 33 were assigned to the AI condition.

4. Results

4.1. Is the intervention effective?

Education literature has shown that compared to traditional, practice-based problem solving, students learn schemas and patterns better when the practice problems are paired with worked examples (Cooper & Sweller, 1987; Sweller & Cooper, 1985; Atkinson et al., 2000). As an instructional device, worked examples show learners an expert's problem solution. We therefore hypothesize that it is effective to improve people's capability of identifying the persuasion techniques on their own by showing them texts that have persuasion techniques annotated. Our results prove that this hypothesis is true. We detail our analysis processes and results in the remainder of this subsection.

We assessed participants' abilities to identify propagandistic segments and persuasion techniques before and after the intervention. Their performance was measured based on two factors: (1) the accuracy of their identification results and (2) their confidence in the identification. To gauge accuracy, we categorized participants' identification results into three scores, as outlined in Table 1.

Table 1. Coding schema for participants' performance on each instance they identified as to contain persuasion technique usage

Score	Label	Definition
1	Correct	The participant correctly identified a propagandistic text segment and the persuasion technique
0	Partially correct	The participant correctly identified a propagandistic text segment but mislabeled the persuasion technique
-1	Error	The participant misidentified a text segment to be propagandistic

Three co-authors coded all 1,887 identified instances. We assessed a participant's accuracy in pre- and post-test performance by analyzing their score distribution in identified instances separately for each test. We then employed the Wilcoxon Signed Rank Test to determine if the pre- and post-test performances differed significantly. As indicated in Table 2, although there was an increase in a participant's correction rate, it was not statistically significant. However, there was a significant decrease in the participant's partial correction rate and error rate. These findings suggest that the annotation-based intervention effectively improved participants' precision in identifying persuasion techniques.

Table 2. Comparison between pre- and post-intervention accuracy at participant level

Measurement for participant's accuracy	Pre-test average	Post-test average	Wilcoxon Signed Rank Test P value
Participant's correction rate	0.3489	0.4110	Not significant
Participant's partial correction rate	0.4214	0.3524	< 0.001
Participant's error rate	0.2297	0.1814	< 0.05

The comparison of participants' confidence levels in identifying persuasion techniques before and after the intervention further supports this finding. We conducted a Mann-Whitney U test to compare the difference between participants' pre- and post-average confidence levels. The statistical test results revealed a significant increase in participants' confidence in identifying persuasion techniques and text segments (pre-test average = 4.1841, post-test average = 4.4246, $p < 0.001$).

To further explore the impact of the annotation-based intervention on the detection of specific persuasion techniques, we compared participants' average accuracy and confidence for the identification they made under each persuasion technique. The analysis results are presented in Table 3 below.

Table 3. Comparison of pre-intervention and post-intervention. The score is calculated based on the coding criteria (Table 1)

Persuasion Technique	Accuracy			Confidence		
	Pre-test Average	Post-test Average	Mann-Whitney U test	Pre-test Average	Post-test Average	Mann-Whitney U test
Loaded language (N = 397)	0.1203	0.2841	$p < 0.05$ effect size: 0.43	3.9398	4.4621	$p < 0.01$ effect size: 0.34
Appeal to fear (N = 287)	0.2366	0.6907	$p < 0.01$, effect size: 0.33	4.3118	4.5825	$p < 0.05$ effect size: 0.41
Name calling/ labeling	0.2895	0.3981	Not significant	4.500	4.4466	Not significant

(N = 179)						
Exaggeration (N = 396)	-0.1231	0.0299	$p < 0.05$, effect size: 0.44	4.0923	4.2786	Not significant
Flag-waving (N = 421)	0.3017	0.2764	Not significant	4.2671	4.5600	$p < 0.01$, effect size: 0.39
Others (N = 207)	0.0561	-0.2	$p < 0.05$, effect size: 0.59	4.1308	4.2900	Not significant

