Empowering Users to Handle Misinformation in Podcasts

by

Ee Xuan Tan

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday December 13, 2024 at 14:00 PM.

Student number:	4907531	
Project duration:	March 15, 2024 - December 13, 2024	
Thesis committee:	Prof. dr. ir. U. Gadiraju,	TU Delft, Chair, Thesis Supervisor
	Prof. dr. ir. Z. Yue,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

First and foremost, I would like to express my heartfelt gratitude to my parents and my brothers for all the sacrifices they have made. Without their support, I would not be in the position I am today. Thank you.

I am deeply grateful to my Thesis Supervisor, Professor Gadiraju, for your invaluable guidance, constructive feedback, and the time you dedicated to supporting my work. I would also like to extend my appreciation to Garrett Allen for offering me endless support and always prioritizing my learning and growth throughout this journey. A special thanks to Professor Yue for taking the time to be a part of my thesis committee. Finally, I want to thank all the friends I have made along the way that have made the experience better.

Ee Xuan Tan
Delft, December 2024

Abstract

Podcasts are a rapidly growing medium for information sharing, but their audio and one-way communication format presents unique challenges in addressing misinformation. This thesis explores how to empower podcast listeners to identify and respond to misinformation effectively. Study I investigates listening habits, user trust, confidence, and behavioral responses to misinformation in podcasts through a survey of diverse participants. Key findings highlight gaps in user confidence, the impact of demographic factors, and preferences for incentives to flag misinformation. Study II builds upon these insights to design, implement, and evaluate three interventions—PAUSE, ALERT, and VOLUNTARY—aimed at optimizing user engagement in flagging misinformation. A labeled podcast dataset was created to facilitate this task-based experiment. The findings offer insights into the design of user-centric misinformation detection systems. Interventions have shown potential in empowering users to identify misinformation in podcasts. Although, whether they are able to address misinformation in podcasts effectively remains uncertain and needs further exploration. This work not only addresses a significant gap in the literature but also lays the groundwork for future innovations in combating misinformation in podcasts.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Misinformation 3
2.2 Existing Methods for Handling Misinformation 4
2.3 Crowd Sourced Solutions for Detecting Misinformation 5
2.4 Feedback Elicitation Design 6
2.4.1 User Engagement 6
2.4.2 Existing User Feedback Strategies 7
2.4.3 Audio Interventions 8

3 Study I: How do users respond to misinformation in podcasts? **9**

3.0.1 Demographics and Listening Habits 9
3.0.2 Trust and Confidence 9
3.0.3 Reaction to Misinformation in Podcast 10
3.0.4 Incentives to Flag Misinformation in Podcasts 10
3.1 Methodology 10
3.1.1 Cognitive Bias Checklist 10
3.1.2 Participant Recruitment 12
3.1.3 Pilot Study 12
3.2 Study Results 12
3.2.1 Demographics and Listening Habits 12
3.2.2 Trust and Confidence 13
3.2.3 Reaction to Misinformation in Podcasts 15
3.2.4 Incentives to Flag Misinformation in Podcasts 16
3.2.5 Limitations 17
3.3 Conclusion 17

4 Study II: How can we optimize user engagement to flag misinformation? **19**

4.1 Hypotheses 20
4.2 Experimental Setup 21
4.2.1 Independent Variables 21
4.2.2 Dependent Variables 21
4.2.3 Podcast Player Interface 22
4.2.4 Cognitive Bias Checklist 23
4.3 Labeled Podcast Dataset 24
4.4 Participant Recruitment 25
4.5 Results and Discussions 25
4.5.1 Demographics 25
4.5.2 Intervention User Engagement 25
4.5.3 Intervention User Experience 27
4.5.4 Flagging Accuracy 29
4.6 Future works 31
4.6.1 labeled podcast dataset 31
4.6.2 Application Design 31
4.6.3 Flagging Interventions 31
4.7 Conclusion 31

5 Conclusion **33**

A Study I: Consent Form **35**

B Study I: How do users respond to misinformation Survey **37**

C Study I: Chi Square Test - Genre/Listening Intent and Misinformation Reaction **39**

D Podcast Data creation prompts **41**

E Podcast Counts by Topic and Truth Label **43**

F Study II: Consent Form **45**

G Study II: Podcast Player Task Description **47**

H Study II: Post-task survey **49**

I Study II: Survey results - Intervention Engagement **51**

1 Introduction

In the digital age, podcasts have emerged as a powerful medium for sharing information, ideas, and stories [68]. With millions of episodes available on diverse topics [66] and the number of listeners nearly doubling since 2019 [65], podcasts have become a daily source of information covering a large range of subjects such as politics, sport, and news. Spotify, the largest audio podcast platform gathers over 100 million regular listeners [64]. However, with this vast and rapid spread of information, the challenge of misinformation has become increasingly important, especially when it poses a threat to people.

One example of this is the controversy surrounding 'The Joe Rogan Experience' podcast during the COVID-19 pandemic. The show hosted virologist Robert Malone, who expressed controversial views questioning the safety and efficacy of COVID-19 vaccines [54]. This episode sparked widespread outrage, leading healthcare professionals to urge Spotify to address COVID-19 misinformation on its platform [55]. With millions of listeners tuning in, podcasts discussing critical topics such as health have the power to shape public perceptions, influencing how people understand, trust, and respond to information.

The problem of misinformation extends beyond health-related topics, impacting other parts of society. Misinformation has the potential to alter voter behavior, undermine trust in institutions, and polarize societies. This is especially relevant in the recent 2024 U.S. elections, as misinformation has previously played a role in shaping public discourse and influencing outcomes, such as during the 2016 U.S. election, which faced significant challenges [3]. Its impact has also been observed globally, such as in the 2018 Italian elections, where misinformation had influenced voter perceptions and decisions [14].

Furthermore, with tools like Google's NotebookLM enabling rapid content creation, such as generating podcasts from notes in minutes [26], addressing misinformation becomes progressively more challenging to deal with. How we manage and respond to misinformation in growing mediums like podcasts becomes increasingly critical for navigating the digital information space.

Misinformation in the context of podcasts poses a distinct challenge for two key reasons. First, the rapid spread of information is attributed to the ease with which one can create and distribute their podcast to a wide audience. As a result, unverified claims can reach an audience more quickly. Second, within the podcasting ecosystem, the dynamic between publishers and audiences is different from social media platforms. While anyone can create and share content, audiences typically lack a direct way to respond. This limits the collective fact-checking potential seen on other social media platforms, such as Facebook and X (formerly known as Twitter) where content publishers and viewers interact by engaging in conversations, through real-time messaging or comment sections [15].

In addition to the rapid rate at which podcasts scale and the different dynamics between publishers and listeners, they differ significantly from traditional forms of media like social media. Podcasts convey information primarily through audio, making them distinct in how misinformation can be identified and managed. Conventional methods, such as visual cues in the forms of hyperlinks, citations, or warning labels [42], are commonly used to warn of potential misinformation in visual media. However, in an audio format, implementing similar cues presents unique challenges, as auditory warnings could be disruptive and distracting for listeners. This distinct format needs new approaches to be able to deal with the challenges of misinformation.

