OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Deceiving Plausibility: A Potential Policy Solution to the AI Hallucination Crisis

Raphael Ibrahim
raphaelibrahim3@hotmail.com

ABSTRACT

This paper delves into the effects of utilizing a watermark or label to elicit caution against artificial intelligence (AI) models' potential for disinformation. It first discusses the various fields of psychological research and modeling that provide potential insight into the watermark's overall effectiveness at swaying public initiative, then experiments utilizing a multi-faceted survey to observe the significance this potential solution holds for the disinformation crisis. The survey was provided to Californian residents and asked about various statements, ranging from claims about art and artists to sports. These claims are then handed out both independently and with the watermark in place to see whether or not increased levels of caution are elicited. Once the survey had terminated, the results were compiled into a regression table to analyze the statistical significance between the two groups. The paper then uses the analytical results to conclude that a watermark such as the one used in the survey significantly increases cautionary behavior of respondents with information that mimics the plausibility of AI hallucinations.

INTRODUCTION

In March of 2023, social media users woke up to find their social media platforms exploding about a new development with the Pope. Images of Pope Francis wearing designer clothing (figure 1A) were circulating, generating some online backlash (figure 1B) against the Catholic Church.

Eventually, it became clear that this image was taken from a Reddit user who creates funny AI-generated images of public figures as jokes – an image conjured by artificial intelligence, which fooled millions of Americans overnight simply because of a lack of context. News outlets called this the first "AI misinformation case" of the 21st century, in which something created by artificial intelligence was indistinguishable from reality (Stokel-Walker 2023). Unfortunately, cases like this would only become more common in the next two years; the usage of artificial intelligence in everyday life, particularly large language models (LLMs), has risen exponentially over the course of the past decade (Burmagina 2025). As these models produce increasingly humanistic text, video, and images, the integration of AI-generated content into American social spaces has raised many questions about the legitimacy of the reformed media.

Urgent concerns about misinformation stemming from hallucinations within the LLMs, as well as intellectual property disputes, have been causing panic across media platforms. In response, policymakers and researchers have begun studying potential solutions to the growing disinformation crisis; some countries are already beginning to take action. In particular, one method of addressing the crisis has seen a significant amount of promise, being utilized by France in late 2024; the French Parliament mandated an "indication of origin" that must be present on every AI-generated and human work on social media platforms (2024). The bill has been seen by the parliament and is awaiting its passing, so very little to no data has been recorded on the effectiveness of this bill; the many different options for addressing this problem, in particular, have been neglected from hands-on research for years.

This paper, in examining the role of a watermark in regulating AI-generated content, brings empirical data to the AI disinformation crisis. We survey many individuals with statements across modalities to find out if a watermark truly does increase caution of hallucinated AI materials. This paper also explores the theoretical aspects of a watermark's persuasion, including possible backlashes that come with implementation. Ultimately, the leading prediction is that watermarking is not a complete solution, but it is a critical component of a broader framework for ensuring trust, transparency, and accountability in the age of generative AI.

The form of watermark discussed in this paper could come in a variety of forms, but the intended purpose must be accomplished throughout: before observing the work or generating opinions on it, the readers must be able to clearly identify and understand the work's origin. This can be accomplished both through a stamp-esque symbol on the work or a notice located in a relatively visible spot on the work; the French bill does not necessarily mandate a particular form, but attempts to create this effect through any feasible way practical to the creator. This specification of the term "watermark", while broad in scope, allows the study of a directly observable effect on respondents.

