Department of Digitization (DIGI)
M.Sc. in Business Administration and Data Science

Master's Thesis

FINAL EXAMINATION PROJECT

Analyzing the Effects of AI-Generated Images and Labels on the Believability of News Headlines

---

Author:
JENO TOTH

Supervisor:
ROB GLEASURE

September ???, 2024

Character count: ???
Word count: ???
Pages: ???

Keywords: AI-Generated Images, AI labeling, Fake News, Experiment, Galvanic Skin Response, Eye-tracking, NeuroIS

Abstract

This thesis explores the impact of images generated by Artificial Intelligence (AI) and AI labels on the believability of news headlines. With the increasing risks of AI being used to create fake news, it is crucial to understand how consumers perceive such news. The research explores how AI-generated images and labels warning of the use of AI affect believability. A 2x2 experimental design with repeated measures was used in a laboratory setting, not only surveying perceived believability but also measuring bodily metrics with the use of eye-tracking, Galvanic Skin Response (GSR), and heart rate monitoring. Two key theories, cognitive dissonance, and the illusory truth effect, were applied to frame the research. Results indicate that AI-generated images decrease believability, while labels have mixed results. Post hoc analysis also looked at the role political and AI attitudes play and the potential dissonance that they cause.

Acknowledgements

I would like to thank Rob Gleasure for supervising this thesis, providing invaluable insights and supporting me continuously throughout the process. I would also like to thank Angelika Uta Kensy Tziatziou for her relentless helping in the laboratory as well as Lotte Lie Duestad and Hanne Celine Foss for their collaborative work. I would also like to thank all who spent time participating in the experiment. Lastly, I would like to thank my partner, my family, and my friends for their unending encouragement and support.

Table of Contents

1 Introduction

Over the past decade, fake news and misinformation became a known issue, weakening social trust and democratic institutions among other problems (Morgan, 2018). While the phenomenon is not new, with the spread of fake news dating back centuries, its prevalence became widespread with the newfound ability for masses to share information (Burkhardt, 2017). Social media platforms made it possible for the average person to spread (mis)information in seconds. Despite this, many users treat information found here as factual. In recent years, the percentage of Americans who used social media regularly to get news increased to over 50% (Aïmeur et al., 2023), and while platforms have tried to combat the spread of fake news, oftentimes these attempts did not change much (Moravec et al., 2018). The effect of fake news can be observed in events such as the 2016 United States presidential election (Bovet and Makse, 2019) or the COVID-19 pandemic (Kim et al., 2021). Numerous studies focused on the characteristics of fake news, focusing on the speed of its spreading (Vosoughi et al., 2018) as well as its believability (Pennycook and Rand, 2017). As we will see, these two features are correlated (Moravec et al., 2018), suggesting that the more believable a fake news article is, the faster it will spread on social media.

With the advent of Artificial Intelligence (AI), the way fake news is created has further evolved. Tools such as Large Language Models (LLMs) can generate text based on prompts that are indistinguishable from human-written content. While still in its early phases, AI image generators can already create photo-realistic images. Such methods to generate content have a marginally lower cost than alternatives, and can often be used without any restrictions. Social media bots have been utilized previously to spread misinformation, but their capacities were limited (Vosoughi et al., 2018). With these new advancements, it is expected that their influence will grow (Kolomeets et al., 2024). There are a number of reasons for using such bots, as they can be both

monetarily profitable (Kshetri and Voas, 2017) and politically influential (Park et al., 2024) while having low costs (Martens et al., 2018).

An important research topic in this realm is analyzing the use and effect of labels. After some backlash, social media platforms began implementing limitations to posts in order to combat the spread of misinformation (Martens et al., 2018). The most common way they achieve this is by flagging fake news with a label, warning users of disputed claims. While this is a step in the right direction, its effects are unclear, with no change in user behavior in some cases (Moravec et al., 2018). In recent developments, Meta has also taken the step to label AI-generated content on Facebook and Instagram (Meta, 2024). Currently, this is done in collaboration with users in order to increase transparency, therefore it does not influence malicious accounts' behavior.

1.1 Purpose and Research Question

While previous research has focused on how AI-generated images (Kolomeets et al., 2024) and warning labels (Martens et al., 2018) influence the believability of posts on social media platforms, these questions have not been combined. The purpose of this thesis therefore is to analyze how AI-generated images and labels signalling the use of AI images influence the believability of news headlines. In order to answer this, laboratory experiments were conducted at Copenhagen Business School's NeuroLab. A survey containing 48 news headlines was created, with a 2x2 design with repeated measures of what participants see for each headline. The first condition was whether the image they see is the real image associated with the news headline or an AI-generated image created with a prompt. The second condition either showed a label warning of an AI-generated image or the headline was not labeled. Participants had to rate how believable, truthful, and credible they found each headline. During the experiment, their emotional responses were measured with the use of eye-tracking, Galvanic Skin Response (GSR), and heartbeat sensors.

The two theories the research applied were cognitive dissonance (Festinger, 1957) and the illusory truth effect (Hasher et al., 1977). The research was constructed in a way that include fake news as well as political topics. Indeed, half of the news headlines participants read were fake news (based on independent fact-checkers), 16 were liberal-favorable, while another 16 were conservative-favorable (as well as including 16 non-political headlines as a baseline). Cognitive dissonance has been shown to be an influential theory in fake news research. In line with the belief-disconfirmation paradigm (Festinger et al., 1956), labeling fake news has been shown to be less efficient in groups who agree with the content of the news they see. The illusory truth effect has been shown to be especially prominent in social media. A single exposure of information is enough to increase believability, even if it is not true (Pennycook et al., 2018). On social media, information flows quicker than in traditional media, which prompts consumers to use mental shortcuts when evaluating the truthfulness of the news they see. This leads them to believe in information they have previously seen (Fazio et al., 2015).

The thesis continues with six chapters. Chapter 2 will take a look at the literature surrounding fake news, AI, and research methodologies using eye tracking and GSR. Chapter 3 will overview the theories of cognitive dissonance and illusory truth effect, formulating the hypotheses to be answered by the thesis. Chapter 4 will describe the preliminary steps taken to set up and conduct the experiment, as well as outline the methods of the data analysis. Chapter 5 focuses on the results of the experiments, focusing on the hypotheses as well as post hoc analysis in order to fully utilize the data available. Chapter 6 will discuss the findings and connect them to the literature. Finally, Chapter 7 concludes the thesis, summarizing the findings.

2 Literature Review

2.1 Fake news

2.1.1 *Defining Fake News*

In some form, fake news has been around long before social media, however, it has gained especially significant importance in the last decade (Burkhardt, 2017). Despite this, there is no single clear definition of what "fake news" is. Shu et al. (2017) provides a detailed overview of how different literature distinguishes it, as well as attempts to define the concept. The two main features that they prioritize in their paper are authenticity and intent, and based on this, they define fake news to be "a news article that is intentionally and verifiably false". This definition treats satire, rumors, and even misinformation differently from fake news, emphasizing how the earlier examples are not necessarily done with malice. This definition also gives way to clearly identifying fake news due to its verifiability, in contrast to some conspiracy theories, for example (Shu et al., 2017). This is important later on when we talk about algorithms used to detect fake news.

Zhou and Zafarani (2020) provides two definitions, one of which is similar to the one by Shu et al. (2017). According to this, "Fake news is intentionally false news published by a news outlet" (Zhou and Zafarani, 2020). In both definitions above, intent plays an important role. However, Zhou and Zafarani (2020) argues that with the large number of people sharing information that we have seen in recent years, it has now become hard to distinguish between maliciousness and misinformation. Furthermore, many fake news datasets upon which research builds also mainly focus on verifiability, leading to a second, broader definition: "Fake news is false news" (Zhou and Zafarani, 2020). Throughout this thesis, this definition will be used.

2.1.2 *Fake News in Literature*

Fake news is not something that only started appearing in the digital age. Indeed, Burkhardt (2017) traces the evolution of fake news from pre-printing press eras, where misinformation was controlled by the elite, to the digital age, highlighting the profound shift in the spread and impact of fake news facilitated by the internet. The paper emphasizes how the democratization of information and economic incentives for sensational content have aggravated the proliferation of fake news, highlighting the challenges in distinguishing between real and fake news in the current media landscape. This historical perspective can improve our understanding of fake news as a complex phenomenon that already existed before technological advances, offering insights into the persistent nature of misinformation and its implications for society.

Burkhardt (2017) evaluates how social media and the internet have amplified the spread and impact of fake news, examining technological and social factors that contribute to its proliferation. Before the internet era, the paper categorizes people into two groups: those who had widespread access to sharing information with the masses, and those who had not. The first group became prominent with the invention of the printing press and had a monopoly on information spreading even in the 20th century via mass media. Meanwhile, the majority of the population could only spread information via word of mouth and letters, only reaching a few others at a time. With the internet, this gap between the groups decreased, and finally, with social media, it almost completely diminished. While this is an advancement in the democratization of information, it also created the ability for an average person to share misinformation with the masses. Overall, this resulted in more fake news reaching people, as well as the creation of networks that could polarize its members.

The paper concludes by emphasizing the critical need for media literacy and the potential roles of technology and individual responsibility in combating fake news. The author summarizes trends in the two main ways fake news can be combated by

algorithms, fake news detection and bot network detection. They also list a number of practices that individuals could apply to be more aware, such as learning how search engines and algorithms work, and how to evaluate information, emphasizing the need to teach students to be more aware of this phenomenon (Burkhardt, 2017).

Numerous studies have been conducted about specific aspects of fake news. Kim et al. (2021) summarized a chunk of the literature by identifying 2,277 fake-news-related scientific articles from 2020 and selecting close to 200 of them randomly to have a comprehensive and representative sample of the current literature. The paper summarized trends of fake news creation, both its methods and reasons. One of the main catalysts identified was the transformation from passive to interactive news consumption, which led to increased emotions and polarization of consumers (Bowman and Willis, 2003). The other facilitator Kim et al. (2021) identified was the misuse of AI technology. Even before the widespread use of Large Language Models (LLMs), social bots were used to generate and spread text (Shao et al., 2018), while image generators were already used to alter images and videos. The analysis of Kim et al. (2021) not only identified the facilitators but also the purposes of creating fake news. Most of the literature focused on ideological purposes such as influencing elections (Berghel, 2017), but it was also shown how fake news could be used for monetary purposes and panic reduction. Regarding monetary purposes, Kim et al. (2021) presents literature on how creators use fake news to generate excess website clicks, therefore generating more revenue (Kirby, 2016). Fear and panic reduction were a significant reason too, mainly due to the COVID-19 pandemic. Here, fake news that showed alternative, usually unscientific, methods to combat illness spread at a fast rate, since it was comfortable for readers to believe their contents (Lavorgna et al., 2018).

Kim et al. (2021) also summarized methods for identifying fake content. They presented both qualitative methods, such as literature focusing on identifying fake news by its grammatical content and punctuation (Ruchansky et al., 2017), and quantitative methods such as training machine learning models to identify AI-tempered images

(Gupta et al., 2013). Finally, the paper looked at what can be done to combat the phenomenon. They identified four key directions for future research, such as the development of socio-technical models, a deeper understanding of fake news consumption behavior, preemptive decision-making and action support, and increased education to improve digital media literacy.

Lazer et al. (2018) address the complex nature of fake news and the need for a multidisciplinary response. The article delineates fake news as fabricated content mimicking legitimate news formats without adherence to journalistic integrity. Highlighting the erosion of trust in traditional news media, the study underscores the complexity of combating misinformation in the digital landscape. It explores the historical context of journalistic norms, the rise of the internet as a double-edged sword for information dissemination, and the nuanced challenges posed by the proliferation of fake news. The authors propose empowering individuals through critical media literacy and advocating for structural changes to mitigate fake news exposure, emphasizing the critical role of internet platforms in shaping the information ecosystem.

Martens et al. (2018) analyzed fake news through the lens of digital transformation. Their analysis found the same results as above, confirming that fake news spreads faster and generates more advertising revenue because of this. They identify that this phenomenon can cause a market failure, as generating and spreading fake news has a low cost with this increased marginal revenue. Here, the emphasis on content editors is decreasing while the role of news distributors increases to maximize clicks. This, in turn, decreases the credibility of all editors, increasing uncertainty in the news realm. To counter this, and evade the market failure, the paper looks at possible policies to intervene. While the authors suggest that there may be some self-correction in the market, such as Facebook introducing methods to combat fake news in order to attract advertisers, the paper also suggests giving increased resources to fact-checkers, and increasing cooperation between them, as well as having more initiatives to increase media literacy.

It may be puzzling to see how fake news can spread so effectively. Kopp et al. (2018) proposed an information-theoretic model to capture what effect a small portion of deceivers have on social systems. They used a game-theoretic background with an evolutionary Iterated Prisoner's Dilemma (Axelrod and Hamilton, 1981) serving as the platform, where agents could evolve their strategies over time. Multi-generational simulations were conducted, where initially the majority of the agents were cooperative, and only a small portion were deceivers; these deceivers were modeled to act as "fake news spreaders". Multiple simulations were conducted, with different costs to deception (Kopp et al., 2018). While the simulations were run only on 50 agents, the authors argue that this can represent real life, where the majority of people are just passively observing smaller groups who actively partake in debates (Campan et al., 2017). The paper found that even a small portion of initial deceivers could shift a large fraction of the cooperative population to be exploitative, which can shift the equilibrium of cooperatives and exploitatives further in later stages. This shift, however, largely depended on the cost of deception. Kopp et al. (2018) showed that while deceivers are effective at low costs, they struggle when it is more difficult to deceive an agent. In the case of fake news, this cost comes in two forms. First, if the cost of spreading information on social media increased (which is currently close to zero), spreading fake news would be more difficult. Second, cooperative agents in the simulation only punished exploitative agents in the event that the deception failed. Instead, a strategy where all information is analyzed more thoroughly by consumers could help slow down the spread of fake news.

Kshetri and Voas (2017) also applied a theoretical framework to the economics of fake news, where they based their research on the costs and benefits of engaging in the creation of fake news. The benefits identified were monetary and psychological, while the costs are direct investment costs, opportunity costs, and psychological costs, as well as the expected cost of being arrested or convicted. From this basis, the paper looked at how these values can be influenced, aiming to result in higher costs than benefits for malicious agents. Kshetri and Voas (2017) identified three groups of agents: consumers,

creators, and arbiters. Similarly to previous research, they found that more informed consumers and higher costs of spreading fake news can tip the balance. Arbiters were categorized into three categories, social, economic, and legal (Wiesenfeld et al., 2008). These groups all could decrease the speed of news spread for fake content. Social arbiters, such as PolitiFact, can flag content, giving consumers more information. Economic arbiters, such as social media companies can improve fake news detection on their platforms and increase punishments for their spreading, however, this is not monetarily beneficial for them, since it results in lower rates of engagement and thus lower revenues. Finally, the paper urges law agencies to give the combat of fake news higher priority, increasing the expected cost of conviction for fraudsters. Kshetri and Voas (2017) sets clear guidelines for a number of different agents in their framework. Since its publication, a number of steps in the right direction have been taken, such as the improvement of detecting and flagging fake news by social media platforms (Pennycook and Rand, 2017).

2.1.3 *Fake News and Social Media*

Figueira and Oliveira (2017) outlines the state of fake news on social media, focusing on detection approaches categorized into content-based, source-based, and diffusion-based. The paper first walks through the case of a browser extension that looks through the user's Facebook feed, labeling content to be verified or not verified. This example showcases how content- and source-based algorithms work. The text of a post (as well as the texts in images) can be validated via search engines to see if they came from a reputable source. Similarly, links can be directly checked, to see if the sources are reputable.