As indicated by these results, the annotation-based intervention proves most effective in aiding individuals to identify two specific persuasion techniques in news content: loaded language and appeal to fear. Participants not only demonstrate increased accuracy in detecting these techniques but also exhibit higher confidence in their judgments following the intervention. The intervention also aids in the identification of the exaggeration technique, although participants' confidence levels do not show significant improvement statistically. While the capability to identify the flag-waving technique does not exhibit notable improvement, participants display enhanced confidence in their judgments post-training. Interestingly, the annotation-based intervention appears less effective in aiding the identification of the name-calling technique, although participants already express considerable confidence in their ability to identify it.

In summary, our findings suggest that the annotation-based intervention can effectively enable participants to identify propagandistic texts and their corresponding techniques with greater precision and confidence, with the effectiveness varying depending on the type of technique.

4.2. Do people trust AI-generated annotations the same as those by experts?

We conducted ANCOVA tests and found no significant difference in participants' performance gain from the intervention (i.e., the difference between their accuracy and confidence in pre- and post-tests) across experiment conditions, while controlling for participants' familiarity with the excerpt topic. Additionally, we observed no significant difference in participants' trust towards the information they were presented with between the two intervention conditions. Therefore, we conclude that the intervention condition did not have a significant impact on our study. Based on this result, we combined data from both conditions for subsequent analysis tasks.

4.3. How are individuals' justificatory reasoning styles affected by the intervention?

When participants identified a persuasion technique used in the experiment, they were required to provide a reason to justify their identification. We conducted a thematic analysis on 1,703 collected responses regarding how individuals justified their identification of propagandistic texts. Three co-authors participated in the coding sessions. Initially, they independently coded 100 randomly selected identification instances and subsequently convened to discuss the coded results. Through an iterative process, they developed the coding schema (refer to Table 4). Following this, one of the co-authors and an undergraduate assistant coded another 100 randomly selected instances, achieving an inter-coder reliability score of 92%. The two coders then individually coded the remaining 1,487 instances and resolved discrepancies between them.

Table 4. Summary of the types of justificatory reasons provided by participants

Code Category for Justificatory Reasoning	Definition and Example
Content-based	Participants justified their decision focusing on the content, e.g., word choices, sentence structures, and/or aspects of language use that fall under the definition of a given persuasion technique (N = 1,373, 80.6%)
- Quoting words or phrases from the excerpt (N = 246)	*Identified text*: "I withdrew the United States from the terrible, very expensive, one-sided Paris Climate Accord." *Identified technique*: Loaded language *Reason provided*: "'terrible', 'very expensive'"
- Naming or providing definition of the technique (N = 247)	*Identified text*: "And you know that was all about the accord, because they were charging people tremendous amount of money and sending that money all over the world to countries that they never heard of , and the people got tired of it." *Identified technique*: Exaggeration *Reason provided*: "Exaggeration"

- Interpreting the excerpt content in relation to a technique (N = 880)	*Identified text*: "we've made ambitious investments in clean energy, and ambitious reductions in our carbon emissions. We've multiplied wind power threefold. We've multiplied solar power more than thirtyfold." *Identified technique*: Others *Reason provided*: "Repeating we've" over and over again"
Intention-based	Participants justified their identification by identifying the manipulative intentions of the authors from how the language was used in the content. (N = 228, 13.4%) *Identified text*: "we've made ambitious investments in clean energy, and ambitious reductions in our carbon emissions." *Identified technique*: Exaggeration *Reason provided*: "Using the word 'ambitious' to emphasize their determination about the word"
Ideology-based	Participants evaluated the content against their preconceived ideology on the focal issue being discussed in the excerpt, without referring to the content. (N = 102, 6.0%) *Identified text*: "punish our workers, our producers, and manufacturers" *Identified technique*: Exaggeration *Reason provided*: "No one was facing corporal punishment"

To examine the relationship between justificatory reasoning and participants' identification performances, as well as the effectiveness of our annotation-based interventions, we conducted a series of Mann-Whitney U tests for identification accuracy and confidence in pre- and post-tests.