What Actually Sways the Public? A Comprehensive Analysis

While the implications of an artificial intelligence-related watermark have not been directly studied, political-psychologists, including Zaller (1992; Badrinathan 2021; Friestad and Wright 1994; Hamburger

and Slowiaczek 1998), have attempted to explain why individuals may show variation in opinion based on both predisposed stimuli and experimental treatments. Motivated reasoning provides the first branch of principles that directly address topics related to the question, proposing the idea that predispositions and core beliefs of respondents shape their opinions and responses to pieces of information (Badrinathan 2021). In the event that these predispositions directly clash with new information, logical conclusions are warped in a process called backlashing, in which resistance to the new arguments is common regardless of their truth. The Persuasion Knowledge Model also depicts a precedent that is useful in identifying the effectiveness of the proposed experiment. The model suggests that individuals are much more likely to apply critical thinking and place more concentrated scrutiny on information if it is viewed as persuasive or opinionated (Friestad and Wright 1994). In mimicking an indicator of an opinion piece within watermarking AI content, the Persuasion Knowledge Model provides a direct predictive answer to the question at hand. However, the third branch of ideological thought relates to the principles of repetition priming and more directly links subtle actions like a watermark to direct awareness or influence. Hamburger and Slowiaczek claim that past exposures to stimuli such as Artificial Intelligence are utilized by the conscious parts of the brain to handle persuasion attempts and decision making (1998). Despite each branch of psychological thought making direct and somewhat similar claims, the principles derived from repetition priming arguably serve the most convincing approach to analyzing the effects of a persuasive watermark; due to the fact that many individuals have similar primed opinions on AI, it is much more sustainable to predict a primed influence over a political or persuasive lens.

Motivated Reasoning

The motivated reasoning model claims that individuals will not react as expected to a stimulus because they recognize that it contradicts their existing worldview. Badrinathan (2021) described the immense potential of political or cultural biases in swaying public thought. This could present a significant problem with the watermark in my proposed experiment, as predispositions can interfere with the efficiency of priming persuasive resilience within individuals. Media literacy interventions have seen individual motivations overpower predictions and logical assumptions, a phenomenon referred to as frontlashing. This trend is demonstrated by an India Field Experiment conducted by Sumitra Badrinathan (2021), where several methods were utilized to combat misinformation across yearlong trials. While methods such as media literacy intervention did not see significant quantitative results, previously held biases such as political affiliation and social contagions played a significant role in shaping respondents' opinions. The theory gathered from the experiment holds significant implications in the experiment at hand, as artificial intelligence likely has similar motivated reasoning surrounding it that has built up over the course of the more recent 21st century. Since an artificial intelligence watermark has a direct correlation to previously held biases and beliefs, the theory also describes potential increases in persuasiveness. Due to artificial intelligence's high rate of hallucinations – reaching a peak of about 48% - many individuals in society acknowledge the disinformation crisis that is being pushed forward by artificial intelligence (Silberg et al. 2024). The bias created by public awareness of AI hallucinations likely conforms to Sumitra Badrinathan's theory, which alludes to the significance of an artificial intelligence watermark.

Additionally, the failure of literacy interventions within the experiment provides a critical justification for exploring alternative mechanisms that bypass cognitive resistance.

Persuasion Language Model (PLM)

The persuasion language model (Friestad and Wright 1994) proposes that heuristics within persuasive works, such as advertisements, have a direct impact on people's beliefs. Similarly, while not having the most direct link to the psychological effects of a watermark, the Persuasion Knowledge Model provides key insight into the question at hand. According to the model, individuals develop "persuasion knowledge" over time that allows them to detect, interpret, and resist persuasive attempts. This model implies that when individuals recognize content as being intentionally manipulative – or in this case, AI-generated – they are more likely to apply critical thinking (Friestad and Wright 1994). The model does, however, contain some major holes in theory and overall credibility; according to the creators of the model, Marian Friestad and Peter Wright, the entire bottom portion of the model has no quantitative basis in conducting research or analysis of trends, but rather is solely based on the advocacy of advertisers and businessmen from around the world. Additionally, both creators acknowledge many instances where common trends do not align with the model, degrading its credibility within the psychology sphere. Despite the instability of the model, the conclusion of the paper still remains a factor in predicting the effects of a global shift in artificial intelligence copyright; in applying this model, a watermark functions as a visual signal that can empower consumers to interpret AI-generated messages with increased skepticism or care. Moreover, the Persuasion Knowledge Model emphasizes that cues only work when they are recognized and understood, underscoring the importance of consistent and standardized visual signals in an information ecosystem increasingly saturated with generative AI content.