In the second part of the paper, Figueira and Oliveira (2017) looks at how a diffusion-based algorithm can be built, and what challenges do these algorithms face. The proposed high-level algorithm takes in four inputs: the content text, the poster ("Who"), the location based on the region of the poster and the locations mentioned in

the text ("Where"), and the time of the posting ("When"). From the text, Natural Language Processing (NLP) methods are used to find the topic. These points of information are combined to create a "Fact", which is then checked against trusted sources in that topic, to see if a piece of content can be validated or not. While this method can help raise awareness of some fake news, it has a number of drawbacks. Comparing content to a "trusted source" assumes that there is a way to establish whether a source is reputable or not. For this, they recommend a scoring system that is up-kept by trustful agents, and using this score to validate. The paper also emphasizes the importance of free speech, where, for example, satire content could be flagged as "not validated", worsening its appearance even though its main goal was not to spread information. Because of these issues, Figueira and Oliveira (2017) argues that while algorithms are necessary to combat large amounts of fake news, human oversight is important as well.

One of the most notorious examples of widespread fake news was the outcome of the 2016 US presidential election. Bovet and Makse (2019) looked at the effect of fake news on Twitter regarding this event, which can serve as a demonstration of fake news on social media. They used a dataset of 30 million tweets that contained links to news websites, categorized these sites based on volume, and determined their leaning and whether they were fake via independent fact-checkers. Users were grouped based on what types of news articles they shared, and the habits of different user groups were compared. The paper concludes that between 12-15% of the news shared was fake or extremely biased. By looking at the timestamps and news shared, they identified bots and found that bots sharing fake news were more active than those sharing other news sources. Regarding political stances, Bovet and Makse (2019) found that while major center and left-leaning accounts were the ones influencing center and left-leaning users, this causality was flipped and the behavior of small right-wing accounts influenced the activity of major right-leaning fake news-spreading accounts.

Regarding social arbiters, Moravec et al. (2018) conducted an experimental analysis

to see how independent fact-checkers can present their results in order to achieve maximum effect. The paper builds on Kahneman (2011)'s System 1 and System 2 cognitions. System 1 cognition is understood to be a quick judgment, especially prominent when reading news on social media, while System 2 is a more deliberate process requiring cognitive effort. Moravec et al. (2018), in an experimental setting, looked at how flags in social media aimed at either System 1 or System 2 (or both) influence the believability of news, as well as the likelihood that the reader will engage further with it.

After reading through the first set of news headlines, participants were informed in an awareness training about a "stop sign" and its meaning, indicating that the news they were reading might not be completely true. Moravec et al. (2018) intended this sign to be a flag for System 1 thinking, while another method, a text saying "Declared Fake by 3rd Party Fact-Checkers", intended to affect System 2 cognition. As a control, Facebook's flag system was also included, which has been previously shown to have little effect on believability (Pennycook and Rand, 2017). Finally, for some headlines, participants saw both flags for System 1 and System 2 cognition at the same time. The paper found that in all cases, after the awareness training, participants were less likely to believe news with the same flags as before. Furthermore, both of Moravec et al. (2018)'s flags were more effective than Facebook's implementation, and they achieved the largest effect with the combination of the two, resulting in participants being more skeptical about believing and engaging with fake content.

To look at the exact speed fake news spreads, Vosoughi et al. (2018) conducted a data analysis of the spread of true and false news on Twitter, looking at 126,000 stories tweeted by 3 million people over a decade. Their findings reveal that false news spreads significantly more rapidly and widely than true news across all types of information, particularly false political news. When controlling for external factors, the paper found falsehoods to be 70% more likely to be retweeted than true content. This extensive diffusion of falsehood is attributed not to bots, but to human nature's propensity for

novelty and emotional involvement, with the paper concluding that while accounts managed by bots are an important reason for the phenomena, they are not the primary drivers of the spread of misinformation.

Since Vosoughi et al. (2018) showed that not only bots influence the difference in the spread of true and fake news, Ali Adeeb and Mirhoseini (2023) looked at the impact of affect on humans when perceiving fake news via a systematic literature review. While they found that previous research highlighted the need to investigate the role of emotions within misinformation research (Lewandowsky et al., 2012), only a few papers focused on this. Ali Adeeb and Mirhoseini (2023)'s analysis concludes with three avenues for future research. These are the effect consuming of fake news on affect and behavior, the relationship between emotion and the belief in fake news, and the study of specific emotions in the context of consuming and sharing fake news.

One paper looking at emotions, Serrano-Puche (2021), examines affective polarization in social media, looking at how emotional content influences individuals. The paper finds that not only are groups formed and divided based on discourse, but emotions are further deepening this division. Emotions also drive engagement through interactions such as likes and comments, meaning that emotional content has a further reach. Martel et al. (2020) also showed a positive correlation between emotion levels and belief in fake news. Based on these trends, Serrano-Puche (2021) concludes with the importance of improving media literacy, especially informing social media users of the role emotions play in their behavior.

2.2 Artificial Intelligence

Artificial Intelligence (AI) has rapidly evolved in recent years from a field of computer science to a cornerstone of modern business and technology, influencing both industry and academic disciplines (Mariani et al., 2023). It had a transformative impact on fields

such as healthcare, finance, and education, and is considered one of the key drivers for businesses in the digital age (Păvăloaia and Necula, 2023).

Due to its expansive nature both in its influence and in its different forms, it is difficult to pin down an exact definition for Artificial Intelligence. Sheikh et al. (2023) overviews the history of AI and its relation with society in order to come up with a definition. The paper describes the challenge of separating the definition of AI from being based on human intelligence, arguing that the latter is also not clearly defined and that the research of the two concepts has been intertwined. AI is an evolving field, where in the past fixed rules were used, but with the advancements of Machine Learning and Deep Learning models, this changed into pattern recognition in data, thereby changing what AI is. Sheikh et al. (2023) concludes that AI is best defined by the European Commission's High-Level Expert Group on Artifcial Intelligence (2019) as "systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy - to achieve specific goals".

There are a number of different applications for Artificial Intelligence, the two main areas of focus being prediction and generation. AI can be used to predict, among others, stock market prices (Vijh et al., 2020), diseases (Ghaffar Nia et al., 2023), and weather Dewitte et al. (2021). These complex models use vast amounts of past data, both including the dependent and the explanatory variables, in order to infer the dependent variable in the future based solely on the independent inputs (Annor Antwi and Al-Dherasi, 2019). Generative AI on the other hand is a method of pre-training models on large datasets in order to be used in real-time to generate new information. It can be used for fast content creation with interfaces that enable users to input a wide array of commands (Gupta et al., 2024).

2.2.1 *Training AI Models*

To gain a deeper understanding of how AI works, it is advantageous to take a look at how models are created. Training AI models is a multi-step process that begins with the collection and preprocessing of large amounts of data relevant to the problem domain, which serves as the foundation for the model's learning capabilities (Geron, 2019). Once prepared, the data is used to train the model using various algorithms that allow the AI to learn patterns, relationships, and structures within the data. This training process can involve supervised learning, where the model is trained on labeled data, or unsupervised learning, where the model identifies patterns in unlabeled data (Geron, 2019). Additionally, reinforcement learning may be used, where the model learns to make decisions by receiving rewards or penalties based on its actions. Throughout this process, the model adjusts its internal parameters to minimize errors and improve its predictive accuracy (Geron, 2019).

Machine Learning (ML) and Deep Learning (DL) are two critical subsets of AI training, each offering unique capabilities in model training and operation (Kalota, 2024). ML, which focuses on the development of algorithms that enable machines to learn from data, is often employed in predictive analytics to make informed decisions or predictions based on historical data. ML techniques include supervised learning, where models are trained on labeled data, and unsupervised learning, where they identify patterns without predefined labels (Geron, 2019). Deep Learning, a specialized subset of ML, involves the use of artificial neural networks with multiple layers—referred to as deep neural networks—that enable more complex and abstract data representations. DL is particularly effective in tasks such as image and speech recognition, where traditional ML techniques might struggle. The strength of DL lies in its ability to automatically learn features from raw data, making it a powerful tool for developing sophisticated AI models capable of handling large-scale and complex data (Sarker, 2022).

In state of the art AI models, transformers are a deep learning model architecture

that has revolutionized various fields such as natural language processing NLP, computer vision, and audio processing. Originally introduced by Vaswani (2017), transformers utilize a mechanism known as self-attention, which allows them to weigh the importance of different elements in an input sequence, enabling the model to focus on relevant parts of the data for a given task. This architecture consists of an encoder and a decoder, each made up of layers that include multi-head self-attention mechanisms and position-wise feed-forward networks. The encoder processes the input sequence to generate a set of key-value pairs, which are then used by the decoder to produce the output sequence. The key innovation of transformers is their ability to handle long-range dependencies in data without relying on recurrent or convolutional layers, making them highly efficient and powerful for tasks that involve sequential data (Lin et al., 2022).

2.2.2 *Generative Artificial Intelligence*

Generative AI is a subset of AI that has seen the biggest increase in both relevance and capabilities over the past years. Gupta et al. (2024) thoroughly summarizes current research trends in the topic, focusing on both technical aspects such as image generation and Generative Pre-Trained Transformers (GPTs), as well as data privacy and security. GAI models are both capable of training on a single type of data (such as text, images, or audio), or on multimodal methods involving different data types connected. Single modal methods allow for example text generation based on prompts (Iqbal and Qureshi, 2022), or alteration of images (Himeur et al., 2022). The second approach allows for the creation of tools that are used in everyday life, such as text-to-speech (Kaur and Singh, 2023) or text-to-image generators (Zhang et al., 2021).

Kolomeets et al. (2024) explores current Generative AI's place in social media by looking at how bots who use AI-generated images perform compared to traditional bots and humans. The paper finds that such bots are quicker to engage, which is a disadvantage for them as this results in significantly quicker detection and therefore being blocked. The paper controls for image content, separating groups based on, for

example, if there is a face on a photo, and if so, is it real, photoshopped, or AI-generated. Since the bots' image usage is connected to their other characteristics, Kolomeets et al. (2024) does not control for algorithms. Since bots that utilize AI images are relatively new, their algorithms are less developed, leading to the aforementioned worse performance. The paper finishes by concluding that the market for social media bots is hesitant to switch to bots that use AI images, both because of their worse performance and also because of the decreased level of trust that was seen with such bots.

Another paper examining AI deception focuses less on images and more on AI models' ability to deceive (Park et al., 2024). The paper summarizes current literature on both a wide range of AI applications' abilities to win strategic games as well as highlighting how large language models can deceive humans. The paper showed that when instructed, AI models can successfully be used for malcontent and that newer models are improving in this regard. The paper highlights 3 risks: "fraud, political influence, and terrorist recruitment" (Park et al., 2024). Regarding political influence, the paper showcased how AI can be used to generate and spread fake news, as well as directly influence politicians.

2.2.3 *AI Ethics*

Despite its rapid development and widespread adoption, AI also presents several challenges, particularly in the realms of ethics, data privacy, and security (Gupta et al., 2024). The capability of GAI to produce highly realistic content raises concerns about the potential for misuse, such as the creation of deepfakes and misinformation (Mirsky and Lee, 2021). The reliance on large datasets for training these models brings to the forefront issues related to data privacy and the ethical use of information. Initiatives have been taken in order to mitigate this, such as the introduction of Federated Learning (FL) (Konečnỳ et al., 2016). FL aims to enable multiple organizations to collaborate in Machine Learning, thus increasing transparency over data usage. While such an initiative can help companies collude, due to the large barrier to enter ML model

training, this solution might still result in a lack of clarity for individuals.

2.3 Experimental Research

Ross and Morrison (2003) provides an overview of experimental research as well as outlining its methods. The paper briefly summarizes the main focus of experiments, creating standardized procedures in order to maximize internal validity. This allows experimenters to control for external factors, which in turn leads to the assumption that the difference between two groups (usually a control and a treatment group) is due to the effect of the treatment. Ross and Morrison (2003) distinguishes 5 types of experimental designs: True Experiments, Repeated Measures, Quasi-experimental Designs, Time Series Design, and Ex Post Facto Design. This thesis uses Repeated Measures. The design administers all treatments to all subjects, allowing for both between subject and within treatment analysis.

Keselman et al. (2001) looks at different statistical methods associated with repeated measures. The paper describes the requirements for conventional univariate tests, as well as proposes alternatives such as multivariate tests and univariate tests with adjusted degrees of freedom. The latter option is a robust option in case group sizes are equal, and allow for easy interpretation of the outcomes. As we will see later in Chapter 3, groups for the analysis are either separated by exogenous setup (half the articles are fake, there are 16 liberal and 16 conservative-leaning articles), by the random scenarios (such as headlines randomly being assigned either a real or an AI-generated image), and by artificial groups (political and AI attitudes will be used to split the participants into two groups based on the median person). All of these leave group sizes to be equal in every analysis, therefore the univariate tests with adjusted degrees of freedom (Keselman et al., 2001) will be used when analyzing the results.

2.3.1 *Eye-tracking*

Apart from survey data, bodily signals such as Eye Tracking, Galvanic Skin Response (GSR) and heart rate will be analyzed in order to quantify emotional response. Vasseur et al. (2019) reviews eye-tracking literature in Information Systems (IS) showcasing an increased trend in its usage. The paper also identifies the most important metrics, such as Fixation, Fixation Count, and Fixation duration, as well as describing a number of other variables used in research, such as Mood analysis and Pattern recognition. Vasseur et al. (2019) also summarizes the findings of other papers, mentioning increased fixation when unconventional stimuli is present.

Joseph and Murugesh (2020) encapsulates a number of metrics regarding eye tracking, highlighting their uses in analysis. Some of these, such as measures focusing on eye movement and pupil dilation, are not relevant to our research, since participants are tasked with reading, a monotonous task. The paper describes areas of interest as predefined parts of the screen that can be used for later analysis to categorize where participants focused on the most. Heat maps, which can show this in more detail post-test, may also be considered, however, the paper emphasizes that they are most advantageous when it is unclear what is expected of subjects' behavior beforehand. Since the areas of interest such as the news title and the headline image can be set in advance, these will be used throughout the analysis. Joseph and Murugesh (2020) also details fixation metrics that will be used such as time to first fixation and dwell count.

2.3.2 *Galvanic Skin Response*

Galvanic Skin Response (GSR) is used to detect emotional valence and arousal Sanchez-Comas et al. (2021). It measures surface resistance on the skin by passing through a microcurrent of electricity through a pair of nodes. The nodes are connected to the palm or fingers since they are densely populated with sweat glands. In a time series analysis, it is able to detect changes caused by sweating and epidermis, connecting

them to the stimuli participants are faced with.

Paul et al. (2020) analyzes in detail how GSR can be used to recognize emotion. The paper highlights that GSR cannot be used to detect a specific emotion, rather it focuses on the intensity of emotions. As such, in research, it is recommended to focus on one specific area and see how emotional levels change in that circumstance. Paul et al. (2020) also recommends using GSR in sync with other measures such as heartbeat rates and body temperature in order to increase data quality.

3 Theory

3.1 Cognitive Dissonance

Cognitive dissonance, a term first coined by Festinger (1957), describes the psychological discomfort experienced by individuals when they hold two or more conflicting beliefs simultaneously. This discomfort is a significant psychological tension that arises when there is an inconsistency between one's beliefs, attitudes, or behaviors. Festinger's theory has since become a cornerstone in social psychology, providing a framework for understanding the processes by which people strive to achieve internal consistency and resolve conflicts within their cognitive framework (Harmon-Jones and Mills, 2019).

Festinger (1957) explains cognitive dissonance through the example of smoking. An informed person who smokes understands that it is bad for their health, all the while still smoking. Festinger (1957) lists four mechanisms which can diminish dissonance. The first is the change in behavior. In our case, this means that the person stops smoking, knowing that it is unhealthy for them. Secondly, they may change their thought, believing that smoking is not harmful, or has positive effects. Third, they might add additional behaviors or cogitations, such as believing that while smoking is unhealthy, they are living a healthy life otherwise (new behavior) and accidents can also cause harm (new cognition). Finally, Festinger (1957) describes trivialization, where the importance of the conflict is downplayed, reducing dissonance.