We found that content-based justificatory reasoning was the most common strategy adopted by participants. When employing this strategy, participants also demonstrated high accuracy in the pre-test. We speculate that participants were able to grasp the general idea of the persuasion techniques before encountering the intervention materials and thus made intuitive decisions. In the post-test, we observed a significant increase in both their accuracy and confidence, indicating that the interventions enabled participants to recognize language signals in the content related to persuasion technique usage and become more confident in their justifications.

However, for the intention category, while there was a significant increase in identification accuracy, there was no change in confidence. This suggests that the annotation-based intervention helped participants better detect and understand the author's intention in relation to persuasion technique usage but did not necessarily increase their confidence in doing so.

Regarding the use of personal ideology to justify persuasion technique identification, participants tended to overidentify persuasion technique usage in the pre-test, but accuracy improved after the intervention. There was also an increase in participants' confidence when they used their personal ideology to justify their identification. These results suggest that the intervention enabled participants to better identify persuasion technique usage even when relying on their own beliefs rather than analyzing the material to justify their identification.

In summary, we observed a general increase in identification performance across different justificatory reasoning styles. However, the effectiveness of the intervention varied for different styles. Table 5 summarizes the test results.

Table 5. Comparison of pre- and post-test propaganda identification accuracy and confidence for different justificatory reasoning provided by the participants

Justificatory reasoning	**Accuracy**			**Confidence**		
	Pre-test Average	Post-test Average	Mann-Whitney U test	Pre-test Average	Post-test Average	Mann-Whitney U test
Content-based	0.1495	0.2762	$p < 0.01$, effect size: 0.45	4.1584	4.4205	$p < 0.01$, effect size: 0.41
Intention-based	0.0924	0.3945	$p < 0.01$, effect size: 0.38	4.3025	4.4404	Not significant
Ideology-based	-0.1915	0.1637	$p < 0.05$, effect size: 0.38	4.1915	4.4546	$p < 0.05$, effect size: 0.39

4.4. Does individuals' prior knowledge or belief mediate the training?

Media messages influence people's beliefs and biases (Banerjee & Kubey, 2013). People's existing beliefs and biases also influence how they evaluate the information (Pennycook et al., 2020), which may affect mediate the effectiveness of our AI-based annotation intervention. Our analysis shows no significant correlation between individuals' firmness of attitude regarding the topic in the content and their level of agreement with the excerpt content, indicating that participants' existing attitudes towards the topic did not significantly relate to their evaluation of the excerpt. Regarding the number of identifications participants made, we observed a significant negative correlation with the excerpt rating in the pre-test ($r = -0.19$, $p < 0.01$) and a positive correlation with topic attitude firmness ($r = 0.18$, $p < 0.01$). However, in the post-test, these correlations decreased and disappeared respectively.

We interpret these results in two ways. Firstly, we argue that participants' existing knowledge of the topic and their evaluation of the text influence how many instances of propaganda they tend to identify. If they agree more with the content, they may hold a more positive attitude towards the text and are likely to identify fewer instances of propaganda use; conversely, if they have high topic familiarity and firmness, they may be more critical of the content and identify more instances. Secondly, the decrease or lack of correlation in the post-test suggests that the intervention may have a leveling effect on participants with different pre-existing familiarity and attitudes towards the content. It is possible that the intervention enabled participants who are less familiar with the topic to be less reserved when evaluating the excerpts, while cautioning those who are more familiar or hold stronger attitudes to be less aggressive. However, our current study was not able to further test these hypotheses, and the results also suggest that other factors contribute to their identification accuracy. We suggest that future research further investigate the phenomenon we observed regarding participants' confirmation bias and intervention effects.