This study aims to tackle this problem by exploring the possibilities of empowering listeners to be able to identify misinformation in podcasts, despite the inherent challenges of the audio format and the one-way nature of communication in this medium. To tackle this research gap; we aim to answer the following question:

- **How can we empower users to handle misinformation in podcasts?**

The goal of this study is to develop an intervention that empowers users to effectively flag misinformation. To achieve this, we first examine existing methods for addressing misinformation in podcasts and other social media platforms. Additionally, we investigate how users react to misinformation, explore approaches to elicit feedback from users, and analyze what drives user engagement. This exploration revealed a gap in the current understanding of how users respond to misinformation, particularly within the podcasting context. To address this gap, we begin by answering the following question:

- **Study I: How do users respond to misinformation in Podcasts?**

To answer this question, we conducted an exploratory survey on the Prolific platform to investigate users' levels of trust, confidence, and reactions to misinformation in podcasts, as well as their opinions on potential incentives for flagging misinformation. This study makes a key contribution by providing a preliminary understanding of user responses to misinformation in podcasts, thereby addressing a critical gap in the existing literature. Using the insights gained from this survey, we inform the design of interventions aimed at empowering users to flag misinformation effectively, directly contributing to the objectives of our second study.

- **Study II: How can we optimize user engagement to flag misinformation?**

To address this research question, we developed three interventions and conducted a task-based experiment to evaluate their effectiveness. The primary aim of the experiment was to assess user engagement and accuracy across the three different interventions while considering the overall user listening experience. To facilitate this, we created a labeled podcast dataset as part of this study. This dataset was generated using the LIAR text dataset [72], combined with GPT-4 [49] and Google's Text-to-Speech tool [16], to produce short podcast segments with annotated timestamps. This dataset is the first contribution to this study and hopefully provides a foundation for future research exploring misinformation in podcasts.

Using the dataset, we designed a two-part experiment. In the first part, participants listened to six different podcast segments and were tasked with identifying as many instances of misinformation as accurately as possible. In the second part, participants completed a post-task survey to evaluate their experience with the intervention. This approach allowed us to gain insights into the effectiveness and disruptiveness of each intervention. This leads us to the second contribution of this study: the exploration, development, and evaluation of interventions aimed at optimizing user engagement for flagging misinformation in podcasts.

2

Related Work

2.1. Misinformation

The field of misinformation research faces challenges in establishing clear definitions, terms like "misinformation," "disinformation," and "fake news" are frequently used without consistency [12]. Broda and Strömbäck [12] defines misinformation as false information shared without intent to harm, typically due to misunderstandings or incomplete knowledge. At the same time, disinformation involves intentionally fabricated or manipulated content designed to mislead for purposes such as political or financial gain [8]. Fake news refers to intentionally false or misleading information designed to resemble real news, characterized by low factual accuracy and a strong intent to deceive. In this study, we focus on empowering users to address information that is false, regardless of intent. Consequently, the term "misinformation" will be used as an umbrella term to encapsulate all related concepts.

Misinformation affects trust, fuels division, and creates challenges across various domains such as politics, economics, and science [2]. Belief in misinformation is attributed to a mix of social, cognitive, psychological, and environmental influences [8]. Beauvais [8] emphasizes the role of social media platforms, where unfiltered, high-speed dissemination of information, especially during major events like elections and the COVID-19 pandemic, increases exposure to misinformation. These dynamics are further compounded by psychological factors, such as an attraction to emotionally charged content, which contributes to individuals' susceptibility to misinformation [41]. Among these effects, cognitive biases can play a role in shaping how information is processed and accepted. Biases such as confirmation bias, the illusory truth effect, and motivated reasoning increase individuals' susceptibility to repeated exposure to false information. The illusory truth effect leads people to accept information as true simply because it feels familiar through repetition, regardless of its accuracy. Motivated reasoning further amplifies this effect by influencing individuals to interpret information in ways that affirm their pre-existing beliefs or identities. Combined with echo chambers, where individuals are exposed only to similar opinions and beliefs while often excluding opposing viewpoints, these biases contribute to the rapid spread of misinformation online [22].

Echo chambers reinforce existing beliefs and limit exposure to diverse perspectives, creating isolated networks that amplify misinformation [7]. Ecker et al. [22] discusses a key strategy to counter the illusory truth effect and prevent repeated exposure from increasing its perceived accuracy is tagging misinformation early.

Ecker et al. [22] discusses tagging misinformation early as a key strategy to counter the illusory truth effect and prevent the perceived accuracy from increasing with repeated exposure. Furthermore, platform interventions, such as accuracy prompts, can encourage users to verify information before sharing, reducing the spread of misinformation driven by familiarity [71]. Similarly, addressing echo chambers requires promoting critical thinking by teaching individuals to evaluate evidence critically and providing transparent, well-sourced counter-information to expose false claims. Reducing social validation loops, such as limiting algorithmic reinforcement of likes and shares can also disrupt the echo chamber effect by exposing users to a broader range of perspectives [58]. Lee et al. [33] and Guo, Schlichtkrull, and Vlachos [27] explores varying effectiveness between debunking (fact-checking after exposure) and prebunking (preemptive strategies), depending on the context and type of misinformation. Proposed solutions include promoting media literacy, regulating social platforms, and establishing partnerships with fact-checkers to limit the spread of misinformation.

The challenges posed by misinformation are magnified by cognitive biases, social dynamics, and the rapid dissemination of information, as seen on social media platforms [22, 7]. Platform-level interventions, like tagging misinformation and implementing accuracy prompts, have been proposed to mitigate these effects by encouraging users to verify content and engage critically with information [71, 58]. However, the effectiveness of such solutions often depends on the context, highlighting the importance of understanding how users interact with misinformation in specific environments.

Existing studies, such as Bode and Vraga [11] work on misinformation during COVID-19, provide insights into how users engage with misinformation in social media contexts. Their findings reveal that users respond to misinformation by witnessing corrections, actively correcting others, or being corrected themselves. However, these responses often have limited effectiveness due to biased corrections, a lack of credible sources, or the influence of social dynamics on how corrections are perceived. For example, witnessing corrections frequently has little impact on observers' beliefs, and being corrected rarely changes misperceptions.

Additional research on social media platforms like X (formerly Twitter), Facebook, and Instagram has explored the challenges users face in recognizing and responding to misinformation. Users on these platforms often struggle to identify false information due to cognitive biases, limited media literacy, and the subtle nature of some misleading content. This "recognition gap" means that many individuals are unaware that the content they encounter is false, reducing the likelihood that they will challenge or correct it. Even when misinformation is recognized, key barriers such as social concerns (fear of conflict or embarrassment), effort-related considerations, and a lack of knowledge to effectively counter false claims further prevent users from acting [31].

Unlike social media, podcasts are an audio-first, one-way communication medium, which presents unique challenges for misinformation. For instance, users cannot rely on text-based fact-checking tools or visual cues, making it harder to evaluate the accuracy of spoken content in real-time. Additionally, the conversational tone and perceived credibility of podcast hosts may influence trust and make misinformation harder to detect [37]. While research has explored these challenges extensively in the context of social media, little is known about how users respond to misinformation in podcasts, a medium that has rapidly gained popularity.