Cooper and Fazio (1984) revisits the concept of cognitive dissonance, expanding on the theory by incorporating the nuances of personal responsibility and foreseeability in the arousal of dissonance. The paper emphasizes that cognitive dissonance arises not merely from holding contradictory beliefs but significantly from engaging in actions that lead to adverse outcomes, for which the individual perceives personal responsibility. This

responsibility is intricately linked to the foreseeability of the consequences of one's actions. Cooper's work advances our understanding of dissonance by detailing how the interplay of decision-making, perceived control, and anticipation of outcomes contributes to the experience of dissonance, offering deeper insights into the mechanisms driving attitude change and behavior modification.

Harmon-Jones and Mills (2019) provides an overview of the four main paradigms in dissonance research that aim to examine the effects and consequences that receiving inconsistent information has on individuals. These are the Free-Choice Paradigm, the Belief-Disconfirmation Paradigm, the Effort-Justification Paradigm, and the Induced-Compliance Paradigm. The Free-Choice paradigm is concerned with exploring the dissonance individuals experience after making decisions. In a number of cases, if someone is confronted with having to choose between two alternatives, they are likely to explore both the positives and the negatives of each option and make a choice based on this. This, however, results in the subject knowing the negative aspects of their choice, as well as knowing the positive aspects of the option they did not pick, causing dissonance (Harmon-Jones and Mills, 2019).

The Effort-Justification Paradigm examines the dissonance that arises when individuals voluntarily engage in an unpleasant activity to achieve a desirable outcome (Harmon-Jones and Mills, 2019). The dissonance arises from the conflict between the cognition that the activity is unpleasant and the cognition that the individual willingly engaged in it. To resolve this dissonance, individuals often justify the effort they put in by exaggerating the desirability of the outcome. In an experimental setting, Aronson and Mills (1959) found that women who underwent a severe initiation to join a group subsequently rated the group as more interesting and worthwhile than those who only experienced a mild initiation.

The Induced-Compliance Paradigm explores the dissonance that arises when individuals are asked to engage in behavior that contradicts their prior beliefs, especially

when they perceive that they have freely chosen to engage in this behavior. In these cases, the individuals experience dissonance because their behavior is inconsistent with their prior beliefs. To reduce this dissonance, individuals may change their attitudes to align more closely with their behavior (Harmon-Jones and Mills, 2019). This paradigm was first used in Festinger and Carlsmith (1959). In their experiment, participants were paid a sum of money to lie about the enjoyment of a boring task. Counterintuitively, those paid a small amount experienced greater dissonance and reported actually enjoying the task more than those who were paid a larger amount. This was explained by the paradigm, stating that the insufficient justification for lying led them to alter their attitudes to resolve the dissonance.

In the realm of fake news research, the most notable paradigm is perhaps the Belief-Disconfirmation Paradigm. It describes the event and the consequences where an individual or a group has a prior belief which is then disproved beyond doubt, and how a group may still preserve their prior beliefs afterward (Harmon-Jones and Mills, 2019). An example of the paradigm was described in Festinger et al. (1956). A group was studied who believed that a flood would be imminent. This message was delivered by a woman to whom the "prophecy was supposedly transmitted by beings from outer space". The experiment examined what happened when the flood did not happen. Those who were alone at the time did not fall victim to the paradigm, however, subjects who waited in groups continued to do so. These people were then told by the woman that she received a message that the flood was stopped because of the group's existence. While before this the group did not engage in proselytizing, after the message they started reaching out to others (Festinger et al., 1956).

Belief can mean that even after being disconfirmed by facts, individuals can still resonate with their beliefs, forming groups with the same misinformation. These groups can use the facts opposing their views to further increase their beliefs. Just by a single person voicing the rejection of the fact, the whole group will become more likely to reject it too (Wolters et al., 2021).

Figl et al. (2019) explore the efficacy of fake news flags in influencing the believability of social media posts. Their study indicates that while such flags can affect perceptions, their effectiveness is nuanced, interacting significantly with the source's reputation. This suggests that the credibility of information on social media is complex, influenced by both the content's presentation and the user's preconceptions about the source. This research underscores the importance of considering psychological factors like cognitive dissonance when designing interventions to combat misinformation online.

3.2 Illusory Truth Effect

The concept of illusory truth was first introduced by Hasher et al. (1977). The paper showed true (e.g.: Lithium is the lightest of all metals) and false (e.g.: The People's Republic of China was founded in 1947) statements were proposed to subjects who had to rate their credibility. The tests were repeated 3 times in a 2-week interval, with some of the questions (both true and false) repeating in the sessions, while others were only seen once. The paper finds that for both true and false remarks, repetition significantly increases the credibility of plausible statements.

Pennycook et al. (2018) tested the extent of the illusory truth effect specifically with regard to fake news articles. They also find, similarly to Hasher et al. (1977), that a single prior exposure is enough to increase perceived credibility. Their research demonstrates that even a single exposure to fake news can significantly increase perceived credibility, a phenomenon that persists across time and is robust even when the news articles are identified as potentially disputed or are in conflict with the reader's political ideology. This highlights the potent influence of repetition on belief formation, underscoring the challenges in combating misinformation in an era dominated by social media. The findings suggest that interventions aimed at correcting misinformation, such as fact-checking labels, may not be as effective as hoped, particularly in mitigating the

effects of repeated exposure to false claims. This work not only contributes to our understanding of cognitive biases but also offers critical insights into the dynamics of news consumption and belief formation in contemporary society, pointing towards the need for more nuanced strategies to foster critical media literacy and reduce susceptibility to fake news.

Fazio et al. (2015) looked at the constraints of the illusory truth effect. While previous research indicated that knowledge of topics and facts can alleviate the effect, Fazio et al. (2015) found in an experimental setting that participants would often rely on fluency instead when evaluating the truthfulness of statements. The paper conducted two experiments among university students using a number of known and unknown statements, both true and false ones. In the first experiment, subjects first had to indicate their interest in the topic of the statement that they were presented with. After this, they had to rate a new set of statements (with some overlap from the first phase) based on how truthful they found them. At the end of the experiment, the subjects' knowledge was tested and they had to answer the statements in a multiple-choice setting. The paper found that in all cases, for both generally known and unknown statements, for both true and false statements, and for both statements to which subjects knew the answer and to which they answered incorrectly, prior exposure (meaning that they saw the statement in the first stage) increased the truth rating they gave in the second stage.

In their second experiment, Fazio et al. (2015) used the same statements in a different setting, but instead of a scale to rate truthfulness in the second stage, a binary option was given. In the case of known true statements, there was no significant increase in the proportion rated "true", but in all other cases, there was. Both experiments led to the paper's conclusion that even when subjects knew the right answer to a statement, they often ended up with "knowledge neglect" and relied on heuristics to answer.

Brashier et al. (2020) looked at the prominence of the illusory truth effect in an experimental setting among young (university student) participants. Four experiments

were conducted in a similar fashion to Fazio et al. (2015), with two phases in each: an "initial exposure" phase and a second "truth rating phase". Subjects saw either factually true statements ("The fastest land animal is the cheetah") or altered false statements ("The fastest land animal is the leopard"). In the initial exposure phase, participants had to rate either their interest in the topic of a statement or the truthfulness of it. In the second phase, they were then prompted to act as fact-checkers and analyze the truthfulness of another set of statements (some were the same as they saw earlier, while some differed).

Brashier et al. (2020) found that if subjects were prompted to rate the truthfulness (instead of their interest) in the first phase, they were able to judge the accuracy of statements better in the truth rating phase in sentences they saw before. This was prominent when there was a two-day delay between the phases as well, leading to the paper's conclusion that simply having a mindset of asking "Is this true?" when seeing new information can help alleviate the illusory truth effect when seeing that same information later on.

Newman et al. (2020) built on the previous two papers introducing two additional factors, the Need for Cognition (Cacioppo and Petty, 1982) and truthiness (Newman et al., 2012). Need for Cognition (NFC) is a metric that describes how much individuals like thinking and engaging in cognitive tasks. The reason for this inclusion is to see if the previous results are homogeneous, or if there is a difference between those who have high NFC compared to those who do not. Truthiness explores how a nonprobative photo influences the believability of a statement, showcasing that the attached image increases believability even though it does not pose additional information regarding the statement.

In a number of experiments Newman et al. (2020) replicated both truthiness and the illusory truth effect. Regarding NFC, the paper found only significant results in one case, in an experiment where subjects did not receive warnings about the possibility of false

statements. In this case, people with high NFC were shown to be more prone to illusory truth effects, leading to the conclusion that these participants might have had higher cognitive fluency during the experiment. In all other cases, however, Newman et al. (2020) found no significant difference between subjects with high and low NFC.

Other theories were also considered for this thesis. Signaling theory explores how individuals perceive signals from other individuals or organizations (Spence, 1974). It is often used in evolutionary biology (Hasson, 1997), as well as organizational theory (Reuer et al., 2012). In signaling theory literature, results are often interpreted with a high focus on quantitative analysis, which can omit to look deeper into the cognitive mechanisms that individuals apply when coming across signals (Drover et al., 2018). Since the aim of this thesis is both to quantify the effects of AI content as well as gain a deeper understanding of the cognitive processes that influence news consumers, cognitive dissonance and the illusory truth effect were chosen as the leading theories instead of signaling.

3.3 Hypotheses Development

The study will mainly look at the effects of AI-generated images and AI labels with respect to a number of external factors, such as political attitudes, fake news, and attitudes to AI. Based on previous literature and theory, the following hypotheses are derived.

$\boldsymbol{H_1}$: The use of AI images decreases believability

Previous studies have established the connection between the believability of a news headline and the speed it spreads. Vosoughi et al. (2018) showed that content spread by bots reaches fewer people on average. Regarding images, Kolomeets et al. (2024) showcased that when comparing bots, those who use AI-generated images are less trusted and less efficient. Overall, it is expected that all things held equal, news

headlines with AI-generated images will receive lower scores on the survey.

$\boldsymbol{H_2}$: AI labels decrease believability when real images are present

Moravec et al. (2018) has previously showcased the ambiguous effects of labels on social media posts. AI-generated images are still often distinguishable from real images Abduljawad and Alsalmani (2022), therefore using labels when participants believe to see real images might be counter-effective, implying deception Kopp et al. (2018) and resulting in lower believability scores.

$\boldsymbol{H_3}$: AI labels increase believability when AI images are present

Fazio et al. (2015) has shown that in cases where users have to consume large quantities of information, such as reading news on social media, they use mental shortcuts in order to come to a decision on whether or not they believe a given piece of news. When AI-generated images are shown, seeing the label might increase trust in the overall scenario, and incite participants to believe the news more.

$\boldsymbol{H_4}$: Both AI images and AI labels increase attention and emotional response.

Vasseur et al. (2019) provided a summary of how unconventional stimuli increase attention in eye-tracking research. Since both AI-generated images and AI labels are relatively new concepts for most people, they are expected to be unconventional and therefore increase attention. Regarding GSR metrics, Lyulyov et al. (2024) has shown that AI content received increased emotional response in the realm of marketing research, leading to the expectation that AI-generated images will have similar effects in the context of news headlines.

While the hypotheses build on previous research, their outcomes might not be trivial. For some, it might be possible to argue against the opposite effect. For example, AI-generated content might appear more sophisticated due to the technological advancements associated with it. Based on this argument, the opposite of $\boldsymbol{H_1}$ might be true. Similarly, AI labels may prove to increase skepticism in all cases, resulting in the

opposite of $\boldsymbol{H_3}$. Due to this uncertainty, it is important to conduct the experiment by focusing on the above hypotheses throughout the analysis.

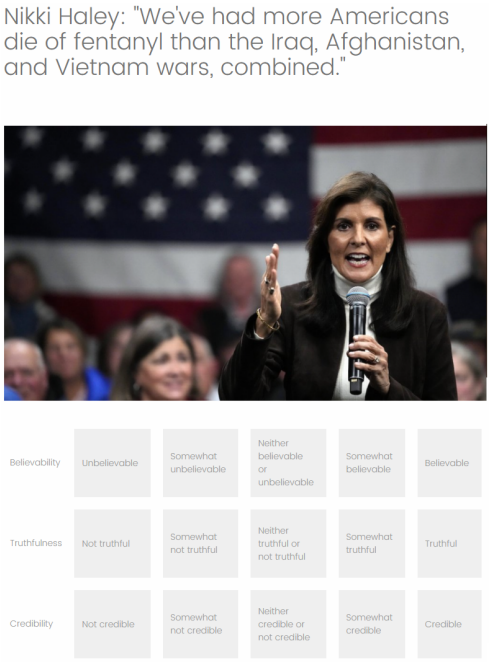
4 Methodology

4.1 Research Framework

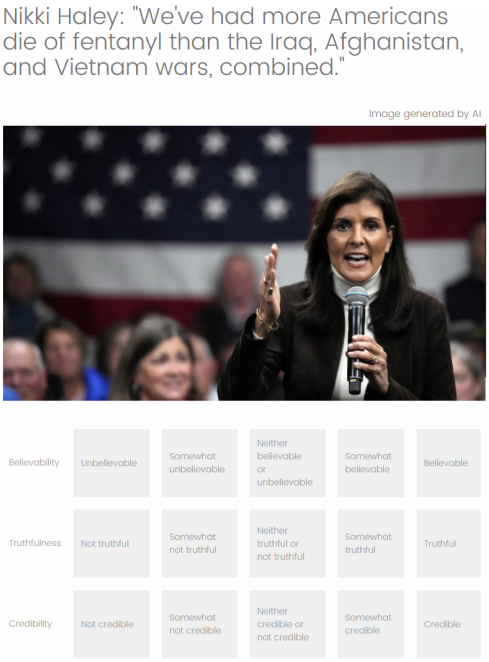
The purpose of this paper is to establish a connection between the use of AI images in news headlines and the believability of these headlines. In order to create data to answer the research questions, surveys were conducted. A similar survey was run in an online setting as well as in a laboratory with tools aimed at measuring body metrics. The online preliminary study served as a baseline, as well as showcasing differences between demographics since the laboratory study was conducted on a smaller, less diverse sample.

4.1.1 *Survey Design*

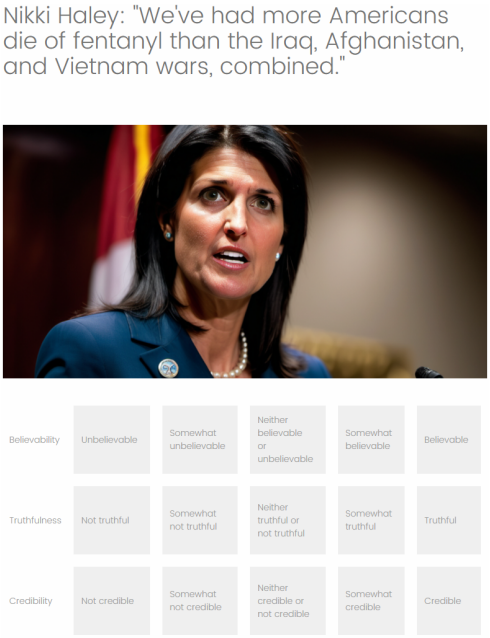
To test our hypotheses, a survey was created where participants saw 48 news headlines in a continuous manner, together with an associated image. This survey then was slightly altered in order to be conductible both online and in a laboratory setting. The experiments adopted a 2x2 design with repeated measures. The first condition manipulated was whether the image was real or AI-generated. All articles had a related headline image, these were used as the "Real" images, while AI-generated images were used as the "Fake" alternative. The second condition manipulated was whether the news headline was presented with a warning "Image generated by AI" (note this meant subjects also rated headlines with real images, labeled as generated by AI). This second condition was adopted to see if labeling images in news articles affects their credibility. For each headline, subjects were asked to score the believability of that headline using the three-item scale from Beltramini (1988) and Moravec et al. (2018). The following images show the 4 potential scenarios a participant could see for a headline (the complete set of news headlines and images can be seen in Appendix A):



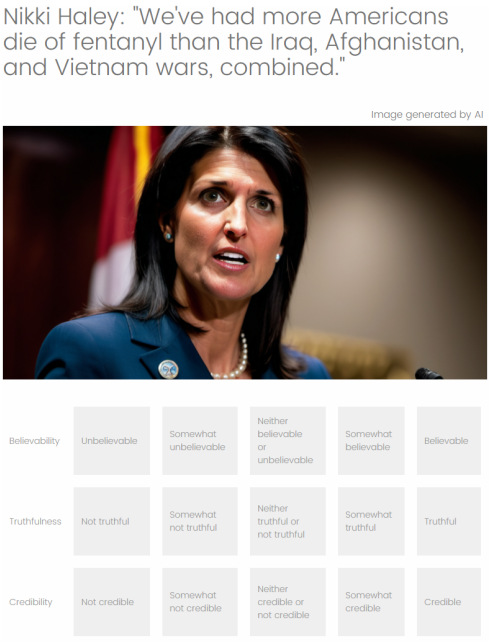
(a) Real image with no label



(b) Real image with label



(c) Fake image with no label



(d) Fake image with label

Figure 1: 2x2 design for a specific news headline

Both experiments also had an introduction and a final questionnaire, while the online experiment also had an attention check. The introduction page can be seen in Figure 2.