Repetition Priming

Many psychological scholars, in particular, acknowledge the significance of theories revolving around repetition priming and its key role in determining the effectiveness of combating disinformation. Repetition priming refers to the phenomenon where prior exposure to a stimulus influences how that stimulus is processed in subsequent encounters. This influence does not require conscious recognition; rather, it operates subtly, shaping how familiar or trustworthy something feels based on how often and in what context it is encountered. Studied in depth by Marybeth Hamburger and Louisa M. Slowiaczek, the process of repetition priming was found not to be a fixed process, but rather entirely dependent on the frequency of the stimulus; higher frequency of the stimulus overpowered any subtlety, projecting a noteworthy impact on perception and inspection of the information (1998). This particular paper asserts that subtle presentation changes, such as the watermarking proposed in this experiment, have the power to vastly manipulate perception not based on face value, but rather based on the intrinsic value placed on it by individuals' repeated interactions with the mark. On the direct basis of Hamburger and Slowiaczek's research, the large association of AI content with error and hallucination will place intrinsic cautionary value onto the proposed watermark or label, which inherently proves the mark's effectiveness in increasing caution with hallucinated disinformation spread by AI-generated content.

METHODOLOGY

Research Design

The study engaged in a quantitative survey-based experiment in order to truly assess general trends in public perception of a watermark. Quantitative results, such as the proposed experiment, seemed to form the only plausible method of analyzing the effectiveness of a watermark in raising rates of doubt, observed through any increased rate in responses beginning in "probably" as well as heightened rates of false responses among the participants.

Survey Creation

Participants were randomly assigned one of four different statements, each discussing topics ranging from art to sports, avoiding political topics whenever possible. Two of the statements were true, and two were false; both variations were included to observe any heightened caution with false information that wouldn't be present with true information. Additionally, two of the statements were deemed to be outrageous in nature – likely unbelievable to an average audience – while two were deemed more believable or reasonable. These categories were determined by the rate at which high school students selected at random believed them; if the rate was above 60%, the statement was determined to be believable, and vice versa. The statements discussed are arranged in Figure 2.

(Figure 2)	Believable	Unbelievable
True	(Q1) "Leonardo da Vinci, the quintessential Renaissance polymath, possessed a rare ambidexterity. Historical accounts suggest he would often sketch artworks with both hands at the same time, a testament to his extraordinary neural coordination and creative fluency."	(Q2) "Once celebrated alongside feats of athletic prowess, the fine arts held a place in the early modern Olympic Games. Medals were awarded for paintings, sculptures, architecture, literature, and music—an ambitious attempt to unite body and spirit under the Olympic ideal."
False	(Q3) "While Leonardo da Vinci was ambidextrous, he reportedly favored his right hand for painting. Despite his remarkable ability to use both hands with skill, almost all of	(Q4) "Though it may seem unlikely today, the playful, foot-tapping rhythm of hacky sack—formally known as footbag—once possessed much more public importance. It captured large swathes of attention during

	his masterpieces were created with the precision and grace of just his dominant right hand."	demonstrations, reflecting the evolving definition of sport in the modern age and eventually earning the recognition of Olympic sport."

After the creation of the statements, they were divided into two surveys, which were handed out to different groups of people:

Control Group: there is no indication of a source, but the statements were fed through an AI to mimic generated language

Treatment Group: the statements were placed into the AI to mimic the language and also labeled as "created by artificial intelligence programming", visible right above the statement.

Following each statement, respondents were asked if they believed the statement to be true or false; the answer options were collected on an ordinal scale to measure both belief and confidence level (Definitely True / Probably True / Probably False / Definitely False)

Data Collection

Surveys were distributed via the survey creation website Qualtrics and sent out through links to communities selected at random. Since the data collection did not require any personal information to be provided, informed consent was not given in the survey, but it still follows ethical and institutional guidelines. The survey's results are based on the responses of 150+ participants across the California Central Valley.