To address this gap, we conducted a survey to explore user trust, concern, and reactions when encountering misinformation in podcasts. The design and findings of our survey are detailed in Chapter 3. Focusing on this relatively unexplored medium, our study seeks to provide new insights into user behavior in podcasts.

2.2. Existing Methods for Handling Misinformation

The process of fact-checking is critical in addressing misinformation. Professional fact-checkers demonstrate that it is a complex, labor-intensive process involving selecting claims, contextualizing content, consulting experts, and preparing reports [46]. Fact-checkers face significant challenges, particularly latency in producing fact-checks and resource constraints that limit both speed and coverage, leaving them unable to keep up with the rapid production of new content. Furthermore, fact-checkers can be affected by perception biases, where one is unconsciously influenced by assumptions and expectations. [43, 27, 18]. Many fact-checkers rely on simple, manual tools, while artificial intelligence (AI) solutions are viewed skeptically due to concerns over accuracy and lack of nuance [46]. Another challenge is the spread of verified information to the public, current methods lack reach due to echo chambers that are formed by social media. Another challenge is the dissemination of verified information to the public; current methods often lack sufficient reach, partly due to the echo chambers formed on social media platforms [35]. Addressing these issues would require more efficient tools, resources, and dissemination methods to strengthen the impact of professional fact-checking.

The work of Guo, Schlichtkrull, and Vlachos [27] on a survey for automated fact-checking discusses the challenges of fact-checking. Given the rapid spread of information and misinformation, manual fact-checking is time-consuming and cannot keep up with the volume of information. The study outlines a structured framework for automating fact-checking, which includes stages such as claim detection, evidence retrieval, claim verification, and justification production. Incorporating these automated stages into user engagement strategies could empower podcast listeners to critically assess content by offering specific, interactive features within the podcast app. For instance, listeners could help verify flagged statements that appear dubious and receive prompts with links to credible sources that either support or refute these claims (evidence retrieval). Additionally, a verification tool could display summaries of related fact-checked information for commonly debated topics within the episode (claim verification), while text-based justifications clarify why certain claims might be misleading (justification production). By seamlessly integrating these stages, listeners would have immediate access to tools that support informed decision-making, enhancing their ability to discern misinformation within the unique audio format of podcasts.

Key challenges identified in the survey include dataset bias, managing multilingual and multi-modal content, and developing real-time fact-checking capabilities. Real-time capabilities are especially important in the podcast context, as information spreads rapidly. While many podcasts are in English, a significant portion are produced in other languages, introducing additional challenges for accurate and inclusive fact-checking [38].

Furthermore, automated fact-checking methods, which are largely text-based, struggle to meet the unique demands of audio formats like podcasts. Limited research and initiatives on non-text formats add to these challenges, making adapting current methods for effectively fact-checking podcast content difficult. Existing models like AFCNR, BRENDA, and ClaimPortal handle only textual input [46], highlighting this limitation. Squash [1], a model designed to process audio by transcribing and cross-referencing statements with fact-checked data, faces critical obstacles, such as frequent transcription errors and a limited fact-check database that often leaves the model idle, and unreliable matching algorithms that sometimes associate claims with irrelevant fact-checks all restrict the model's effectiveness. Additionally, Squash still requires human oversight to manage inaccuracies, hindering its potential for fully autonomous fact-checking [1].

Adair [1] also highlights the importance of integrating human fact-checkers to verify nuanced contexts or ambiguous information that AI models may struggle to assess effectively. It suggests a hybrid model where humans act as a check to ensure accuracy, transparency, and alignment with ethical standards, particularly in situations where models might fail to understand subtle cues such as sarcasm or contextual nuances [27]. Similarly, Sethi [61] emphasizes the need for critical thinking to effectively combat misinformation, pointing out that existing automated tools for detecting fake news primarily focus on identifying hoaxes or tracking viral spread but cannot verify underlying alternative facts. These perspectives underscore the importance of a hybrid approach that combines human expertise with automated systems to address these limitations.

A human-in-the-loop faces similar challenges to professional fact-checkers, such as scalability and resource demands. Social media platforms Facebook and X (formerly Twitter) have addressed this issue by leveraging their user base to flag misinformation, which allows for a more distributed and responsive approach [37]. Facebook further enhances fact-checking efforts by prioritizing and rapidly surfacing high-priority claims, enabling fact-checkers to respond quickly to misleading information whenever it appears [46]. Exploring how a similar approach could be adapted for the podcast context, engaging listeners to flag misinformation in real-time, presents an opportunity to challenge the one-way communication of podcasts.

2.3. Crowd Sourced Solutions for Detecting Misinformation

Crowdsourcing, defined as engaging a large group of people to contribute knowledge, complete tasks, or solve problems [18], has been widely explored as a scalable and cost-effective approach to misinformation detection [47, 10, 32, 39]. It leverages user-generated signals, such as flags, to identify potentially false content, which can then be verified by experts or supplemented with algorithmic techniques. This approach has demonstrated significant advantages in managing the challenges posed by the vast volume of content shared online, making it particularly relevant for contexts like misinformation detection in podcasts [39, 10].

One of the key strengths of crowdsourcing is its scalability, as it distributes the workload across a broad user base. This allows platforms to process large datasets without relying exclusively on resource-intensive expert validation. Additionally, the diversity of perspectives contributed by users enhances the ability to identify misinformation by incorporating varied insights and experiences. For example, Martel et al. [43] investigated the ability of non-experts to identify misinformation and demonstrated that aggregated trust ratings from crowds closely matched the evaluations of professional fact-checkers. Participants were tasked with assessing the trustworthiness of various types of news content, including credible mainstream articles, hyperpartisan pieces, and outright fake news. The study found that non-experts could reliably differentiate between these categories, with their collective judgments effectively identifying misinformation.

In citizen journalism, crowdsourcing has proven its potential to improve accuracy and transparency. During the Greek COVID-19 pandemic, citizen-generated content provided a more accurate understanding of events compared to traditional media, which were also susceptible to spreading misinformation [32]. The 'Nea Smyrni riots' highlighted how videos and social media posts from non-professional sources shed light on critical issues, such as police violence, that were initially downplayed by mainstream outlets. This example shows how crowdsourcing can challenge traditional media by leveraging diverse perspectives to present a more accurate representation of events [32].

Apart from hybrid solutions with experts, we also see hybrid solutions with algorithms that leverage the strengths of both human input and computational solutions. These systems integrate crowdsourced judgments with advanced algorithms to improve scalability and accuracy across different domains. For instance, Tschiatschek et al. [69] demonstrated how the DETECTIVE algorithm uses Bayesian inference to assess user reliability and filter out misleading signals, tackling environments where there are adversarial users. Similarly, Nanath and Olney [47] showed how combining crowdsourcing with machine learning enhanced fraud detection in job postings, with hybrid models outperforming standalone algorithms by incorporating crowd-generated metrics like Net Promoter Scores. Such hybrid solutions allow for the efficient handling of large datasets while capturing nuanced human perspectives. Hybrid models, such as those outperformed standalone algorithms, showcasing the applicability of crowdsourcing in enhancing algorithmic predictions. These findings show the potential of combining human input with algorithms to tackle complex challenges across various domains.