Figure 2: Introduction page for all participants

The final questionnaire was based on Pennycook et al. (2018). First age, gender, nationality, occupation, and whether the participant wore vision correction during the study were asked. Then, the political preferences were examined, with two prompts: "On social issues, I am..." and "On economic issues I am...", each having a 5-scale Likert Scale from "Strongly liberal" to "Strongly conservative". Participants were then asked to indicate which news sources they use, they could select multiple of the following: "Online news websites", "Social media platforms", "Newspapers (print or digital)", "Podcasts", "Friends or family word-of-mouth", "Online forums", and "Other, please specify". Finally, participants' views on AI and their impulsivity were measured on a 5-scale Likert Scale from "Strongly disagree" to "Strongly agree" (Turel and Kalhan, 2023). The following statements were proposed: "Using AI is a good idea", "Using AI is a wise idea", "Using AI is a desirable thing", "Using AI is beneficial", "I say things

without thinking", "I spend more money than I mean to", "I am often impatient" and "I make 'spur of the moment' decisions".

4.1.2 *Article Selection*

The news articles used were gathered from PolitiFact (https://www.politifact.com/, last accessed: 06/19/2024). PolitiFact manually looks at verifiable statements in a number of news articles, as well as in tweets and claims from politicians. Based on their truthfulness, they receive a rating between "True", "Mostly True", "Half True", "Mostly False", "False" and "Pants on Fire". 24 of the news articles used were either "True" or "Mostly True", while 24 were "False" or "Pants on Fire". Regarding political standings, 16 of the articles were left-leaning, 16 were right-leaning, and 16 were not based on politics (all three categories had 8 true and 8 fake articles). Since PolitiFact is mainly concerned with United States politics, the majority of the articles used in the experiments were also related to them. After considering the time it would take for a participant to read through 48 whole articles, only the headlines were left as the texts for the experiments. This mimicked the way news articles are displayed on social media platforms.

4.1.3 *Image Creation*

After testing a number of image-generating AI software, MUSE AI (miramuseai.net, last accessed: 08/18/2024) was used for generating the images for the news articles. This platform builds on Midjourney AI's models. These models use transformative architecture, trained on text-labelled images in order to create new images based on text inputs (Derevyanko and Zalevska, 2023). While other state-of-the-art models such as DALL-E or Stable Diffusion might have resulted in more realistic images, these software have restrictions regarding political impersonations (Abduljawad and Alsalmani, 2022). For each news article, the used website received the following prompt: "Generate a headline image for the following news article:" followed by the news article's headline.

The model was set to generate "Realistic" images. 4 images were generated for each article, in order to avoid issues (such as multiple people having the same face generated, or other features being distorted), before selecting one to be in the experiment.

4.1.4 *Preliminary Experiments*

Two surveys were conducted on Qualtrics, an experience management platform (https://qualtrics.com/, last accessed: 08/25/2024). The first questionnaire was not concerned with the articles, only with how realistic the AI-generated images look. Participants were presented with 48 images in a continuous manner, one randomly selected for each headline. A statement appeared below the images, saying "This image has been tampered with Artificial Intelligence (AI)". Subjects had to select a value on a Likert-scale between 1 and 7, where 1 means "Strongly disagree", and a score of 7 means "Strongly agree". Participants also faced an attention test between the 28^{th} and 29^{th} headlines. The test consisted of an image they previously saw with the text "Attention test! Select "Strongly agree"!". This was done in order to enhance data quality by eliminating participants who were not fully paying attention. The result of this survey was then used to not only have a binary variable for whether or not an image is AI-generated but to allow insights to see if the difference in believability changes as the difference between a real and an AI-generated image diminishes.

The other online survey was one adapted from the experiment described in Section 4.1.1. The main addition here was the inclusion of an attention test in a similar matter as described above, where participants faced the following prompt: "Attention test! Select "Believable", "Truthful" and "Credible"!". The resulting data of this survey was in a similar structure as the laboratory experiment, allowing a feasible comparison of the two.

4.1.5 *Laboratory Experiment*

A number of measurements were taken in order to ensure participants were not agitated before the experiment and not disturbed during it. Windows were kept closed and the blinds rolled down before their arrival. A small movable wall was put between the participant and the experimenter in order to avoid distractions and to keep the answers anonymous. The scene participants were faced with upon entering the lab can be seen in Figure 3.

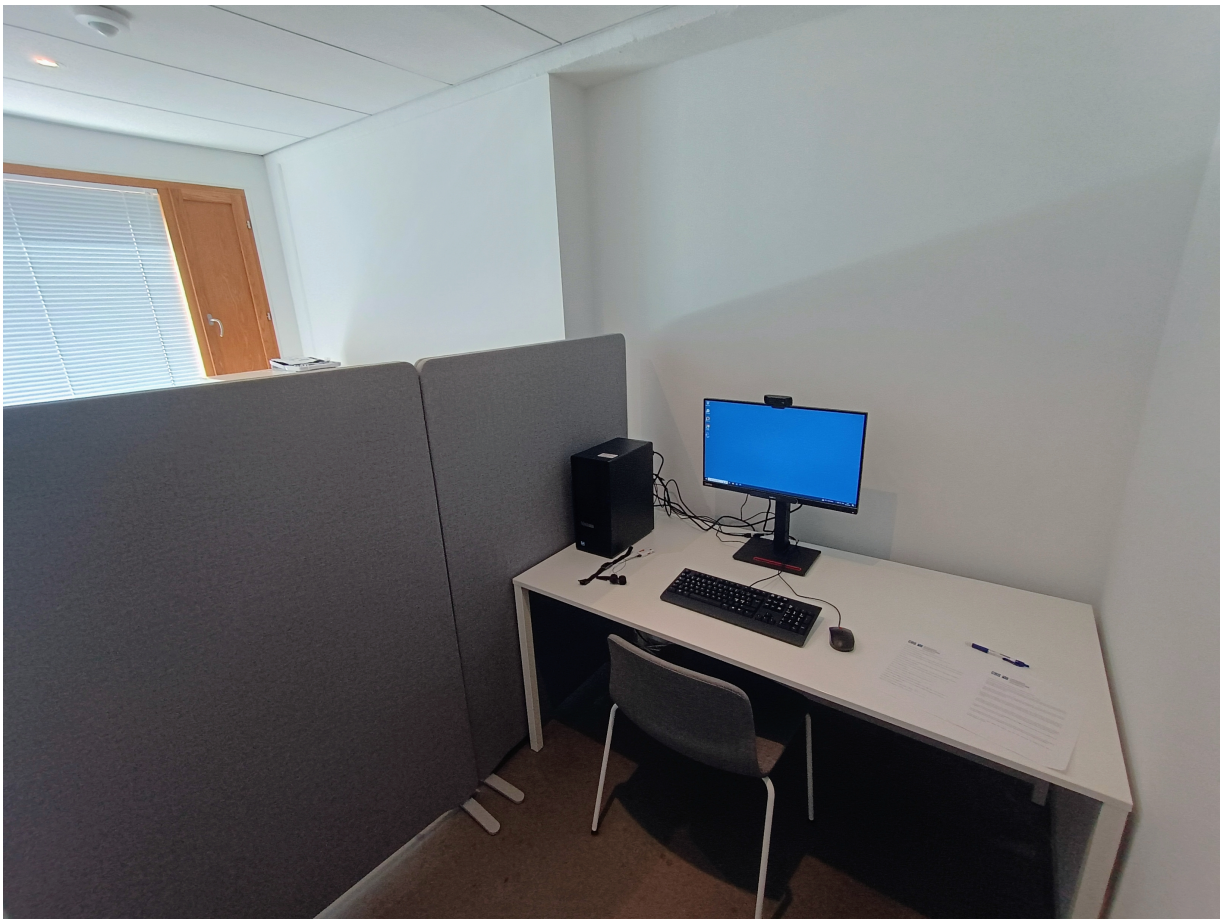


Figure 3: Laboratory setting

Before the experiment began, participants were instructed to read through a page detailing the experiment, its purpose, and what measures were collected. After this, they were asked to read through a consent form outlining anonymity as well as the potential to withdraw consent at any time.

After this, the participants were seated so that the monitor was 60-64 centimeters from their eyes and GSR and heart rate trackers were attached to their non-dominant hands. Eye-tracking calibrations were conducted before beginning the study. Tobii Pro Nano Eye Tracker and Shimmer3 GSR were used to record bodily signals. The survey and data recording was run locally using iMotions (https://imotions.com/, last accessed: 09/09/2024) software.

At the start of the experiment, a grey screen appeared for 2 minutes, this gave time for the experimenter to sit on the other side of the movable wall seen in Figure 3 and for the participant to get into a neutral state. After the survey, the trackers were removed from the subject's hand and as a reward for participating, they were given a voucher for a movie ticket worth 100 DKK. The experimenter took notes during the study about the quality of the eye tracker calibration as well as other information (for example it was noted if the participant wore glasses).

4.2 Data Preprocessing

After conducting the surveys, the output data needed to be transformed in order to be analyzable. In total, 6 datasets were used, 3 containing the survey results (the results of the image tampering survey, the online survey, and the lab experiment), as well as 3 tables containing data regarding eye tracking, GSR metrics, and heart rate metrics in relation to the lab experiment.

The three survey databases were all in wide format with a large number of columns (115 for the image rating survey, 616 for the online survey, and 612 for the lab survey) and a number of rows equal to the participants who filled out the survey (as well as a header row). This raw data contained many empty values, since each question had its own column, and each subject only saw one of four questions randomly selected. A

helper table was therefore used, with each question's ID corresponding to three values: "Question number", "AI generated", and "AI labeled", where the first value is between 1-48 and the latter two are binary, depending on the image and whether it is labeled or not. With this, the data was transformed into a long format, where each participant had 48 rows assigned to them for the 48 questions they answered, each row containing their personal data (such as age, nationality, attitudes to AI), the three question values described above and the scores they gave for believability, truthfulness, and credibility.

After this transformation, it was possible to analyze and preprocess the data further. The Likert-scale values were in text format (such as "Strongly agree"), and these were transformed into numerical values, similar to the questionnaire answers at the end regarding political stance, impulsivity, and attitudes to AI. Based on these values, for the two online experiments, those who failed the attention check were omitted. In order to have a single main variable, the average of believability, credibility, and trustworthiness was calculated for each row, creating the "score" variable. In a similar fashion, political attitudes and AI attitudes were also aggregated, these averages were used to create two binary variables, "pol_binary" and "attAI_binary". For each participant, these values were 0 if their average was below the median of all participants and 1 otherwise (meaning that pol_binary was 0 for more liberal and 1 for more conservative participants while attAI_binary was 0 for those who opposed AI more and 1 for those who did not). The resulting preprocessed tables were exported in order to be further studied.

The other three tables regarding bodily data were organized in a long format broken up by each question for each respondent. The main preprocessing step for these databases was therefore extracting the question number and question ID from text labels, in order to be connected to the results of the survey data later on. A number of rows solely contained events (such as the beginning of an experiment), these were removed.

4.3 Data Processing

The data preprocessing, as well as further manipulation of the data and statistical tests, were done using the programming language R, via its integrated development environment, RStudio. The following packages were used in the environment for the analysis: dplyr, ggplot2, ggpubr, glmmTMB, lavaan, lme4, readr ,readxl, semTools, sjmisc, sjPlot, stargazer, stringr, WRS2. ChatGPT was used to assist with debugging the code in order to ensure accurate outcomes.

Throughout Chapters 5 and 6, the results of a number of t-tests will be presented. To calculate these results, R's t.test() command was used, resulting in the output of Two Sample Welch t-tests (Welch, 1947). These tests aim to see if there is a significant difference between the mean of two groups that do not necessarily have equal variance. Such a test can be used for example in the case where we want to see if the believability scores are different for news with AI-generated images and with real images. The test calculates a t statistic via the following formula:

$$\boldsymbol{t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

Where $\boldsymbol{\bar{X}_1}$ is the sample mean of the first group, $\boldsymbol{\bar{X}_2}$ is the sample mean of the second group, $\boldsymbol{n_1}$ is the sample size of the first group, $\boldsymbol{n_2}$ is the sample size of the second group, $\boldsymbol{s_1^2}$ is the sample variance of the first group, and $\boldsymbol{s_2^2}$ is the sample variance of the second group. The calculated t statistic can be interpreted using a t distribution, resulting in a p-value. The null hypothesis of the test is that the means of the two groups are not significantly different. If the p-value is smaller than a chosen level, we reject this hypothesis. Throughout this paper, a p-value of 0.01 was chosen to indicate a significant difference. If in a test the p-value is smaller than 0.01, then we can reject the null hypothesis and say that there is a statistically significant difference between the means at the 99% confidence interval.

For the laboratory data, separate tests were run for the eye tracking, GSR metrics, and heart rate metrics. For the eye tracking, for each headline three Areas of Interest (AOI) were selected: the image, the title, and the label. This resulted in the data for each headline to be 3 rows, one for each AOI. The selected areas served as guidelines for the program, enabling it to calculate, for example, how long participants looked at the image specifically in a headline. Other eye-tracking measures, such as facial features, were also considered, however, since participants were tasked with a monotone assignment of reading, their expressions did not change significantly throughout the experiment.

For GSR and heart rate metrics, data was available for each headline. For GSR analysis, GSR peaks were analyzed, specifically Peak Count, as well as Peaks Per Minute, since this metric controlled for news title length. For heart rate, average heart rate (Beats Per Minute) was analyzed for each individual news piece.

5 Results

The following chapter details the findings of the experiments, focusing on the three surveys (the image rating survey, the online preliminary experiment, and the laboratory experiment) individually.

5.1 Participant Statistics

In total, 153 people completed the image rating survey, 197 did the preliminary experiment, and 23 did the laboratory experiment. In the preprocessing, those who failed the attention check were omitted, leaving 124 submissions for the image rating survey and 188 for the online study. 2 participants were also eliminated from the laboratory findings since one of the experiments was interrupted and another had no GSR data available.

Table 5.1 contains descriptive statistics for the online preliminary experiment as well as the laboratory study (the initial image rating survey did not have an extensive questionnaire at the end).

Table 5.1: Descriptive Statistics of the Experiments

Participant statistics		
	Online Experiment	Laboratory Experiment
Mean Age	44.30	24.19
σAge	13.36	2.81
Female %	48.40	42.86
Mean Political Attitude	2.51	2.14
σPolitical Attitude	1.13	0.75
Mean Attitude to AI	3.60	4.10
σAttitude to AI	1.05	0.57

As we can see, there is a significant difference between the demographics of the two experiments. This can be explained by the fact that the majority of lab participants were university students, while the Qualtrics survey was set to include a wider

demographic. The participants of the laboratory experiment were younger, politically more liberal, and had a more positive attitude to Artificial Intelligence on average.