4.5. Is PRTA user-friendly and trustworthy?

We delved into the usability of the Prta system and participants' trust in the system, particularly in the backend AI models responsible for generating propaganda annotations. On average, participants rated the usability of the Prta system at 68.79 on the SUS scale (with a scoring range of 1-100; SD = 17.39), indicating a positive assessment of system usability. Additionally, they gave a rating of 3.64 (on a 5-point scale; SD = 0.51) for their trust in the propaganda classification algorithm used by the system, suggesting a perceived trustworthiness of the algorithm. These results suggest that incorporating AI technology, especially through the PRTA system, to support media literacy interventions can be a viable and well-received approach by users.

5. Discussion

Our findings indicate that while AI's performance on propaganda detection is inferior to that of domain experts, people's perception of AI today tends to be more trusting than doubtful. It is possible that the public is unaware of the technical challenges that need to be overcome in developing highly specialized AI as opposed to general AI technology.

Our study also provides empirical evidence supporting the use of AI-based annotation technologies to train people to become more critical thinkers. This implies that with competitive performance compared to domain experts, we can apply such technologies in many training contexts in the future, such as in formal classrooms that teach students persuasive communication.

While we have a sufficient sample size for each condition (39 for the human condition and 33 for the AI condition) according to the central limit theorem (Chang, Huang, & Wu, 2006), most participants were undergraduate or graduate students, rendering the sample not representative of social media users. Consequently, the generalizability of our findings and the effectiveness of the intervention for the public are in question. We speculate that for social media users who are less educated than those in our sample, a different representation of the instructional intervention may be more effective, such as the inclusion of visuals alongside textual instructions. As researchers on semiotics maintain, people can generate and interpret meaning from various forms of information (Mingers & Willcocks, 2014, 2017), one way to address this limitation is to design a study that explores the effectiveness of instructional materials with various representations.

Another limitation of the study is that our experimental materials do not feature as many instances of logical fallacy-based techniques as those of emotion appeal-based techniques. Logical fallacy-based techniques influence people's minds through faulty logic, such as black-and-white fallacy and causal oversimplification. Compared to emotion appeal-based techniques, which often exhibit strong language cues in the text (e.g., words of strong

emotions, repeated words, and phrases, etc.), these techniques may be harder to detect. While our intervention has shown how to help individuals identify propagandistic texts and persuasion techniques applied in them, it may not be as effective when dealing with logical fallacy-based propaganda. We thus call for future investigations to examine the effectiveness of instruction-based interventions in helping people identify logical fallacy-based propaganda.

6. Conclusion

In this study, we explored the potential of an AI-based annotation intervention to help people recognize manipulative texts and the corresponding persuasion techniques in online propaganda. Our experiment results showed that the intervention was effective. Furthermore, its effectiveness did not differ whether the participants were told the annotations were conducted by AI or by domain experts, suggesting the potential of AI-powered tools in training critical thinking. Our study also discovered that participants evaluate the materials based on three aspects: the content of the material, the author's intention they inferred from the material, and their personal ideology related to the topic of the material.

Acknowledgement

This project is funded by first author's professorship endowment fund.

7. References

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of educational research*, 70(2), 181-214.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology*: *general*, 149(8), 1608–1613.

Banerjee, C. S., Kubey, R. (2013). Boom or Boomerang: A Critical Review of Evidence Documenting Media Literacy Efficacy. In Angharad N. Valdivia & Erica Scharrer (Eds.), *The International Encyclopedia of Media Studies*, First Edition. Oxford, UK: Blackwell Publishing Ltd.

Beltramini, R. F. (1988). Perceived Believability of Warning Label Information Presented in Cigarette Advertising. *Journal of Advertising*, 17(2), 26-32.

Boyd, J. (2003). The Rhetorical Construction of Trust Online. *Communication Theory*, 13(4), 392–410.