Despite its advantages, crowdsourcing faces several limitations that require careful attention. A key challenge lies in balancing monetary incentives with ethical considerations and task suitability [59]. While monetary rewards effectively engage participants, they often prioritize cost-efficiency over fairness, leading to low wages and potential exploitation of workers. This raises ethical concerns, particularly when participants are undercompensated or lack transparency regarding task expectations and payment structures. Task design plays a critical role; as Bhuiyan et al. [10] highlighted, the suitability of crowdsourced tasks depends on the expertise required. Non-experts excelled in evaluating simpler tasks, such as opinion pieces, but struggled with deeper analyses, like assessing scientific claims. Additionally, over-reliance on financial incentives can overshadow intrinsic motivators, such as personal interest or social impact, which are crucial for ensuring high-quality contributions, especially in complex tasks. This primary motivator could impact the transferability to natural user engagement, tasks like flagging content could result in limited or incomplete data collection. Furthermore, the reliability of crowdworkers varies—while some provide accurate assessments, others may contribute incorrect or adversarial signals, introducing vulnerabilities such as spam or low-quality inputs [69]. One way to address reliability and low-quality inputs, could potentially be to explore hybrid solution with expert validation or AI models.

The studies described in this section highlight the potential of crowdsourcing as a scalable, cost-effective tool for misinformation detection. As an audio-based medium, podcasts present unique challenges due to their one-way communication format and the difficulty of verifying spoken content. Utilizing crowdsourcing offers a means to leverage diverse listener perspectives to flag misinformation effectively. By exploring its effectiveness and discussing its limitations, crowd-sourcing can be a powerful strategy to ensure transparency and accuracy in podcast content.

2.4. Feedback Elicitation Design

2.4.1. User Engagement

User engagement is crucial for social network platforms as it drives loyalty, trust, and advocacy, while also enhancing the platform's perceived quality and utility. In the competitive digital landscape, brands that prioritize user engagement gain a significant advantage, as it not only strengthens customer relationships but also translates into measurable business success [21].

User engagement can take various forms in different contexts, such as through personalized feeds within eHealth networks [34]. Engagement ranged from passive interactions, like clicking on feed items, to active participation, such as contributing blog posts and engaging in forum discussions. By tailoring content to individual users through a predictive ranking mechanism, the system successfully increased engagement and information's perceived relevance. Optimizing user engagement is essential to enhance the user experience by delivering content that resonates with personal interests but also to drive active participation, which is critical for sustaining online communities. In settings like eHealth, where user contributions are vital for knowledge sharing and support, effective engagement ensures the platform remains a dynamic and valuable resource.

The forms of user engagement also vary significantly across social media platforms and content types. For example, richer media formats like videos often drive active engagement behaviors such as commenting, while simpler formats like photos are more effective for passive engagement, such as liking [62]. This variation is platform-specific, with desktop users on Facebook engaging more actively than mobile users on Instagram, demonstrating that engagement behaviors are influenced by both content type and user experience. Optimizing user engagement enhances the reach and effectiveness of content, allowing platforms and creators to tailor strategies for specific audiences. However, potential drawbacks must be considered. Optimizing for engagement metrics can sometimes lead to superficial interactions, where content prioritizes immediate reactions (e.g., clickbait) over meaningful exchanges. Lag effects, where engagement with one post influences subsequent posts, may create pressure for creators to continuously outperform themselves, potentially leading to unsustainable practices. Additionally, platform bias may arise, favoring certain user behaviors or content formats at the expense of others, potentially limiting diversity and thoughtful interaction. Balancing these considerations is key for deeper interactions while maintaining content quality and relevance [10].

While engagement strategies are often tailored to the unique characteristics of social media platforms and their audiences, user engagement extends beyond this. Effective engagement is also heavily influenced by the design and usability of the platforms themselves. Web design elements play a critical role in shaping user experiences, guiding interactions, and fostering sustained engagement. Garett et al. [25] identify seven critical elements of website design that significantly influence user engagement, emphasizing their importance in crafting effective and engaging digital platforms. These elements include site navigation, graphical representation, content organization, the utility of content, purpose, simplicity and readability. The study emphasizes the interconnectedness of these elements. For example, navigation aids, such as visible links, not only enhance usability but also improve the site's organization and visual appeal. While the paper does identify the key elements, the paper also specifies the gap for defining clear, measurable standards for these design elements. However, these findings stress the necessity of optimizing these design elements to optimizee for user engagement, trust, retention, and overall satisfaction. Furthermore, they advocate for tailoring design strategies to align with objectives, ensuring that digital experiences effectively meet user needs [25].

Customer engagement manifests in various forms across social media platforms, influenced by factors such as content type, message framing, and platform-specific user preferences. For example, on Instagram, subjective and positively framed posts tend to drive engagement through likes, shares, and comments, aligning with the platform's younger demographic and visually driven culture. In contrast, on Twitter, objective and neutral posts resonate more with its older audience, fostering engagement through discussions and fact-sharing. The role of influencers also varies; popular influencers with high credibility can significantly amplify engagement by leveraging their reach and trustworthiness. Optimizing customer engagement is essential because it creates stronger emotional and cognitive connections between users and brands. This leads to increased loyalty, advocacy, and trust, which are critical for long-term success. Moreover, tailoring engagement strategies based on platform dynamics and audience behavior allows brands to maximize their impact, ensuring that content not only captures attention but also sustains user interaction. By understanding these diverse forms of engagement and adapting strategies accordingly, brands can enhance their digital presence, foster meaningful relationships, and achieve measurable outcomes like higher retention and improved brand perception [23].

2.4.2. Existing User Feedback Strategies

In the context of user engagement, feedback mechanisms have shown promise in shaping behavior and sustaining attention. Effective strategies include real-time performance feedback, which reinforces positive behaviors, and reward mechanisms, such as gamified points or achievements, that give a sense of accomplishment [67]. Social approval mechanisms, such as peer recognition, further enhance user focus and motivation. However, poorly designed feedback systems risk distracting users, emphasizing the importance of thoughtful and minimally intrusive designs [67].

In software engineering and user experience design, feedback mechanisms can be categorized into push and pull systems, each with distinct advantages and challenges. Push systems actively request input from users through prompts, surveys, or notifications, which, while proactive, risk causing feedback fatigue if overused or poorly timed. Push systems can encourage critical thinking, which has been explored in many different contexts [56, 45, 29]. Prompts and mechanisms that encourage users to question the validity of the information they encounter are crucial in fostering engagement and careful consideration when encountering information [56]. These prompts are designed to shift users toward System 2 thinking, engaging them in deliberate and reflective cognitive processes. Rather than relying on intuitive or automatic judgments, such prompts aim to encourage users to analyze the content critically, evaluating its accuracy and rationale before taking action. For example, awareness training messages or contextual prompts might ask users to reflect on why a particular piece of information is flagged or to identify evidence supporting or refuting its claims [45]. By requiring active engagement, these push systems align with the principles of System 2 cognition, which has been shown to be more effective in mitigating the spread of misinformation [45]. Through thoughtful design, such interventions ensure that users are not only alerted to potential misinformation but are also nudged with the cognitive tools to evaluate the segment carefully.