5.2 Image Rating Survey Results

While at its core, the image rating survey only connects one dependent variable (the score given to the prompt "This image has been tampered with Artificial Intelligence (AI)") and a binary explanatory variable (whether or not the image has been AI generated), a number of other control variables were included to truly capture the relationship. Initially, these included three more variables. A binary "Fake" variable, showcasing whether the article associated with the headline image is fake news (note, that participants did not see any text from the news article). The second variable is "leaning", which is 0 for liberal favorable articles, 1 for neutral news, and 2 for conservative favorable articles. Finally, the third is an exogenous variable created after the survey has been conducted called "Face". This binary variable is 1 when there is a face present in an image and 0 otherwise. This is included due to present AI image generators struggling with the creation of realistic-looking faces (Kolomeets et al., 2024). These 4 variables were included in a multivariate linear regression model in order to see if they were significant. The results can be seen in Table 5.2.

As we can see, the binary variables regarding AI generation, fake news, and having a face in the image are all significantly correlated at the 99.9% confidence interval with the score that subjects gave for image tampering. This is a surprising outcome since subjects had no interaction with the news articles themselves, and it means that the images (both real and AI-generated) may be tied to the authenticity of the articles. Based on the three key binary variables, Figure 4 showcases how the distribution of scores given changes in each specific scenario. We can see from this plot that the variations are larger for AI-generated images and we can also see the scores for AI tampering to be higher for such pictures.

Table 5.2: Image Scoring Linear Regression

	Dependent variable:
	score
AI_generated	**2.270*** **
	(0.049)
Fake	**0.688*** **
	(0.050)
Face	**0.188*** **
	(0.051)
leaning	**−0.047**
	(0.030)
Constant	**2.267*** **
	(0.060)
Observations	5,952
R^2	0.281
Adjusted R^2	0.280
Note:	*p<0.05; **p<0.01; ***p<0.001

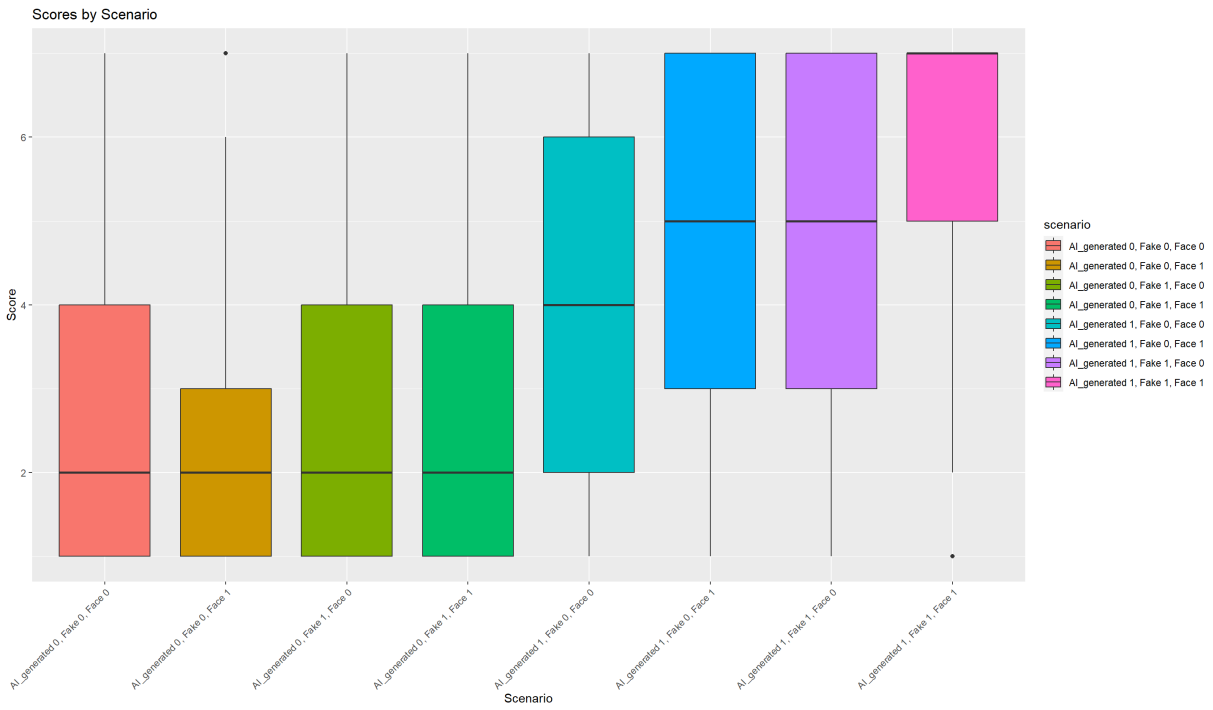


Figure 4: Image Rating Survey Scores

5.3 Preliminary Experiment Results

5.3.1 *Initial Findings*

Before analyzing results focusing on the believability scores, an important aspect to look at is how subject behavior changes throughout participation in the experiment. One might assume that over time participants might realize that the AI labels are put up randomly, and start ignoring them. This could result in a difference between the scores of the first news articles and the remainder. To see if this is the case, the data was split so that the first 12 articles were compared with the last 36. Welch Two Sample t-tests were conducted in the overall data, as well as four slices of the dataset based on whether the image is AI generated and labeled.

Table 5.3: First 12 and Last 36 News Headline Scores

Statistic	Overall	AI 0; label 0	AI 0; label 1	AI 1; label 0	AI 1; label 1
t-value	0.85806	1.3769	0.34096	-0.53205	0.28292
DoF	3852.1	987.63	1022.7	90.875	931.7
p-value	0.3909	0.1688	0.7332	0.5948	0.7773
Mean of First 12	2.948	3.110	3.022	2.867	2.892
Mean of Last 36	2.976	3.021	2.999	2.902	2.873

As we can see from Table 5.3, there is no significant difference between the scores in any of the scenarios (at the 99% confidence interval). In this separation, participants saw on average 6 mislabelled images (real images labeled as AI generated and vice versa). Different cutoffs were also examined, comparing the first 24 with the last 24 articles, as well as the first 36 to the last 12. In all these cases, there was no scenario with a significant difference between the groups at the 95% confidence interval. Similarly, instead of looking at the aggregate score, the individual points of believability, truthfulness, and credibility were also analyzed, once again showing no significant differences. To conclude, contrary to the above assumption, there was no difference in behavior by participants throughout the study. This allows us to analyze the results as a

whole in the rest of this section.

5.3.2 *Believability, Truthfulness, Credibility*

For a number of measurements throughout this section, a score variable is used instead of analyzing the three measures of believability, truthfulness, and credibility individually. While this allows for a prompt analysis of the results, it is also important to look at the three metrics separately. To see why this is the case, t-tests were conducted to confirm that there are statistically significant differences between the aggregated score and the individual metrics (Table 5.4).

Table 5.4: Difference Between "score" Variable and Individual Metrics

Statistic	Believability	Truthfulness	Credibility
t-value	6.1891	-2.1261	-4.1953
DoF	17974	18042	18033
p-value	6.184e-10	0.03351	2.738e-05
Mean of Individual metric	3.083	2.912	2.870
Mean of Score variable	2.955	2.955	2.955

The difference between truthfulness and the score was significant only at the 95% confidence interval, while the difference between believability and credibility compared to the score were both significantly different at the 99% confidence interval.

5.3.3 *Overall Findings*

The three binary variables differing between news headlines are the news type (fake or true), image type (real or AI-generated), and whether or not there is a label. The following table contains the effects of these variables on both the whole dataset, as well as on subsets of the data (Table 5.5).

Table 5.5: Binary Variable Effects on Score

Data	Binary Measure	Mean Score (measure = 0)	Mean Score (measure = 1)	p-value
All	Fake News	3.344	2.566	<0.0001
All	AI Image	3.025	2.886	<0.0001
All	AI Label	2.968	2.942	0.3510
Fake News	AI Image	2.641	2.494	0.0002
Fake News	AI Label	2.568	2.564	0.9252
True News	AI Image	3.398	3.289	0.0035
True News	AI Label	3.363	3.324	0.2925
AI Image	AI Label	2.894	2.877	0.6758
Real Image	AI Label	3.044	3.005	0.3264
Fake News, AI Image	AI Label	2.483	2.503	0.7161
Fake News, Real Image	AI Label	2.658	2.624	0.5392
True News, AI Image	AI Label	3.318	3.258	0.2591
True News, Real Image	AI Label	3.407	3.388	0.7168

As we can see, in all scenarios Fake News and AI Image are significantly influencing the scores, having a negative effect. AI Labels, on the other hand, are insignificant both in the overall data, as well as in subsets of the data. Appendix B shows further analysis focusing on the interchangeability of the binary AI Image variable and the 7-scale scores given in the image rating survey.

5.3.4 *Post Hoc Analysis*

Differences by age and gender

Of the 188 participants in the preliminary study, 91 were female, and 97 were male. One person did not provide data for their age, the remaining had the following statistics: mean: 44.3, median: 43, min: 20, max: 83 years. Figure 5 contains information about the age distribution as well as the binary political leaning.

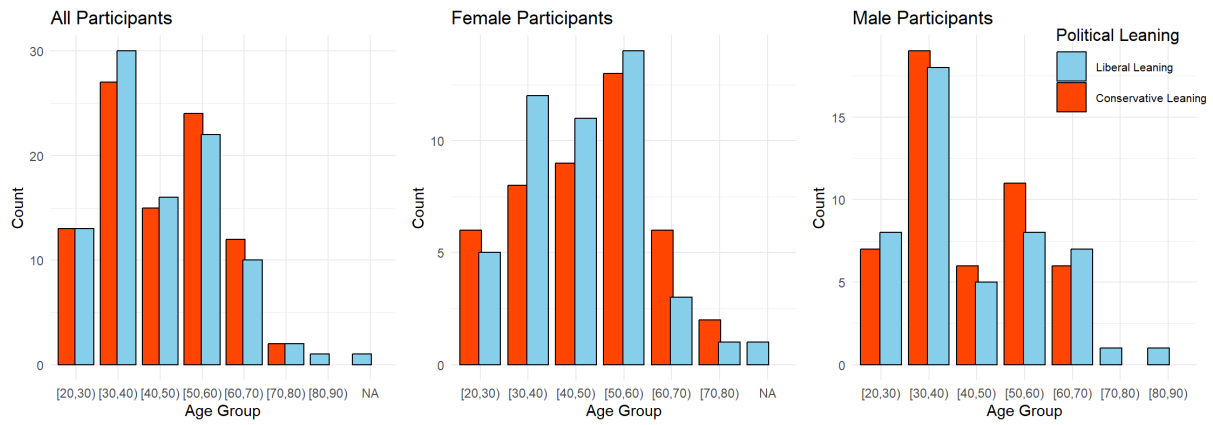


Figure 5: Preliminary Experiment Demographics

Political leaning

As we have seen in previous literature, dissonance may arise in subjects after encountering information that is not supported by their previous views (Harmon-Jones and Mills, 2019). So far in this study, we observed an aggregate decrease in the believability of news headlines when AI images were present, this section will explore if this trend still holds when controlling for political views. The experiment was divided into 6 segments based on the inclination of news articles (liberal favorable, conservative favorable, and neutral) and whether it is fake or not.

The two groups divided by political attitude were first examined to see how they differ. As we can see in Figure 5, there is no large deviation between the ratios of liberal to conservative subjects by age, this is slightly different than what surveys may suggest (Peltzman, 2019), where the ratio of liberal-leaning people is higher among young populations, while lower in older cohorts. Regarding gender, 51.6% of females are more liberal-leaning while this figure is 49.5% for males (remember that the binary political attitude was created by dividing the groups based on the median attitude, so this figure would be 50% for the overall group). Finally, we examined how the scoring behavior of the two groups differed, whether one group tended to give lower scores overall, using a t-test. The liberal-leaning group gave higher scores for truthfulness and credibility at the

99% confidence interval. For believability and the overall score rating, however, these differences were not significant at the 1% interval. The exact figures can be seen in Table 5.6.

Table 5.6: Difference in average scores given by political cohorts

Score Category	Conservative Favorite	Liberal Favorite	p-value
Believability	3.11	3.06	0.0713
Truthfulness	2.95	2.88	0.0092
Credibility	2.91	2.83	0.0099
Overall Score	2.99	2.92	0.0163

To see the difference in how the two cohorts scored regarding different news types, t-tests were conducted using the aggregated "score" variable (Table 5.7). When looking at statistical significance at the 99% confidence interval, in almost all cases, using one of the other three scoring metrics resulted in similar results, with the only exception in the case of true, conservative favorable news, when the difference in believability was not significant ($p = 0.04097$), while in the overall score, it was.

Table 5.7: Difference in scores given for each news type by political cohorts (Preliminary Experiment)

News Category	Conservative Favorite	Liberal Favorite	p-value
Liberal, True	3.27	3.54	<0.0001
Liberal, Fake	2.69	2.83	0.0320
Neutral, True	3.46	3.52	0.3879
Neutral, Fake	2.50	2.38	0.0652
Conservative, True	3.23	3.04	0.0030
Conservative, Fake	2.74	2.17	<0.0001

These results show that there was a deviation in the scores given between the two cohorts depending on what news article they were reading. For neutral-leaning news (both fake and true) there was no statistically significant difference at the 99% confidence level. This was to be expected, since there was no difference in the overall

scoring either, and these articles are not concerned with politics. There was, however, a significant difference in 3 of the 4 other cases, except for liberal-leaning fake news. In these cases, the cohort whose views aligned with the articles gave significantly higher scores than the cohort who opposed them. These cases were further examined in order to see how having AI images and labels influences the results. For each case, 8 t-tests were conducted, 4 for AI images and AI labels separately and 4 tests in a 2-by-2 matrix for AI images and AI labels.

Liberal Favorite, True

Table 5.8: Difference in scores given for each headline type by political cohorts (True, Liberal favorable articles)

Image Category	Conservative Favorite	Liberal Favorite	p-value
Real image	3.30	3.48	0.0663
AI image	3.23	3.61	<0.0001
Not labelled	3.30	3.61	0.0009
Labelled	3.23	3.47	0.0084
Real image, not labelled	3.37	3.56	0.1696
Real image, AI labelled	3.24	3.67	0.0008
AI image, not labelled	3.25	3.39	0.2561
AI image, AI labelled	3.21	3.55	0.0112

Liberal Favorite, Fake

Table 5.9: Difference in scores given for each headline type by political cohorts (Fake, Liberal favorable articles)

Image Category	Conservative Favorite	Liberal Favorite	p-value
Real image	2.79	2.98	0.0459
AI image	2.58	2.70	0.2245
Not labelled	2.75	2.85	0.3006
Labelled	2.63	2.82	0.0485
Real image, not labelled	2.84	3.05	0.1131
Real image, AI labelled	2.75	2.92	0.2196
AI image, not labelled	2.66	2.67	0.9587
AI image, AI labelled	2.49	2.72	0.0864

Conservative Favorite, True

Table 5.10: Difference in scores given for each headline type by political cohorts (True, Conservative favorable articles)

Image Category	Conservative Favorite	Liberal Favorite	p-value
Real image	3.27	3.13	0.1101
AI image	3.19	2.96	0.0120
Not labelled	3.19	3.04	0.0906
Labelled	3.28	3.04	0.0110
Real image, not labelled	3.20	3.06	0.2967
Real image, AI labelled	3.37	3.19	0.1826
AI image, not labelled	3.18	3.01	0.1823
AI image, AI labelled	3.20	2.91	0.0307

Conservative Favorite, Fake

Table 5.11: Difference in scores given for each headline type by political cohorts (Fake, Conservative favorable articles)

Image Category	Conservative Favorite	Liberal Favorite	p-value
Real image	2.80	2.26	<0.0001
AI image	2.67	2.08	<0.0001
Not labelled	2.72	2.12	<0.0001
Labelled	2.76	2.21	<0.0001
Real image, not labelled	2.78	2.25	0.0001
Real image, AI labelled	2.83	2.27	<0.0001
AI image, not labelled	2.66	2.02	<0.0001
AI image, AI labelled	2.68	2.16	0.0002

Attitude to AI

As mentioned in Chapter 4, at the end of the experiment, participants had to answer four questions regarding their attitudes to Artificial Intelligence. These four scores, ranging from 1-5 (1 disagreeing with the use of AI being "good", "wise", "desirable" and "beneficial) were averaged to have a single parameter about participants' views on AI, and based on the median of these values, they were split into two groups (mean: 3.597, standard deviation: 1.05). Table 5.12 looks at the difference between the two groups (Negative Attitude being below median attitude score, Positive Attitude for above) in an overall setting, as well as in 14 subsections based on whether or not the headline image is made with AI, labeled as AI-generated, and a combination of these, together with whether the article is fake news or not.