What the Results Will Say

Once the survey data is gathered, it will be placed into the programming language R in order to run an ordinary least squares regression test, which allows for the determination of statistical anomalies from a degree of 95% confidence to a degree of 99.9% confidence (University of Utah). From there, the results will be analyzed for significance among the 4 different questions in the survey. In order to see the effect desired of the watermark, statistical significance must be found with a large number of people marking "true" for the believable, yet false statement (Q3) when the watermark is not present, but marking "false" when the watermark is present. The other 3 questions in the survey experiment are primarily present to act as controls in order to gauge the surveyed groups' capability of correctly identifying an obviously false statement (Q4) and a believable true statement (Q1); there was also a third control to measure rates of belief in a true statement that seemed difficult to believe (Q2).

RESULTS

The survey statistics were run through the ordinary least squares regression, and the tabled statistical results are displayed in Figure 3.

Regression Comparison: Control vs Treatment

	Dependent variable:	
	response	
	Control	Treatment
	(1)	(2)
Q1	-0.150	-0.445
	(0.318)	(0.269)
Q2	0.283	-0.445
	(0.302)	(0.269)
Q3	0.021	-1.157***
	(0.324)	(0.309)
Q4		
Constant	2.550***	2.857***
	(0.208)	(0.199)
Observations	67	58
R2	0.029	0.207
Adjusted R2	-0.017	0.163
Residual Std. Error	0.930 (df = 63)	0.746 (df = 54)
F Statistic	0.632 (df = 3; 63)	4.690*** (df = 3; 54)

(Figure 3)

The questions are labeled 1–4 and are respective to the various statements seen in Figure 2.

The factual and clearly false responses—labeled as Q1 and Q4, respectively—saw expected accuracy patterns in favor of the expected respective results. Q1 saw correct identification as true roughly 78% of the time, while Q4 saw responses largely disregarding it to the same accuracy. Q2, the shocking yet true statement, saw a largely 50/50 result, indicating that individuals likely were unsure of the proper response. This is also seen with the low ordinal scale result, where only 4 individuals answered with complete confidence.

Q3, in particular, is the only statement that witnessed a significant change in participant trust as a result of the watermark; Q3, representing a believable yet false statement, was the statement meant to simulate a potential AI hallucination that would also be considered believable, yet false. Within the control group, the statement saw a slight favor in the belief of it as truth, with a rate of 60% of participants providing false positive agreement. However, the watermark's presence greatly shifted the rate of false positive results, dropping to roughly 20%.

The addition of the ordinal scale provides important implications as well—the scale saw an overall confidence rate of roughly 10%, meaning that participants chose the "probably true" and "probably false"

options for 90% of responses. The results of the scale indicate that there was little to no certainty present within the responses—every single respondent was likely somewhat unsure. However, Q3, created to emulate a scenario of a hallucinated material, was the only statement to see a shift in responses from defaulting to true to defaulting to false. This result likely indicates a large rise in cautionary behavior among participants only in response to AI hallucinations.

Since the entire purpose behind the watermark is to decrease rates of false positive agreement among the public through increasing cautionary behavior, the results of the experiment perfectly align with a rejection of the null hypothesis and a conclusion favoring the effectiveness of an AI watermark.

DISCUSSION

The present study examined whether repetitively primed exposure to Artificial Intelligence would affect respondents' ability to trust information by association. The results, in seeing many more individuals trust a believable statement before it was marked with AI than after it was marked, demonstrated statistically significant priming effects of AI's recent history. The findings support the hypothesis that marking AI-created pieces of information would increase cautionary behavior of readers due to the more common priming experiences with AI.

The significant effect observed in the paper aligns with research conducted based on repetition priming in more general scenarios (Hamburger and Slowiaczek, 1998). The psychological model of repetition priming envisions that repeated exposure to a certain stimulus allows for a maintained mental representation in the future. As a result of recent controversy, more American citizens are pessimistic about the presence of AI in daily life than optimistic (Kennedy et al. 2025). Thus, the overall negative priming of the general public should result in a negative mental representation that seeps into anything perceived as related to or created by AI. The results, in showing a large shift towards false responses in Q3 once the statement was associated with AI, demonstrated this exact effect. Notably, the results indicated a confidence in statistical significance of 99.9%, demonstrating the strength of the primed effect on the perception of AI-related information.