Pull systems, on the other hand, allow users to provide input at their discretion, typically through feedback forms, but often result in fewer or delayed responses. Maalej, Happel, and Rashid [40] explores hybrid approaches, such as context-aware push systems, that aim to balance these strategies by prompting feedback only during appropriate moments, minimizing disruption while maintaining engagement.

In the context of podcasts, encouraging accuracy assessments and rationales could be integrated into an intervention, prompting users to reflect on the podcast content. Training interventions, such as short, targeted messages before or during the podcast, could educate listeners on identifying misinformation and applying critical thinking. Gamified feedback mechanisms, including real-time acknowledgments for flagging misinformation or tracking their engagement, could sustain user motivation and participation. For example, listeners could receive achievements for accurately identifying misinformation, giving users a sense of accomplishment.

2.4.3. Audio Interventions

Research on audio interventions reveals their potential for influencing user behavior in real-time scenarios. Auditory warnings, when carefully designed, are effective because they capture attention instantly and provide immediate feedback, prompting user reaction. For instance, studies such as [9] demonstrate that auditory signals can serve as a stimulus to discourage certain behaviors, like smoking indoors, while promoting adherence to desired actions. The success of such audio interventions relies on a careful balance between urgency and tolerability. Sounds that are too unpleasant risk disengagement, whereas calibrated signals can elicit behavior change. Interventions like the Mindless Attractor leverage subtle auditory perturbations, such as changes in pitch or volume, to refocus user attention during video-based learning sessions, avoiding the frustration often caused by explicit alerts [4]. Explicit alerts, while noticeable, are perceived as intrusive and disruptive, especially when triggered unnecessarily by false positives, leading to user frustration and loss of trust [4]. Audio interventions like these are particularly effective for podcasts because the medium relies entirely on sound.

3

Study I: How do users respond to misinformation in podcasts?

There is a notable gap in existing literature regarding how misinformation in the context of podcasts is addressed. To bridge this gap, we designed a survey with the primary objective of understanding public response to misinformation encountered in podcasts. By crowdsourcing the survey to participants, this research seeks to provide insights that could help and inform design choices for flagging interventions that empower listeners to actively engage in identifying and flagging misinformation in podcasts. Towards this end, we conducted a survey that aims to gather perspective into the following four categories; **Demographics and Listening Habits**, **Trust and Confidence**, **Reaction to Misinformation in Podcast** and **Incentives to flag Misinformation in Podcasts**. A full list of the survey questions is provided in the Appendix B.

3.0.1. Demographics and Listening Habits

The first portion of the survey captures essential background information about respondents, specifically their age, gender, and education level. The Listening Habits section gathered information about respondents' podcast consumption patterns and preferences. These included how often and how many hours per week they listen to podcasts, the types of podcasts they typically enjoy (e.g., Educational, Entertainment, Political) with the option to specify additional genres, their motivations for listening (e.g., Entertainment, Educational, Relaxation), how they discover new podcasts (e.g., recommendations, podcast platforms), and the activities they engage in while listening (e.g., commuting, exercising). These questions provide context about the demographics and listening behaviors of the respondents, which forms the basis for analyzing the study findings. By capturing characteristics such as listening frequency, podcast preferences, motivations, discovery methods, and activities while listening, this information reveals the audience's characteristics and helps assess how broadly the insights can be generalized. Furthermore, understanding these habits and demographic factors enables a nuanced interpretation of responses, particularly in evaluating how they might influence reactions to misinformed podcast content.

3.0.2. Trust and Confidence

The Trust & Confidence section of the survey seeks to understand participants' trust in podcast content and their level of confidence in identifying misinformation. We examine four dimensions of trust: trust in podcasts overall, trust in podcasts that participants listen to, trust in the methods used to discover new podcasts, and the key factors influencing one's trust. We aim to uncover the role trust plays in shaping behavior around misinformation. Prior research has shown trust can significantly impact how individuals engage with and share information. High trust often leads to overconfidence in the accuracy of information, reducing users' motivation to verify content critically, while low trust may discourage proactive engagement altogether [70]. We investigate whether users recognize an encounter with misinformation in podcasts and how often they believe it occurs. This gives us a view of user perception related to podcast information, which can have implications for the design of an intervention. This exploration can also provide insights into whether users are encountering challenges like echo chambers [7]. This section also delves into the extent to which listeners believe they can accurately identify misinformation, and whether they have confidence in their ability to critically evaluate content. Understanding these dynamics helps us design interventions, described in Chapter 4, that either raise awareness of potential misinformation or simplify the process for contributing feedback, depending on the trust levels observed.

3.0.3. Reaction to Misinformation in Podcast

The Reaction to Misinformation section focuses on how participants believe they would respond when encountering misinformation in podcasts. This part of the survey explores whether listeners are likely to actively adjust their behavior upon recognizing misleading or false information. It investigates specific behavioral shifts, such as a willingness to fact-check the information, avoid the current episode, stop following the podcast series altogether, or even reconsider listening to podcasts in general. Participants are also asked about the sources they typically use to verify information, highlighting their trust in different verification methods. Finally, this section proposed potential intervention forms, PAUSE, ALERT, and VOLUNTARY, which aimed to identify the approach participants considered most acceptable and formed the basis for the interventions described in Chapter 4.

3.0.4. Incentives to Flag Misinformation in Podcasts

The final section of the survey, Incentives to Flag Misinformation, explores factors that could motivate listeners to actively engage in flagging misinformation. Both intrinsic incentives, such as altruistic reasons and a sense of responsibility, and extrinsic incentives, including monetary rewards and access to exclusive content are included. Monetary incentives, in particular, have been shown to enhance motivation and improve accuracy in identifying misinformation by encouraging users to invest more time and effort into content evaluation [57]. Panizza et al. [51] suggests that these incentives effectively shift behavior from passive consumption to active engagement, such as fact-checking or reporting misinformation. By linking rewards directly to accurate outcomes, monetary incentives provide a tangible reason for users to critically evaluate content, making them a powerful tool in combating misinformation. However, it is equally important to explore whether non-monetary incentives, such as altruistic motivations, access to ad-free content, or community recognition, could similarly drive user engagement and action against misinformation. Understanding the relative effectiveness of these approaches could inform the design of a more sustainable intervention to nudge listeners towards more active roles in reporting misinformation.

3.1. Methodology

The survey data was collected through a questionnaire hosted on the Qualtrics platform[1]. Responses were sourced from Prolific.[2] Prolific was chosen for this study because it provides access to a diverse and reliable participant pool, which ensures a broad representation. Prolific was chosen for this study because it provides access to a diverse and reliable participant pool, which ensures a broad representation. The platform is specifically designed for academic research and offers features that facilitate ethical participant recruitment [50]. Prolific's transparency regarding compensation and participant rights aligns with the TU Delft ethical considerations. This also reduces issues such as response bias and dropout rates. These attributes motivated our choice to use Prolific for conducting this study.