Table 5.12: Difference in scores given for each article type based on AI attitudes (Preliminary Experiment)

News Category	Negative Attitude	Positive Attitude	p-value
Overall	2.93	3.01	0.0148
AI Images	2.87	2.92	0.2343
AI Labelled	2.92	2.98	0.2025
Real Image, Not labelled	3.00	3.13	0.0291
Real Image, AI labelled	2.99	3.04	0.3634
AI Image, Not labelled	2.88	2.93	0.3931
AI Image, AI labelled	2.86	2.91	0.4129
Real Image, Not labelled, True	3.36	3.51	0.0601
Real Image, Not labelled, Fake	2.62	2.73	0.2161
Real Image, AI labelled, True	3.39	3.38	0.8666
Real Image, AI labelled, Fake	2.56	2.75	0.0353
AI Image, Not labelled, True	3.33	3.30	0.6877
AI Image, Not labelled, Fake	2.45	2.56	0.2086
AI Image, AI labelled, True	3.27	3.23	0.5895
AI Image, AI labelled, Fake	2.56	2.75	0.0353

As we can see from the table, there is no statistically significant difference in any of the cases, suggesting that AI attitude does not influence people's believability of news.

We have seen previously that on the whole cohort, there was no significant effect of labels. In fact, Table 5.13 shows that this is not specific for either group, here we can see that whether or not there are AI labels does not change the scores given significantly at the 99% confidence interval. For both groups, however, there are smaller scores given for headlines with AI-generated images ceteris paribus.

Table 5.13: Differences in scores given for each type of headline based on AI attitudes

Group	Not Labelled	AI Labelled	p-value
Negative Attitude	2.94	2.92	0.6782
Positive Attitude	3.03	2.98	0.3266
Group	Real Image	AI Image	p-value
Negative Attitude	2.99	2.87	0.0002
Positive Attitude	3.09	2.92	0.0015

5.4 Laboratory Experiment Results

5.4.1 *Initial Findings*

Similarly to the preliminary study, in order to see if results can be interpreted homogeneously throughout the survey, the first 12 scores were compared to the rest of the answers in order to see if there was a change in participant behavior throughout the survey (Table 5.14).

Table 5.14: First 12 and Last 36 News Headline Scores

Statistic	Overall	AI 0; label 0	AI 0; label 1	AI 1; label 0	AI 1; label 1
t-value	-0.44478	-0.46238	0.51993	-0.022141	-1.163
DoF	1329.6	355.06	342.08	330.15	297.5
p-value	0.6565	0.6441	0.6034	0.9823	0.2458
Mean of First 12	2.776	2.979	2.853	2.725	2.515
Mean of Last 36	2.799	3.025	2.803	2.727	2.627

It is interesting to see the difference in scores between the laboratory study and the online survey. Table 5.15 shows that those in-person gave lower scores. One explanation for this could have been that they are more focused, and therefore better at identifying fake content (Pennycook and Rand, 2019). After looking at true and fake news separately, we can see that this is not the case, and every kind of news received lower scores, not just fakes.

Table 5.15: Online and Laboratory scores

News Headlines	Mean Online	Mean Lab	p-value
All	2.96	2.79	<0.0001
Fake	3.43	3.15	<0.0001
True	2.57	2.43	0.0003

Due to the laboratory setting, for this experiment, there was detailed data available on how long participants spent on each page. Therefore, we can see if there was a

difference in the time they spent on headlines in the beginning and in the second half of the experiment. Table 5.16 contains the results of the t-test for all images, as well as separate tests for only real and only AI-generated images.

Table 5.16: Time spent on First 12 and Last 36 headlines (seconds)

Image Type	First 12 Mean	Last 36 Mean	p-value
All	25.013	20.499	<0.0001
Real Image	25.098	20.370	<0.0001
AI Image	24.920	20.631	<0.0001

As we can see, participants spent consistently more time in. This difference was persistent even when the cutoff was after a different number of articles. In the case of the first 36 versus the last 12 articles (all images), participants spent an average of 22.75 seconds on the first 36 articles and 18.27 on the last 12 (p-value <2.2e-16). This shows that the time spent on articles further decreased as participants read more headlines. Figure 6 captures this by looking at the average time spent on each question, together with a linear trend line.

While the trend has outliers, these could be simply because of the differences in the number of words in some news headlines. Overall, we can see that there is some change in subject behavior over time, but because there is no difference in scoring behavior, we can look at the answers as homogeneous over the length of the experiment.

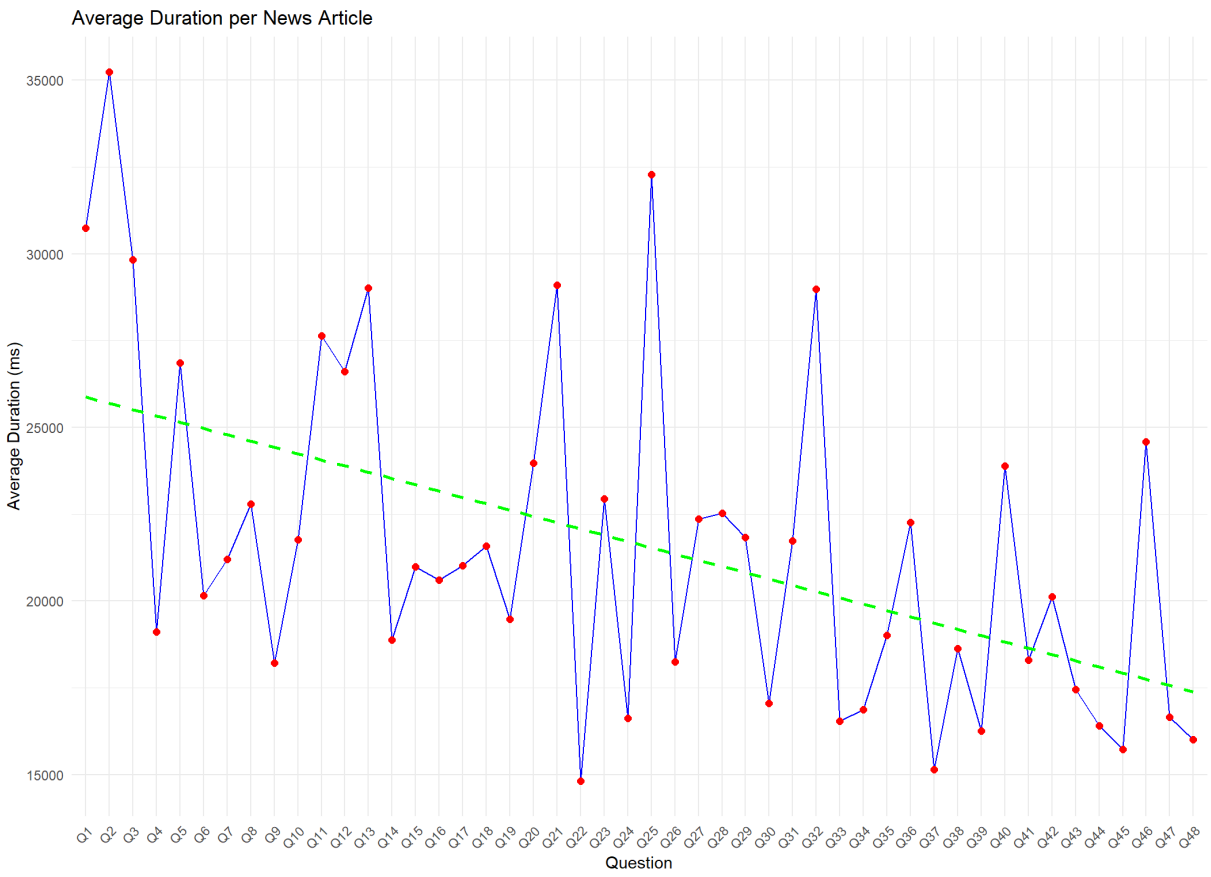


Figure 6: Average time spent on news articles over time

5.4.2 *Believability, Truthfulness, Credibility*

Once again, we examine the difference between the aggregate score metric and the individual scales that subjects had to fill out. The results for this can be seen in Table 5.17. Interestingly, the difference between believability and credibility scores is larger than previously (0.442 compared to 0.213). This is due to the lower scores given to credibility. One potential reason for this could be that participants were more skeptical about news sources, however, the cause for this is unclear and requires further research.

5.4.3 *Overall Findings*

Of the 21 participants, 9 were female, and 12 male. There was a significantly smaller difference in age distribution compared to the preliminary experiment, with the following

Table 5.17: Difference between "score" variable and individual metrics

Statistic	Believability	Truthfulness	Credibility
t-value	-22.326	5.5868	19.069
DoF	3023	3023	3023
p-value	2.2e-16	2.518e-08	2.2e-16
Mean of Individual metric	3.038	2.745	2.596
Mean of Score variable	2.793	2.793	2.793

statistics: mean 24.2, median: 24, min: 19, max: 33. This can be explained by the difference in the recruitment process, where the online study was open to a wide range of demographics, and mainly university students were informed of the laboratory experiment.

Table 5.18 contains information about the believability scores' relationships with the binary variables for fake news, AI image, and AI labels. As we can see, similarly to the online study, Fake News and AI Image are both significant in the t-tests in all cases. Unlike in the preliminary experiment, AI label is significant in some cases here, both in the overall data and in some subsets. The most notable is when news type and image type are controlled. In these cases, we can see that labels are insignificant in most cases, except when true news and real images are present. In this case, labeling the image to be AI-generated (even though it is not), decreases the believability score by 0.261 on average.

Regarding political attitudes, the distribution of the aggregate political score was also smaller in the laboratory study (mean: 2.238, standard deviation: 0.75). As such, the difference between the two groups (after separated by the medium attitude), is also smaller. Table 5.19 summarizes the difference between scores given for the two groups similarly to Table 5.7.

As we can see, there was only one case where there was a significant difference between the scoring behavior of the two groups, in conservative-leaning true news. This

Table 5.18: Binary Variable Effects on Score (Laboratory Experiment)

Data	Binary Measure	Mean Score (measure = 0)	Mean Score (measure = 1)	p-value
All	Fake News	3.151	2.435	<0.0001
All	AI Image	2.918	2.663	<0.0001
All	AI Label	2.875	2.709	0.0001
Fake News	AI Image	2.573	2.287	<0.0001
Fake News	AI Label	2.493	2.377	0.0536
True News	AI Image	3.270	3.031	<0.0001
True News	AI Label	3.236	3.060	0.0021
AI Image	AI Label	2.726	2.601	0.0408
Real Image	AI Label	3.014	2.817	0.0012
Fake News, AI Image	AI Label	2.291	2.284	0.9324
Fake News, Real Image	AI Label	2.657	2.480	0.0328
True News, AI Image	AI Label	3.081	2.9718	0.1620
True News, Real Image	AI Label	3.402	3.141	0.0015

Table 5.19: Difference in scores given for each news type by political cohorts (Laboratory Experiment)

News Category	Conservative Favorite	Liberal Favorite	p-value
Liberal, True	3.19	3.23	0.8217
Liberal, Fake	3.13	2.69	0.0204
Neutral, True	3.17	3.20	0.8351
Neutral, Fake	2.55	2.39	0.2009
Conservative, True	3.35	2.90	<0.0001
Conservative, Fake	2.44	2.11	0.0707

means that the results of the laboratory experiment will not be comparable to the literature focusing on political stances and fake news.

Similarly to the preliminary experiment (Appendix B), it is also worthwhile to see the relationship between the binary AI image variable and the 7-scale average for AI tampering. A similar regression was run with the same control variables in order to see their effects. As we can see, the AI image variable is insignificant now, due to its high correlation with the 7-scale score (0.79).

Table 5.20: Lab Score to AI Image and Image Score Linear Regression

	Dependent variable:
	Score
AI Image	**−0.039**
	(0.070)
Image Rating Score	**−0.094*****
	(0.024)
AI Labelled	**−0.137*****
	(0.041)
Binary AI Attd	**−0.114***
	(0.048)
Fake News	**−0.651*****
	(0.044)
Constant	**3.653*****
	(0.078)
Observations	3,024
R^2	0.111
Adjusted R^2	0.110
Note:	*p<0.05; **p<0.01; ***p<0.001

Regarding attitudes to AI, the surveyed scores were higher for the participants in the

laboratory experiment on average, with less dispersion (mean: 4.095, standard deviation: 0.57). This can mean that similarly to the political attitudes, the results will not be as significant as in the case of the preliminary study.

Table 5.21: Difference in scores given for each article type based on AI attitudes (Laboratory Experiment)

News Category	Negative Attitude	Positive Attitude	p-value
Overall	2.76	2.89	0.0197
AI Images	2.72	2.66	0.4134
AI Labelled	2.65	2.71	0.4281
Real Image, Not labelled	2.93	3.14	0.0023
Real Image, AI labelled	2.82	2.80	0.8801
AI Image, Not labelled	2.68	2.88	0.0478
AI Image, AI labelled	2.63	2.51	0.2583
Real Image, Not labelled, True	3.19	3.58	0.0013
Real Image, Not labelled, Fake	2.61	2.80	0.1617
Real Image, AI labelled, True	3.09	3.36	0.0699
Real Image, AI labelled, Fake	2.51	2.41	0.4754
AI Image, Not labelled, True	3.05	3.17	0.3329
AI Image, Not labelled, Fake	2.26	2.42	0.3105
AI Image, AI labelled, True	2.99	2.92	0.6616
AI Image, AI labelled, Fake	2.15	2.32	0.1530

From the table, we can see that at the 99% confidence interval, there is a difference between the two groups in two cases, both concerning real, unlabelled images, specifically true news. Table 5.22 contains the within-group differences AI labels have in specific cases.

Table 5.22: Difference in scores given for each article type based on AI attitudes (Laboratory Experiment)

News Category	Not Labelled	AI Labelled	p-value
Positive Attitude			
Overall	3.08	2.66	<0.0001
Real Image	3.26	2.80	0.0005
AI Image	2.88	2.52	0.0033
Real Image, True	3.74	3.36	0.0362
Real Image, Fake	2.80	2.41	0.0189
AI Image, True	3.17	2.92	0.1408
AI Image, Fake	2.42	2.15	0.1214
Negative Attitude			
Overall	2.81	2.72	0.0924
Real Image	2.93	2.82	0.1160
AI Image	2.68	2.63	0.4742
Real Image, True	3.28	3.09	0.0361
Real Image, Fake	2.61	2.51	0.2823
AI Image, True	3.05	2.99	0.4844
AI Image, Fake	2.26	2.32	0.5193

5.4.4 *Eye-tracking Results*

The measures from the eye-tracking data can be used to see how participants reacted to different scenarios in a number of ways. Specifically, Time to First Fixation (TTFF, in seconds) and Dwell Count were measured. For each metric, each AOI was looked at individually. For the image and title, both all headlines and only labeled ones were analyzed, while for labels only the labeled headlines were considered (since there was no data for unlabelled headlines). Table 5.23 contains the results for TTFF for each AOI in regards to AI images, Fake news, and the binary attitude to AI (1 being negative to AI and 0 being positive). In both metrics, the only significant factor was AI attitude. The analysis in Appendix C looks at the two groups in more detail in order to see if this difference can be explained by one of the control variables, or if the groups inherently behave differently. As we can see, all the results are insignificant withing the tests, suggesting that AI attitude explains the difference between the groups (Tables **??** and **??**).