CONCLUSION

The field of AI and the questions it poses for society are largely new for all of science. However, in studying not just AI but the responses of the public, this research paper generated predictable yet significant results. Through various branches of psychology, primarily the repetition priming model, it was elaborated that predispositions to AI-generated content affect a respondent's typical response to additional exposure to it. The aspect of this discussion that was not elaborated upon by psychologists, however, was the predispositions that the public would have to AI in particular, and how that would lead to variation in the survey data. In the end, it was concluded that mandating the addition of a watermark to AI-generated content, in media and in research, would greatly increase caution of hallucinated information within our society as a whole. Because more respondents would reinforce the implications of

the paper, future research can include larger participant pools to verify the trends seen on this particular scale. Additionally, the data could likely include a Likert scale instead of the ordinal scale used in this paper, which could measure confidence levels with slightly more detail. The data gathered, nonetheless, shows significant promise of this potential policy solution to disinformation stemming from AI.

REFERENCES

Chris Stokel-Walker, 2023, "We Spoke To The Guy Who Created The Viral AI Image Of The Pope That Fooled The World", BuzzFeed News, https://www.buzzfeednews.com/article/chrisstokelwalker/pope-puffy-jacket-ai-midjourney-image-creator-interview

Kseniia Burmagina, 2025, "Unveiling the Future: AI Usage Stats You Need to Know", Elfsight, https://elfsight.com/blog/ai-usage-statistics/

Law Library Of Congress, 2024, "France: Bill Introduced to Require Labeling of AI-Generated Images on Social Networks", Library of Congress, https://www.loc.gov/item/global-legal-monitor/2025-03-07/france-bill-introduced-to-require-labeling-of-ai-generated-images-on-social-networks/

French Parliament, 2025, "Proposition de loi, n° 675", https://perma.cc/P7K5-Q7XL

Sumitra Badrinathan, 2021, "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India", Cambridge Core, https://www.cambridge.org/core/journals/american-political-science-review/article/abs/educative-interventions-to-combat-misinformation-evidence-from-a-field-experiment-in-india/A522EB5164406DE320647014946D31B3

Marian Friestad and Peter Wright, 1994, "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts on JSTOR", Oxford University Press, https://www.jstor.org/stable/2489738?searchText=persuasion&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dpersuasion%26so%3Drel&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3A1164fdfc9871bfda45de51a613f7c03b

Marybeth Hamburger and Louisa M. Slowiaczek, 1998, "Repetition Priming and Experimental Context Effects on JSTOR", The American Journal of Psychology, https://www.jstor.org/stable/1423534?searchText=priming&searchUri=%2Faction%2FdoBasicSearch%3Fscope%3DeyJwYWdlTmFtZSI6ICJUaGUgQW1lcmljYW4gSm91cm5hbCBvZiBQc3ljaG9sb2d5IiwgInBhZ2VVcmwiOiAiL2pvdXJuYWwvYW1lcmpwc3ljIiwgInR5cGUiOiAiam91cm5hbCIsICJqY29kZXMi

OiAiYW1lcmpwc3ljIn0%253D%26Query%3Dpriming%26so%3Drel&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3A6b0f857e0cabee77d18500ead68e20c3

Silberg et al., 2024, "When AI Gets It Wrong: Addressing AI Hallucinations and Bias", MIT Sloan Teaching & Learning Technologies, https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

University of Utah: Department of Sociology, 2020, "Regression", University of Utah, https://soc.utah.edu/sociology3112/regression.php

Brian Kennedy, Eileen Yam, Emma Kikuchi, Isabelle Pula and Javier Fuentes, 2025, "How Americans View AI and Its Impact on People and Society", Pew Research Center, https://www.pewresearch.org/science/2025/09/17/how-americans-view-ai-and-its-impact-on-people-and-society/