The survey begins with a consent form that provides participants with the necessary information about the study, including its purpose, an overview of question types, the estimated completion time, and the steps taken to maintain confidentiality. The full consent form can be found in Appendix A. As required by ethical guidelines, participants must agree to the terms specified in the consent form before participating in the survey. The ethical considerations that were taken into account follow the TU Delft ethics procedure [3]. This process ensured informed consent, data anonymization, and secure storage to uphold participant privacy and ethical standards throughout the study. This study was formally approved by the HREC committee at TU Delft. As for the style of questions, the survey included primarily multiple-choice and Likert scale questions, chosen for ease of response and to ensure structured data collection. The full questionnaire is in Appendix B.

3.1.1. Cognitive Bias Checklist

Crowdsourcing and using platforms like Prolific have their limitations, particularly in ensuring data quality due to the inherent variability in worker motivation and cognitive biases. Cognitive biases are a significant yet often overlooked factor influencing the quality of crowdsourced data. These biases can negatively

[1] https:www.qualtrics.com

[2] https:www.prolific.com

[3] https:filelist.tudelft.nl/TUDelft/Over_TU_Delft/Strategie/Integriteitsbeleid/Research%20ethics/HREC-Articles_of_Association.pdf

impact data reliability by skewing crowd workers' judgments in tasks that require subjective answers [20].

Draws et al. [20] has proposed the use of a structured checklist to be able to systematically assess potential cognitive biases. While ideally applied before data collection, applying the checklist post-hoc still allows us to identify biases that might have influenced the data collected, allowing us to document and contextualize results. Furthermore, a post-hoc approach provides insights into how different biases may have affected crowd workers' responses. These findings can inform future task designs, ensuring that mitigation strategies are taken into account.

The cognitive bias analysis of the survey revealed several key biases that could potentially influence participant responses which can be found below in 4.1. Self-interest bias was identified as a concern, as participants motivated by monetary incentives might rush through the survey. This was already addressed as a concern before the checklist,To address this, attention-check questions were incorporated to mitigate the risk and ensure data quality. Salience and availability biases were flagged due to the inclusion of specific podcast names in attention checks, which may have disproportionately influenced responses; for future note this can be better addressed by utilizing more generic attention checks, that are still in theme but ones that do will not have influence. Confirmation bias and overconfidence bias were also noted, with participants potentially overestimating their ability to recognize misinformation or being influenced by discussions of misinformation, which was further explored in Chapter 4. Lastly, anchoring effects were identified in scaled questions, and sunk cost fallacy was recognized as a potential issue, with mitigation steps taken for participants to quit at any time, however, it could have been better clarified that participants would be rewarded accordingly to the time spent.

The cognitive bias analysis of the survey revealed several key biases that have potentially influenced participant responses, as outlined in Table 4.1. Self-interest bias was identified as a concern, as participants motivated by monetary incentives might rush through the survey. This issue had already been anticipated, and attention-check questions were incorporated to mitigate the risk and ensure data quality. Salience and availability biases were flagged due to the inclusion of specific podcast names in attention-check questions, which may have disproportionately influenced responses. Specific podcast names could draw attention to participants' personal experiences, shaping their perceptions or decisions. For the future, this can be addressed by utilizing more generic attention checks. Confirmation bias and overconfidence bias were also noted. Participants might overestimate their ability to recognize misinformation or be influenced by discussions of misinformation, an issue that is further explored in Chapter 4. Lastly, the sunk cost fallacy was identified as a potential issue, where participants may have felt compelled to complete the survey to ensure compensation. While steps were taken to allow participants to quit at any time, it could have been better clarified that they would be rewarded proportionally for the time spent.

These findings highlight areas for refinement in survey design to minimize cognitive biases and enhance the validity of future studies.

Cognitive Bias	**Identified? (Yes/No)**	**Impact**
Self-interest Bias	Yes	The survey compensates participants only upon completion, which could incentivize rushed or inattentive responses to complete the survey quickly.
Affect Heuristic	No	N/A
Groupthink or Bandwagon Effect	No	N/A
Salience Bias	Yes	Prominent examples, like specific podcast names in the attention check question (Q25), could disproportionately affect responses.
Confirmation Bias	Yes	Talking about misinformation may induce bias that leads participants to believe that they have come across misinformation more.
Availability Bias	Yes	Prominent examples, like specific podcast names in the attention check question (Q25), could disproportionately affect responses.
Anchoring Effect	No	N/A

Halo Effect	No	N/A
Sunk Cost Fallacy	Yes	Participants might feel compelled to complete the survey once started to ensure compensation, even if they lose interest.
Overconfidence or Optimism Bias	Yes	Questions on confidence in recognizing misinformation (e.g., Q9) may elicit overconfident responses.
Disaster Neglect	No	N/A
Loss Aversion	No	N/A

Table 3.1: Results of applying the Cognitive Bias Checklist to Study I post-hoc.

3.1.2. Participant Recruitment

In total, 110 participants were recruited and each was compensated at an average rate of £9 an hour. All participants passed the attention check successfully and consented to participate in the study. The target audience was selected according to the following criteria to define the scope of the study. Only English-speaking participants were included to maintain consistency in understanding and interpreting the survey questions. Participants were required to be at least 18 years old. Geographically, the study focused on individuals from the United States and the United Kingdom. This focus was important for the transferability of findings, especially given the second research question, which involves experiments using podcast segments in English. While education level was not used as an exclusion criterion, the survey collected information on participants' educational backgrounds. This approach allowed for the analysis of potential educational influences on study outcomes.

3.1.3. Pilot Study

A preliminary pilot study was conducted in two phases, each with 5 participants, to ensure the survey's functionality and to gauge the time necessary for participants to complete the task and get fairly compensated. In the first phase, the expected completion time was set at 10 minutes; however, participants completed the survey on average significantly faster, in only 4.30 minutes. In the second phase, the expected time to complete the survey was updated to give participants 6 minutes. Additionally, an open question was added at the end of the survey. It allowed participants to share their thoughts, concerns, or suggestions about the survey, including question clarity, survey flow, or technical issues. All participants in both phases were compensated at an average rate of £9 per hour.

3.2. Study Results

3.2.1. Demographics and Listening Habits

The survey participants represented a diverse demographic profile in terms of age, gender, and education level. Ages ranged from 21 to 66 years, with an average age of approximately 36 years, and majority (74.5%) of participants were aged 20–40 years. In terms of gender, 57.7% identified as female, 40.5% as male, and 1.8% as non-binary. Education levels varied, with nearly half of the participants (43%) holding a bachelor's degree, 20% having completed a master's degree, and around 30% reporting a high school diploma or equivalent. A smaller subset (3%) had earned a doctorate.