Table 5.23: Time to First Fixation for different AOIs and Binary Measures

AOI	Binary Measure	TTFF (seconds) (measure = 0)	TTFF (seconds) (measure = 1)	p-value
Image	AI Image	2.122	1.899	0.3130
Title	AI Image	0.384	0.447	0.0656
Label	AI Image	9.610	7.989	0.1187
Image (labelled)	AI Image	1.865	1.790	0.7979
Title (labelled)	AI Image	0.444	0.484	0.4784
Image	Fake	1.752	2.273	0.01858
Title	Fake	0.405	0.425	0.5616
Label	Fake	9.643	7.830	0.0757
Image (labelled)	Fake	1.642	2.004	0.2144
Title (labelled)	Fake	0.457	0.470	0.8090
Image	AI Attitude	1.723	2.942	0.0001
Title	AI Attitude	0.411	0.429	0.6796
Label	AI Attitude	9.255	6.897	0.2358
Image (labelled)	AI Attitude	1.539	2.825	0.0021
Title (labelled)	AI Attitude	0.462	0.471	0.9016

Table 5.24: Dwell Count for different AOIs and Binary Measures

AOI	Binary Measure	Dwell Count (measure = 0)	Dwell Count (measure = 1)	p-value
Image	AI Image	7.016	7.075	0.8019
Title	AI Image	5.125	5.032	0.6537
Label	AI Image	1.521	1.586	0.6390
Image (labelled)	AI Image	7.228	7.016	0.5189
Title (labelled)	AI Image	5.032	5.161	0.6532
Image	Fake	7.266	6.823	0.0609
Title	Fake	5.156	5.000	0.0.4393
Label	Fake	1.624	1.478	0.2851
Image (labelled)	Fake	7.453	6.809	0.0506
Title (labelled)	Fake	5.317	4.887	0.1349
Image	AI Attitude	7.307	6.201	<0.0001
Title	AI Attitude	5.323	4.300	<0.0001
Label	AI Attitude	1.642	1.205	<0.0001
Image (labelled)	AI Attitude	7.315	6.455	0.0105
Title (labelled)	AI Attitude	5.271	4.496	0.0052

5.4.5 *GSR Results*

As mentioned in Chapter 4, Peak Count and Peaks Per Minute, a derivative of the first measure, will be analyzed. This is to see if the time it takes to read through an article influences participants' behavior. The following tables contain a basic analysis of the measures when controlling for other variables.

Table 5.25: Peak Count for Different Binary Measures

Data	Binary Measure	Peak Count (measure = 0)	Peak Count (measure = 1)	p-value
All	Fake News	1.671	1.546	0.0461
All	AI Image	1.679	1.534	0.0208
All	AI Label	1.600	1.616	0.8033
All	AI Attitude	1.901	1.517	<0.0001
AI Images	Fake News	1.572	1.496	0.3694
AI Images	AI Label	1.543	1.526	0.8437
Real Images	Fake News	1.768	1.592	0.0565
Real Images	AI Label	1.654	1.705	0.5761

Similar results can be seen in Table 5.26 for the Peaks Per Minute metric.

Table 5.26: Peaks Per Minute for Different Binary Measures

Data	Binary Measure	Peaks Per Minute (measure = 0)	Peaks Per Minute (measure = 1)	p-value
All	Fake News	4.584	4.694	0.4863
All	AI Image	4.732	4.543	0.2299
All	AI Label	4.643	4.635	0.9639
All	Binary AI Attd	5.892	4.248	<0.0001
AI Images	Fake News	4.527	4.558	0.8895
AI Images	AI Label	4.505	4.580	0.7363
Real Images	Fake News	4.640	4.821	0.4158
Real Images	AI Label	4.771	4.690	0.7170

Since in both cases, the most significant variable was attitude to AI, Table 5.27 explores this metric further, separately for those with a positive and negative attitude.

Table 5.27: Peak Count and PPM for Different AI Attitudes

Data	Binary Measure	Peak Count (measure = 0)	Peak Count (measure = 1)	p-value
Pos AI Attd	Fake News	1.942	1.859	0.4756
Pos AI Attd	AI Image	1.880	1.922	0.7199
Pos AI Attd	AI Label	1.961	1.832	0.2646
Neg AI Attd	Fake News	1.586	1.448	0.0598
Neg AI Attd	AI Image	1.614	1.417	0.0070
Neg AI Attd	AI Label	1.480	1.553	0.3221
Data	Binary Measure	Peaks Per Minute (measure = 0)	Peaks Per Minute (measure = 1)	p-value
Pos AI Attd	Fake News	5.713	6.070	0.2182
Pos AI Attd	AI Image	5.996	5.778	0.4518
Pos AI Attd	AI Label	6.112	5.637	0.0938
Neg AI Attd	Fake News	4.231	4.264	0.8591
Neg AI Attd	AI Image	4.326	4.168	0.3887
Neg AI Attd	AI Label	4.151	4.343	0.2932

5.4.6 *Heartrate Results*

Table 5.28 summarizes the results of the heart rate metrics (BPM), for both the complete dataset as well as for only AI-generated images. As we can see, the results here are significant in the case of AI Images, and insignificant otherwise. This can be explained by the fact that the experiments were not intended to evoke strong reactions, and participants were mainly focused on reading news headlines, as even when the difference is statistically significant, there is only a small variation.

Table 5.28: Difference in scores given for each article type based on AI attitudes

Data	Binary Measure	BPM (measure = 0)	BPM (measure = 1)	p-value
All	Fake News	67.30	67.88	0.1180
All	AI Image	67.26	67.94	0.0150
All	AI Labelled	67.69	67.99	0.1893
AI Images	Fake News	67.58	68.11	0.1474
AI Images	AI Labelled	67.58	67.30	0.1534

6 Discussion

The aim of this thesis was to investigate how AI-generated images and labels influence the believability of news headlines in both real and fake news articles. The preliminary study and the laboratory experiment aimed at controlling for several factors such as political leaning, fake news, and attitudes to AI. The following sections discuss the findings of the experiments, highlighting the hypotheses set in Chapter 3.

6.1 Findings

6.1.1 *Image Rating Survey*

As we have seen from the image rating survey results, Midjourney's AI image generator is, in most cases, far from reality. On aggregate, respondents were able to distinguish real images from AI-generated ones. This was especially true when there were faces present, as the image generator struggled to produce lifelike details. This is in line with Kolomeets et al. (2024) who showed that bots who used real or human-edited faces gained more trust than those who used AI-generated ones on social media platforms. It could be worthwhile to repeat the experiments as AI technologies evolve and the difference between real and AI-generated images diminishes. This change could be measured by the decrease in the score and could be implemented as a function of the believability of news.

Interestingly, images associated with fake news articles received higher scores, even without the texts of the articles. While this does not imply causation, one explanation for this could be fake news articles having more eye-catching images associated with them to further increase attention.

6.1.2 *Preliminary Experiment*

While the implementation of the attention check provided some guarantee for reliable data, on average participants in the online study completed the survey 12% faster than in the lab, suggesting that the respondents were not as mindful. Despite this, the group was much more diverse both regarding age and political attitude, allowing us to inspect how different cohorts responded differently.

The overall findings showcase that AI-generated images decreased believability for both true (0.109 difference in mean scores) and fake (0.147) news. On the other hand, putting up the "Image generated by AI" labels had no significant effect in any of the cases (Table 5.5). Regarding the use of the image rating score scale instead of the binary AI Image variable, we could see that the two are both significantly influencing believability, and had a high correlation, suggesting that they are interchangeable.

Our initial findings regarding the differences between political cohorts showcased that there are significant differences between the scorings of liberal and conservative-leaning participants in the cases of true liberal and both true and fake conservative-leaning articles. Similarly to Pennycook et al. (2018)'s results, we could see that even after controlling for every binary variable, there were cases where different groups believed news at different levels. In the cases of liberal favorite true news, when presented with real images labeled AI-generated, liberal-leaning participants scored significantly higher than more conservative subjects. Further investigation looked at the effect of labels in true, liberal favorite news to show that for liberal participants labels increased believability, while it decreased it for conservatives, both when real and when AI-generated images were present. This effect was not as significant in the case of conservative favorite true news, resulting in no difference between the groups.

The biggest deviation between liberals and conservatives was in the case of conservative-leaning fake news. Labels did not have such a strong effect as we've seen

previously, however, both for real and AI images, overall believability was significantly different. When comparing the results of liberal fake news, we can see that this is not caused by the higher average believability among conservative subjects, but rather a much lower average score for liberal-leaning participants.

6.1.3 *Laboratory Experiment*

While there were some deviances in scoring distributions in the laboratory results compared to the preliminary survey, and there was a smaller sample size of 21 participants, similar results were seen overall in the two experiments, with AI images decreasing believability ceteris paribus. When looking at AI labels, we could see significant effects in the case of true news, specifically when real images were present. In this case, AI-labelled headlines on average received a 0.261 smaller believability score, suggesting a distrust in the labels.

Post hoc analysis showed that unlike in the preliminary study, political attitudes were not relevant. This can be explained by the fact that the subjects were less diverse politically, and while they were separated by the median political attitude of the two groups, the groups behaved rather similarly. AI attitude, however, was a significant factor in scoring behavior in the experiment in some cases. We saw overall that in the case of true articles with real images and no labels, those with a positive attitude toward AI scored 0.39 higher. This is a surprising outcome since this scenario has no connections to AI. The construction of the survey can serve as an answer. Since participants also faced scenarios with AI present, those with a negative attitude to it might have felt more deceived, and therefore scored lower. While Kolomeets et al. (2024) explores trust in AI images, the paper does not look at groups separately. Further research therefore should be aimed at exploring behavioral differences between AI positive and AI adverse people.

Apart from scorings, AI attitude was also seen as a difference-maker in emotional responses. Regarding eye tracking, when controlling for AI images and fake news, there

was no significant effect in either time to first fixation or dwell count. However, AI attitude proved to have a significant effect. Regarding TTFF, we could see that those with a negative attitude took longer to first look at the image, but focused on the label earlier when it was present. This group also had lower dwell counts for the images, overall suggesting that they were less interested in headline images than those with a positive attitude (Tables 5.23 and 5.24). When looking at the two groups separately, there was no significant effect of AI images or fake news, suggesting that the discrepancy between the groups is not due to external factors but to different attitudes.

Regarding the GSR metrics of peak count and peaks per minute, the results were in line with eye-tracking data. Separate tests for controlling AI images and fake news did not cause a significant difference in either Peak Count or Peaks Per Minute, however, there was a significant difference between those with a negative and positive attitude to AI.

Similarly to eye tracking, when looking within the groups, this difference vanished in most cases, suggesting that the reason for it is different emotional responses between those with a positive and negative attitude to AI. The only significant effect was when measuring peak count regarding AI images, in this case, those with a negative attitude had less peak counts on average when AI images were present compared to when real images were shown, further suggesting that AI images are not as important to this group.

6.2 Hypotheses Results

$\boldsymbol{H_1}$: The use of AI images decreases believability

Table 5.18 showcases the effect of AI images on both the whole data as well as on fake and true news separately. From this, we can see that on aggregate, having AI images decreased the believability score by 0.255, this did not differ much based on news types (0.286 for fake news and 0.239 for true news). We have also seen that there is no

significant difference between those with positive and negative attitudes to AI when controlled for news articles with AI images (Table 5.21).

$\boldsymbol{H_2}$: AI labels decrease believability when real images are present

In the overall findings, we saw a decrease of 0.197 in the score when looking at the effect of AI labels on real images ceteris paribus. (Table 5.18). Further investigation showed that regarding fake and true news, this effect was only significant in true news, and there was no significant difference at the 99% confidence interval for fake news. This suggests that when presented with fake news, subjects were using other signals to come to a decision. When looking at AI attitudes, we have also seen that labels have no significant effect on those who have a more negative view of AI. This could be explained by their adversity in seeing information related to AI. On the other hand, when examining those with a positive attitude, we could see a significant negative effect of having AI labels when real images were present.

$\boldsymbol{H_3}$: AI labels increase believability when AI images are present

Unlike when real images were shown, AI labels had no significant effect on news headline believability in connection to AI-generated images. There was no statistically significant difference between labeled and unlabelled images at the 99% confidence interval in any of the cases, whether looking at the overall picture or controlling for news types and AI attitudes. This carries an implication that similarly to what Moravec et al. (2018) showed, these simple labels are often not enough to affect consumers.

$\boldsymbol{H_4}$: Both AI images and AI labels increase attention and emotional response.

The final hypothesis concerned eye tracking and GSR measures. Regarding eye tracking, in the majority of the cases, there was no significant difference when controlling for variables such as AI images, AI labels, or fake news, neither in times to first fixation nor in dwell count. The only case when there was a difference was between groups separated by AI attitude. Further analysis showed that this difference was inexplainable via other known variables, suggesting that those with a negative attitude toward AI are

less fixated on AI-generated images in news headlines.

6.3 Theoretical Contributions

Spreading fake news and generating images using AI both have marginally low costs (Martens et al. (2018), Abduljawad and Alsalmani (2022)), while simultaneously promising both financial and political returns (Park et al., 2024). This comes at a cost for society, therefore it is important to set up effective barriers to slow down the spread of misinformation. Previous studies (Kim et al. (2021), Lazer et al. (2018)) showcased the need for improved media literacy and support of independent fact-checkers, while Moravec et al. (2018) showed that current labels are often not enough to change readers' perceived believabilities.

6.3.1 *Cognitive Dissonance*

This thesis aimed to analyze the change in behavior when looking at news headlines in a 2x2 design with repeated measures, focusing on AI-generated images and warning labels regarding AI. A preliminary experiment showed that dissonance may arise when reading political news. After controlling for variables, we could still see a difference between the believability scores given by more liberal and more conservative-leaning groups, and while they gave fake news lower scores on aggregate, this was only done to a smaller extent when they had to rate news that aligned with their views. This result supports the one in Figl et al. (2019), with the additional control variables for AI images and labels.

6.3.2 *Illusory Truth Effect*

Participants were not prompted to read the news or behave in any specific manner. For each page, they had to read a news headline of 16.1 words on average, process the

information, and select 3 values. On average, this took them only 21.6 seconds, suggesting that instead of fully comprehending each article's validity, they relied on heuristics to decide. This is in line with Fazio et al. (2015)'s findings. While no significant laboratory outcomes were obtained regarding political attitudes, the results of the thesis showcase that those with a negative attitude to AI-focused less on the images and also rated them less believable in some cases compared to those with a positive attitude. This can be explained by the illusory truth effect. Since participants took mental shortcuts when making a decision, those who trusted AI more were seen to rely on the images more.

6.3.3 *Artificial Intelligence*

Due to the nature of this research, Generative AI tools were both studied from a user and a researcher perspective. As a user who wanted to generate realistic photos of politically sensitive content, we have seen that while state-of-the-art tools (Abduljawad and Alsalmani, 2022) were restrictive, it took little effort to find online tools that had no built-in safeguards. While the images were used for research purposes, this highlights the danger of easily accessible AI tools that can be used maliciously. From a researcher's perspective, results similar to Kolomeets et al. (2024) were replicated. While in general AI images were distinguishable from real photos, this was especially true when faces were present. For developers, this can serve as a guideline on what areas to focus on when improving image-generating tools.

6.3.4 *Fake News*

This research could not have been possible without independent fact-checkers. Martens et al. (2018) argued for the need for increased resources for them, and this thesis supports this argument. When selecting news articles for the experiment, a reason for using headlines from United States politics was the wider availability of fact-checked articles. Throughout the setup of the research, the lack of independent fact-checkers

became a clear issue. This is especially true for smaller countries. The most prominent Danish fact-checking site, TjekDet reviewed only 28 news articles between June-August 2024 (https://www.tjekdet.dk/faktatjek, last accessed: 09/14/2024). Due to the effects of fake news (Kim et al., 2021) it is important that these organizations receive enough resources to identify misinformation. This can also help establish datasets of labeled news articles, increasing the potential of machine learning tools used in fake news research (Ruchansky et al., 2017).