Brooke, J. (1995). SUS - A quick and dirty usability scale. Usability Evaluation in Industry, 8. Bulger, M., & Davison, P. (2018). The Promises, Challenges, and Futres of Media Literacy. *Journal of Media Literacy Education*, 10(1), 1–21.

Chang, H. J., K. Huang, and C. Wu. (2006). Determination of sample size in using central limit theorem for Weibull distribution. *International Journal of Information and Management Sciences*, 17(3), 153-174.

Chen, S. J., Xiao, L. (2023). Automatic detection of Aristotle's persuasion strategies in the online misinformation content: a multi-label classification approach, *Journal of Information Science*, https://doi.org/10.1177/01655515231169

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction*, 39(9), 1727-1739.

Collins, C. (2004). *Education for a just democracy: The role of ethical inquiry*. PhD thesis, School of Education, University of South Australia, Adelaide

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79, 347-362.

Da San Martino, G., Shaar, S., Zhang, Y., Yu, S., Barrón-Cedeño, A., & Nakov, P. 2020). Prta: A System to Support the Analysis of Persuasion techniques in the News. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 287–293

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-Grained Analysis of Propaganda in News Article. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 5636–5646.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6), 261–262.

Hobbs, R., & McGee, S. (2008). Teaching about Propaganda: An Examination of the Historical Roots of Media Literacy. *Journal of Media Literacy Education*, 6(62), 56–67.

Hristakieva, K., Cresci, S., Martino, G. D. S., Conti, M., & Nakov, P. (2022). The Spread of Propaganda by Coordinated Communities on Social Media. *14th ACM Web Science Conference*, 2022, 191–201.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635.

Jowett, G. S., & O'Donnell, V. (2012). What is propaganda, and how does it differ from persuasion? In *Propaganda & Persuasion*, 1-48. Sage Publishing.

Knight, S. and Collins, C. (2006). Cultivating reason-giving: The primary purpose of education? *International Journal of Humanities*, 3(2): 187–194

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review*, 62(2): 155–178

Kuhn, D., & Udell, W. (2003). The Development of Argument Skills. *Child Development*, 74(5), 1245–1260.

Li, J., Ye, Z., & Xiao, L. (2019). Detection of Propaganda Using Logistic Regression. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 119–124.

Da San Martino, G., Cresci, S., Barron-Cedeno, A., Yu, S., Di Pietro, R., & Nakov, P. (2020). A Survey on Computational Propaganda Detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Survey track,* 4826-4832.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2), 337-359.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709–734.

Meyrowitz, J. (1998). Multiple Media Literacies. *Journal of Communication*, 48(1), 96–108.

Miller, C. R. (1939). The techniques of propaganda. From "how to detect and analyze propaganda," an address given at town hall. The Center for learning.

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, 1, 191-229. Academic Press.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, 2(2), 1-25.

Mingers, J., & Willcocks, L. (2014). An integrative semiotic framework for information systems: The social, personal and material worlds. *Information and Organization*, 24(1), 48-70.

Mingers, J., & Willcocks, L. (2017). An integrative semiotic methodology for IS research. *Information and Organization*, 27(1), 17-36.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11), 4944-4957.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review*, 23(3), 393–404.

Simons, H. W. (1976). *Persuasion: Understanding, practice, and analysis*. Addison Wesley Publishing Company.

Soong, H., Lee, R., & John, G. (2012). Cultural differences in justificatory reasoning. *Educational Review*, 64(1), 57–76.

Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence, Draft NISTIR 8332, National Institute of Standards and Technology, U.S. Department of Commerce, https://doi.org/10.6028/NIST.IR.8332-draft

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2, 59-89.

Tandoc Jr, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3), 381-398

Torok, R. (2015). Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. *Proceedings of Australian Security and Intelligence Conference*. 58- 65.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Weston, A. (2018). *A Rulebook for Arguments*. Hackett Publishing.

Xiao, Z., Zhou, M. X., Chen, W., Yang, H., & Chi, C. (2020). If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14