Most survey participants can be categorized as regular casual podcast listeners, individuals who engage with podcasts frequently but for shorter durations. Specifically, 66.4% of participants reported a habitual engagement with podcast content, listening to podcasts more than once a week. However, despite this frequent listening, over 70.9% of participants indicated that their total listening time was less than three hours per week, suggesting that their podcast sessions are typically concise. To further understand these patterns, an analysis was conducted to explore potential correlations between participants' listening behavior and their demographic characteristics, such as age, gender, and education level. This exploration aimed to uncover whether demographic factors influenced listening frequency and provide insights into whether certain demographic groups engage differently with podcasts. Relationships between demographic factors and listening habits were analyzed with a Pearson's correlation test [60]. Pearson's correlation measures the linear relationship between continuous or ordinal variables, providing the strength and direction of any potential associations [60]. The results of this analysis, shown in Table 3.2, were based on encoding listening frequency and listening hours per day as ordinal variables, with values assigned incrementally from least to most frequent.

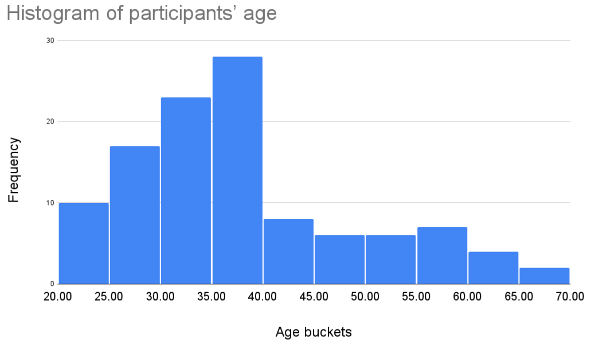


Figure 3.1: Histogram of participants' age

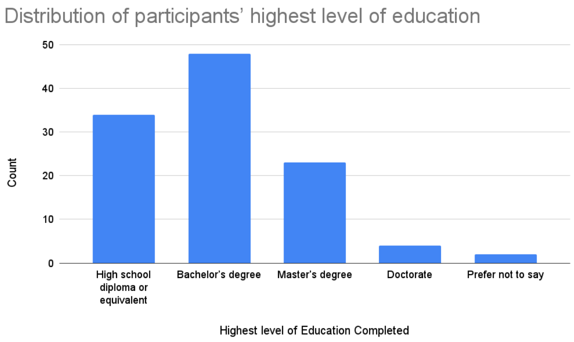


Figure 3.2: Distribution of participants' education level

Variable	**Age (r, p-value)**	**Gender (r, p-value)**	**Education (r, p-value)**
Frequency	$0.166, 0.208$	$-0.107, 0.358$	$-0.046, 0.315$
Hourly Frequency	$0.121, 0.208$	$-0.089, 0.358$	$-0.097, 0.315$

Table 3.2: Pearson's Correlations for Listening Frequency and Hourly Listening Frequency with Demographic variables.

These findings indicated weak and non-significant associations. Although a weak positive correlation was observed between age and listening frequency ($p = 0.082$ and $p = 0.208$) and a directional trend for males listening more frequently than other gender groups ($p = 0.267$ and $p = 0.358$), these results did not reach statistical significance. Similarly, education levels showed negligible and non-significant correlations with listening habits ($p = 0.636$ and $p = 0.636$). Given the non-significant p-values, these findings do not provide sufficient evidence to conclude that demographic factors influence podcast listening behavior.

Furthermore, results showed that Entertainment, Education, and News are the most popular podcast genres, with a few participants mentioning more niche genres such as Spirituality and True Crime. To explore whether demographic factors, such as age, gender, and education level, influenced participants' genre preferences, we conducted a Chi-Square Test of Independence [44]. This test was chosen because it is appropriate for analyzing associations between categorical variables, allowing us to determine whether there is a significant relationship between demographics and preferred podcast genres [44]. The results of the test revealed a p-value of 0.995, suggesting that we do not reject the null hypothesis. This indicates no significant association between demographics and preferred podcast genres.

In addition to genre preferences, we explored how participants discover and engage with podcasts. We find that the primary ways people discover new podcasts are through Social media, Recommendations from Friends and Family, or Platform Suggestions. Moreover, a substantial portion (70.9%) of participants listen to podcasts while doing housework or chores, with other common activities being traveling, walking, or exercising. Only under 18.2% reported just listening. One participant mentioned in response to the open feedback question that their response to misinformation in podcasts is influenced by their current activity, such as driving, cooking, or walking, which limits their ability to engage actively. They noted they might make a mental note of the misinformation but often continue listening without taking further action. This context highlights two key considerations for the remaining section: first, participants may not fully immerse or focus on the content while listening, and second, their multitasking context shapes their ability to interact with or engage meaningfully with the podcast player interface.

3.2.2. Trust and Confidence

The majority of participants expressed trust in the information presented in podcasts, with a higher level of trust for podcasts they regularly listen to. Specifically, trust in podcasts overall was reported by 55% of respondents, increasing to 78% for familiar podcasts they listen to regularly. Confidence in recognizing misinformation was notable, with 70% of participants expressing confidence in their ability to identify misinformation.

Arin, Mazrekaj, and Thum [5] observed that factors such as age, income, and political orientation influence misinformation detection abilities. Older, higher-income, and left-leaning individuals tend to

Response	Trust in Podcast (%)	Trust in Podcast I Listen To (%)	Confidence (%)
Strongly Disagree	0.0%	0.0%	0.0%
Disagree	6.4%	2.7%	9.1%
Neutral	38.2%	19.1%	20.9%
Agree	54.5%	68.2%	56.4%
Strongly Agree	0.9%	10.0%	13.6%

Table 3.3: Percentage distribution of responses for trust in podcasts, trust in podcasts participants listen to, and confidence in identifying misinformation.

demonstrate greater proficiency, while younger and right-leaning individuals often face challenges. In contrast, our findings reveal that trust in podcasts and confidence in identifying misinformation were not strongly associated with demographic factors overall. Table 3.4 indicates weak and non-significant correlations between general trust in podcasts and age, gender, or education level. Similarly, trust in the podcasts participants listened to showed no significant associations with these demographic factors. Confidence in identifying misinformation was also not significantly influenced by age or gender. However, education level demonstrated a statistically significant, albeit relatively weak, correlation ($r = 0.193, p = 0.044$), suggesting that participants with higher education tend to feel more confident in their ability to identify misinformation. It is important to note that Arin, Mazrekaj, and Thum [5] highlight how increased confidence may not always translate to accuracy, as overconfidence can be a critical issue. This tendency toward overconfidence can lead individuals, including those with higher education, to unintentionally spread misinformation.

Variable	Age (r, p-value)	Gender (r, p-value)	Education (r, p-value)
General Trust in Podcasts	$0.071, 0.463$	$0.041, 0.668$	$-0.051, 0.597$
Trust in Personal Podcast Choices	$0.054, 0.573$	$-0.064, 0.509$	$-0.085, 0.380$
Confidence in Identifying Misinformation	$-0.108, 0.259$	$-0.064, 0.509$	$0.193, 0.044$

Table 3.4: Pearson's Correlations for Trust and Confidence Variables with Demographics.