6.3.5 *Information Systems*

The thesis also contributes to the research corpus of Information Systems. Regarding eye-tracking, while Vasseur et al. (2019) mentioned an observed increased attention when unconventional stimuli are shown, neither AI-generated images nor AI labels replicated these findings. Further research in this area focusing on either news headlines or AI images in other contexts might help answer if this lack of attention was due to the monotonous task of reading, or if AI-generated images have no significant effect on eye-tracking data in general. Similarly, while Lyulyov et al. (2024) found increased arousal levels when analyzing GSR metrics, the results for this were also inconclusive in our research. Based on this thesis, for both metrics, it is recommended for future research to control for attitudes to AI, since there was a significant difference between those with positive and negative attitudes.

6.4 Limitations and Future Work

6.4.1 *Limitations*

There are a number of trivial ways the research could have been improved. Having a more diverse, larger group of participants could lead to improved insights regarding political attitudes, however, this is costly and difficult to execute. The news headlines

were also concerned with US politics, while the participants were mainly European. This discrepancy was due to more extensive resources available for fact-checking regarding the United States, while the laboratory was based in Copenhagen. It could be a worthwhile avenue to explore how results change when participants are from the same country or region from which the news headlines are gathered.

6.4.2 *Future Work*

With the continuous improvements to AI image generators, it is expected that such images will be harder and harder to distinguish from real photos. Therefore the same experiment could be conducted in the future with state-of-the-art generators in order to see how results change as a function of image realism.

The concept of truthiness (Newman et al., 2012) is connected tightly to this thesis. While it has been shown that simply attaching an image to a statement increases believability, this has only been shown regarding real images. Research focusing on how AI-generated images influence this phenomenon could help understand more the influence pictures have on believing information. Potentially a similar research to this thesis could be connected with simply changing the binary measure of showing either real or AI images to showing real, AI, or no images, to see how the three scenarios are connected.

Finally, while the way news headlines were created in this research aimed to look similar to how they might appear on social media, they did not mimic exactly what a user might see on a platform. Research often mimics scenarios that participants might see on social media (such as a Facebook feed, Villota and Yoo (2018)), however, the purpose of this thesis was to look at the effects of AI images and labels in general. Future research can focus on the same concept, but with more specifications, either framing headlines as social media posts or experimenting with different types of labels.

6.4.3 *Policy Recommendations*

The main motivation behind this thesis was to understand the role of AI in the context of fake news, with a focus on social media. As AI tools become more accessible and more efficient, their potential dangers will increase. 4 groups of stakeholders are identified: social media platforms, AI platforms, social media consumers, and policymakers.

While social media platforms have introduced measures to fight misinformation, often these were shown to be inefficient (Moravec et al., 2018). With regards to AI content, this thesis has shown that right now headlines with AI images are less believable, and therefore expected to reach fewer people (Kolomeets et al., 2024). However, platforms should monitor them closely, focusing on separating real and AI-generated content, since the latter is cheaper to produce, and therefore can be exploited if it becomes more believable in the future. Meta (2024)'s initiative to have creators label AI content is a step in the right direction, however, it is necessary to focus more on those with a malcontent.

Platforms working on building AI models should also focus on setting up appropriate limits. Creators who aim to generate content for malcontent should be limited to what they can do. Often image generating platforms already stop users from being able to create certain images, such as ones depicting political figures (Abduljawad and Alsalmani, 2022), however, there are still online sites where this can be avoided, such as MUSA AI, the platform used for this thesis. While offline solutions are hard to identify and work around, setting up limits on major online sites is a step in the right direction, since it discourages the majority of people from engaging in creating harmful content.

Social media consumers should be aware of their implicit ways of engaging with news online. The illusory truth effect means that they might take shortcuts when fully understanding content, however, while it requires more attention, simply being aware of this can help overcome biases. Consumers should also be aware of the increasing number

of AI-generated content online and know that they are potentially used in adverse ways.

While social media and AI platforms can change their policies based on market incentives, often these are not enough to combat misinformation and the social harm it causes. As such, external interventions may be needed. Policymakers should encourage platforms to develop tools to identify fake news and to stop AI use for harmful intentions. Machine learning tools can be useful for fake news detection (Gupta et al., 2013), however, training sets are needed for this. One way policymakers can help is by increasing funding to independent fact-checkers (Martens et al., 2018). Finally, as previous literature (Serrano-Puche (2021), Burkhardt (2017), Kim et al. (2021)) suggested, a focus on improving media literacy can help consumers be more informed and decrease the spread of fake news.

7 Conclusion

Fake news is already a social issue (Morgan, 2018) and with its costs being potentially lowered by cheaper and more accessible AI tools, its effects are expected to increase Martens et al. (2018). It is therefore important to gain an understanding of how AI-generated images influence the believability of news headlines. This thesis focused on exploring the effects of AI images and labels in an experimental setting using a 2x2 design with repeated measures.

The thesis found that AI-generated images decrease the believability of both fake and true news. In a preliminary setting with a diverse range of participants, it was found that political attitude influenced the magnitude of this decrease. In line with previous studies of cognitive dissonance (Figl et al., 2019), the score for perceived believability decreases less when participants' views align with the news and more when they oppose it. This result could not be repeated in the laboratory setting, however, post hoc analysis showed that attitudes to AI influenced believability, as well as an emotional response. While attention and emotional response did not see changes within subjects (as measured by eye-tracking and GSR), those with a positive attitude to AI showcased more bodily responses than those with a negative attitude.

Labeling AI-generated content proved to have mixed effects. While they did not have an influence on AI-generated images, they decreased believability when real images were present (both in the cases of fake and true news). This was explained by the deception that participants might have perceived.

In conclusion, AI-generated images decrease the believability of news headlines. This implies that such news spreads slower on social media (Kolomeets et al., 2024) and therefore those with malicious intent might avoid using them for now. Despite this, it is expected that such images will become more realistic in the future, and due to their

lower costs, they might be the preferred way to generate harmful content soon. As such, it is important for both social media and AI platform owners to set up appropriate measures to combat this phenomenon. Both previous literature (Moravec et al., 2018) and this thesis show that simple labels are inefficient for signaling. As such, both more noticeable labels (Pennycook and Rand, 2017) and other methods should be explored in this regard. Policymakers should also focus on setting up guidelines as well as exploring how consumers can be warned against this phenomenon most effectively.

Bibliography

Abduljawad, M. and Alsalmani, A. (2022). Towards creating exotic remote sensing datasets using image generating ai. In *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 84–88. IEEE.

Aïmeur, E., Amri, S., and Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Ali Adeeb, R. and Mirhoseini, M. (2023). The impact of affect on the perception of fake news on social media: A systematic review. *Social Sciences*, 12(12):674.

Annor Antwi, A. and Al-Dherasi, A. A. M. (2019). Application of artificial intelligence in forecasting: A systematic review. *Available at SSRN 3483313*.

Aronson, E. and Mills, J. (1959). The effect of severity of initiation on liking for a group. *The Journal of Abnormal and Social Psychology*, 59(2):177–181.

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.

Beltramini, R. F. (1988). Perceived believability of warning label information presented in cigarette advertising. *Journal of Advertising*, 17(2):26–32.

Berghel, H. (2017). Oh, what a tangled web: Russian hacking, fake news, and the 2016 us presidential election. *Computer*, 50(9):87–91.

Bovet, A. and Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature Communications*, 10(1).

Bowman, S. and Willis, C. (2003). How audiences are shaping the future of news and information. *We Media*.

Brashier, N. M., Eliseev, E. D., and Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, 194:104054.

Burkhardt, J. M. (2017). Combating fake news in the digital age. *Library Technology Reports*, 53(8).

Cacioppo, J. T. and Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131.

Campan, A., Cuzzocrea, A., and Truta, T. M. (2017). Fighting fake news spread in online social networks: Actual trends and future research directions. *2017 IEEE International Conference on Big Data (Big Data)*, 2017.

Cooper, J. and Fazio, R. H. (1984). A new look at dissonance theory. *Advances in Experimental Social Psychology*.

Derevyanko, N. and Zalevska, O. (2023). Comparative analysis of neural networks midjourney, stable diffusion, and dall-e and ways of their implementation in the educational process of students of design specialities. *Scientific Bulletin of Mukachevo State University. Series "Pedagogy and Psychology*, 9(3):36–44.

Dewitte, S., Cornelis, J. P., Müller, R., and Munteanu, A. (2021). Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction. *Remote Sensing*, 13(16):3209.

Drover, W., Wood, M. S., and Corbett, A. C. (2018). Toward a cognitive view of signalling theory: Individual attention and signal set interpretation. *Journal of management studies*, 55(2):209–231.

Fazio, L. K., Brashier, N. M., Payne, B. K., and Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5):993–1002.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Festinger, L. and Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, 58(2):203–210.

Festinger, L., Riecken, H. W., and Schachter, S. (1956). *When prophecy fails.* University of Minnesota Press.

Figl, K., Kießling, S., Rank, C., and Vakulenko, S. (2019). Fake news flags, cognitive dissonance, and the believability of social media posts. *Proceedings of the Fortieth International Conference on Information Systems, Munich, Germany, 15–18 December 2019.*

Figueira, A. and Oliveira, L. (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121.

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc., 2nd edition.

Ghaffar Nia, N., Kaplanoglu, E., and Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3(1).

Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. (2013). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. *Proc. of the 22nd International Conference on World Wide Web.*

Gupta, P., Ding, B., Guan, C., and Ding, D. (2024). Generative ai: A systematic review using topic modelling techniques. *Data and Information Management*, 8(2):100066. Systematic Review and Meta-analysis in Information Management Research.

Harmon-Jones, E. and Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. *Cognitive dissonance: Reexamining a pivotal theory in psychology (2nd ed.).*

Hasher, L., Goldstein, D., and Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1):107–112.

Hasson, O. (1997). Towards a general theory of biological signaling. *Journal of Theoretical Biology*, 185(2):139–156.

High-Level Expert Group on Artifcial Intelligence (2019). A defnition of ai: Main capabilities and scientifc disciplines.

Himeur, Y., Rimal, B., Tiwary, A., and Amira, A. (2022). Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. *Information Fusion*, 86:44–75.

Iqbal, T. and Qureshi, S. (2022). The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528.

Joseph, A. W. and Murugesh, R. (2020). Potential eye tracking metrics and indicators to measure cognitive load in human-computer interaction research. *Journal of scientific research*, 64(01):168–175.

Kahneman, D. (2011). Thinking, fast and slow. *Macmillan*.

Kalota, F. (2024). A primer on generative artificial intelligence. *Education Sciences*, 14(2).

Kaur, N. and Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, 56(7):5837–5880.

Keselman, H., Algina, J., and Kowalchuk, R. K. (2001). The analysis of repeated measures designs: a review. *British Journal of Mathematical and Statistical Psychology*, 54(1):1–20.

Kim, B., Xiong, A., Lee, D., and Han, K. (2021). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *L. Lavorgna (Ed.), PLOS ONE*, 16(12):e0260080.

Kirby, E. J. (2016). The city getting rich from fake news. *BBC News*, 5.

Kolomeets, M., Wu, H., Shi, L., and Moorsel, A. (2024). The face of deception: The impact of ai-generated photos on malicious social bots.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Kopp, C., Korb, K. B., and Mills, B. I. (2018). Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to

Kshetri, N. and Voas, J. (2017). The economics of "fake news.". *IT Professional*, 19(6):8–12.

Lavorgna, L., De Stefano, M., Sparaco, M., Moccia, M., Abbadessa, G., Montella, P., Buonanno, D., Esposito, S., Clerico, M., Cenci, C., Trojsi, F., Lanzillo, R., Rosa, L., Morra, V. B., Ippolito, D., Maniscalco, G., Bisecco, A., Tedeschi, G., and Bonavita, S. (2018). Fake news, influencers and health-related professional participation on the web: A pilot study on a social-network of people with multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 25.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI open*, 3:111–132.

Lyulyov, O., Pimonenko, T., Infante-Moro, A., and Kwilinski, A. (2024). Perception of artificial intelligence: Gsr analysis and face detection. *Virtual Economics*, 7(2):7–30.

Mariani, M. M., Machado, I., and Nambisan, S. (2023). Types of innovation and artificial intelligence: A systematic quantitative literature review and research agenda. *Journal of Business Research*, 155:113364.

Martel, C., Pennycook, G., and Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1).

Martens, B., Aguiar, L., Gomez, E., and Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. *SSRN Electronic Journal*.

Meta (2024). Our approach to labeling ai-generated content and manipulated media. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/ last accessed: 09/09/2024.

Mirsky, Y. and Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41.

Moravec, P., Kim, A., and Dennis, A. R. (2018). Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *SSRN Electronic Journal*.

Morgan, S. (2018). Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, 3(1):39–43.

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., and Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin amp; Review*, 19(5):969–974.

Newman, E. J., Jalbert, M. C., Schwarz, N., and Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of need for cognition. *Consciousness and Cognition*, 78:102866.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. (2024). Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988.

Paul, T., Bhattacharyya, C., Sen, P., Prasad, R., and Shaw, S. (2020). Human emotion recognition using gsr and eeg.

Peltzman, S. (2019). Political ideology over the life course. *SSRN Electronic Journal*.

Pennycook, G., Cannon, T. D., and Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12):1865–1880.

Pennycook, G. and Rand, D. G. (2017). Assessing the effect of 'disputed' warnings and source salience on perceptions of fake news accuracy. *SSRN Electronic Journal*.

Pennycook, G. and 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. The Cognitive Science of Political Thought.

Păvăloaia, V.-D. and Necula, S.-C. (2023). Artificial intelligence as a disruptive technology—a systematic literature review. *Electronics*, 12(5):1102.

Reuer, J. J., Tong, T. W., and Wu, C.-W. (2012). A signaling theory of acquisition premiums: Evidence from ipo targets. *Academy of Management Journal*, 55(3):667–683.

Ross, S. and Morrison, G. (2003). *Experimental Research Methods*. Routledge.

Ruchansky, N., Seo, S., and Liu, Y. C. (2017). A hybrid deep model for fake news detection. *Proc. of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*.

Sanchez-Comas, A., Synnes, K., Molina-Estren, D., Troncoso-Palacio, A., and Comas-González, Z. (2021). Correlation analysis of different measurement places of galvanic skin response in test groups facing pleasant and unpleasant stimuli. *Sensors*, 21(12):4210.

Sarker, I. H. (2022). Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3(2).

Serrano-Puche, J. (2021). Digital disinformation and emotions: exploring the social risks of affective polarization. *International review of sociology*, 31(2):231–245.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1).

Sheikh, H., Prins, C., and Schrijvers, E. (2023). Artificial intelligence: Definition and background. *Research for Policy*, page 15–41.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Spence, A. (1974). Market signaling: Informational transfer in hiring and related screening processes.

Turel, O. and Kalhan, S. (2023). Prejudiced against the machine? implicit associations and the transience of algorithm aversion. *MIS Quarterly*, 47(4):1369–1394.

Vasseur, A., Léger, P.-M., Senecal, P. D., et al. (2019). Eye-tracking for is research: A literature review.

Vaswani, A. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vijh, M., Chandola, D., Tikkiwal, V. A., and Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia computer science*, 167:599–606.

Villota, E. J. and Yoo, S. G. (2018). An experiment of influences of facebook posts in other users.

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

Welch, B. L. (1947). The generalization of 'Student's' problem when several fidderent population varlances are involved. *Biometrika*, 34(1-2):28–35.

Wiesenfeld, B., Wurthmann, K., and Hambrick, D. (2008). The stigmatization and devaluation of elites associated with corporate failures: A process model. *Academy of Management Rev.*, 33(1):231–251.

Wolters, H., Stricklin, K., Carey, N., and McBird, M. K. (2021). The psychology of (dis)information: A primer on key psychological mechanisms. *Center for Naval Analyses*.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.