The analysis of trust and confidence in relation to podcast listening frequency revealed varying degrees of correlation. As shown in Table 3.5, trust in podcasts showed a weak positive correlation with both listening frequency and hourly frequency, though neither is statistically significant. In contrast, trust in the podcasts that participants personally listen to demonstrated a statistically significant positive correlation with listening frequency ($r = 0.271, p = 0.004$). This suggests that individuals who listen to podcasts more frequently are more likely to trust their selected podcast choices. These findings emphasize the role of personal podcast selection in shaping trust, while confidence in identifying misinformation appears to be less influenced by how often or what type of content participants engage with.

These findings highlight the role of personal podcast selection in shaping trust, which could lead to two contrasting interpretations. On one hand, users may gravitate toward echo chambers, potentially suffering from the illusory truth effect, where repeated exposure to similar content reinforces biases and misinformation. On the other hand, frequent listeners who can identify misinformation might carefully curate their podcast selections to minimize exposure to unreliable content. However, the reality likely lies somewhere in between: people tend to listen to podcasts they enjoy and inherently trust, occasionally identifying misinformation but likely missing it at other times. This highlights the potential for further research to further understand how trust and podcast selection interact and whether these patterns enlarge or mitigate the risks of misinformation.

Lastly, while 25% of participants believe they never encounter misinformation in podcasts, 38% think they do, with no statistically significant correlation to demographic factors. Two-thirds (66%) trust the sources from which they discover podcasts, mostly through social media, friends, and family. The main factors influencing trust in podcasts are host and source credibility, with some participants also considering guest reliability and the host's expertise.

Variable	Frequency (r, p-value)	Hourly Frequency (r, p-value)	Genre (r, p-value)	Reaction to Misinformation (r, p-value)
Trust in podcast	$0.171, 0.073$	$0.151, 0.115$	$0.063, 0.341$	$-0.041, 0.608$
Trust in podcast I listen to	$0.271, 0.004$	$0.123, 0.201$	$-0.002, 0.978$	$0.002, 0.977$
Confidence	$0.164, 0.086$	$0.034, 0.728$	$-0.005, 0.937$	$-0.054, 0.499$

Table 3.5: Pearson's Correlations between Listening Habits, Trust/Confidence Variables, Genre Preference, and Misinformation Reaction.

3.2.3. Reaction to Misinformation in Podcasts

From Table 3.6, we can see that 24.6% reduce their listening to an episode when they encounter misinformation, and 20.0% of participants stop listening. A small number indicated they do not care about the misinformation. Similarly, 19.1% stopped listening to the entire podcast series after encountering misinformation, while 28.2% reduced their listening frequency. More than half (55.5%) of the participants keep misinformation in mind when listening to other podcasts and over 21.8% do not care. In all situations, over 36% of listeners continue to listen but remain mindful of the misinformation.

Response	Current Episode (%)	Podcast Series (%)	Other Podcast Series (%)
I do not care	1.82%	2.73%	21.82%
I keep it in mind but continue listening	43.64%	36.36%	55.45%
I somewhat reduce listening	10.00%	13.64%	7.27%
I reduce listening	24.55%	28.18%	10.91%
I stop listening	20.00%	19.09%	4.55%

Table 3.6: Participant responses to encountering misinformation across different podcast contexts.

Participants displayed varying reactions to encountering misinformation in podcasts, with a discrepancy between those who continued listening and those who reduced or stopped their listening. To understand the factors driving these behaviors, we first explored the relationship between demographic factors and participants' reactions. A Chi-Square test revealed no significant associations between demographic factors and reactions to encountering misinformation across podcast episodes ($p = 0.500$), entire series ($p = 0.082$), and other series ($p = 0.772$). These findings indicate that demographic factors alone are not strong predictors of participants' responses to misinformation, suggesting that other influences, such as genre preferences or listening intent, may play a bigger role. Genre preference could impact how misinformation is perceived because certain podcast genres, such as news or educational content, are typically consumed with an expectation of accuracy and factual integrity. When misinformation is encountered in these genres, listeners may feel more betrayed or disillusioned, prompting stronger reactions such as reducing or stopping their engagement. Conversely, in genres like entertainment or storytelling, where factual accuracy may not be the primary focus, listeners might be more forgiving or indifferent to misinformation, leading them to continue listening despite recognizing its presence. Similarly, listening intent can shape reactions to misinformation. Listeners who tune in for relaxation or entertainment may prioritize the enjoyment of the content over its factual accuracy, making them less likely to disengage upon encountering misinformation. On the other hand, those who listen for staying informed or educational purposes may be more critical, leading to a stronger reaction.

In both cases, we conducted a Chi-Square test to examine the association between genre and reaction. While the results showed a significant association in both cases (see Appendix C). A standardized adjusted Pearson residual analysis [44] showed this was between the ”Other” option and the ”I do not care” cell in both tests. Each having a extremely small number of participants in these categories, making the findings unreliable and limiting the generalizability of these results. This highlights the need for caution when interpreting associations driven by such small sample sizes.

further revealed notable differences among participants who selected ”Other” and those who indicated that they did not care. However, the number of participants in these categories was extremely small, making the findings unreliable and limiting the generalizability of these results. This highlights the need for caution when interpreting associations driven by such small sample sizes.

However, the number of participants in these categories was extremely small, making the findings unreliable and limiting the generalizability of these results. This highlights the need for caution when interpreting associations driven by such small sample sizes.

In both cases, we conducted a Chi-Square test, while the Chi-Square test showed a significant association between genre and reaction in a podcast episode. A standardized adjusted Pearson residual analysis [44] revealed notable differences among participants the association was between the ”Other” genre and ”I do not care” cell (see Appendix C). However, this result is unreliable due to the small sample size (1 participant) who selected ”Other,” making it difficult to draw meaningful conclusions from this specific finding. A similar scenario was observed for listening intent, where the association was between the ”Other” listening intent and the ”I do not care” cell. This result was also unreliable due to a small sample size (1 participant). For reactions to entire podcast series or other podcast series, no significant associations were found, indicating that genre and listening intent had no observable influence in these cases. This finding suggests that the topic or listening intent may not significantly influence how listeners evaluate and respond to misinformation. The lack of significant associations for entire podcast series and other podcast series indicates that listeners might approach misinformation in a relatively consistent manner, regardless of their specific listening intent or the genre of the content. This suggests that listeners may care about addressing misinformation regardless.

This concern is reflected in participants’ reactions to misinformation, with 72% of participants claiming to research it further, while 36% ignore it, and less than 1% report it. This indicates a notable discrepancy: while the majority are willing to invest effort in researching misinformation, very few take the step to report it. This suggests potential barriers to reporting, such as a lack of trust in reporting systems, the perceived effort involved, or unfamiliarity with the process. One participant expressed that even if they can identify misleading information, they believe experts might be better placed to explain why, which acts as a barrier to personally flagging misinformation. These findings highlight an area with potential for further exploration and intervention.

Most participants (77%) verify misinformation through online searches, 36% consult academic journals or research papers, and only 13% do not typically verify misinformation at all. Building on these findings, we examined whether participants’ verification sources varied based on their education level. A Chi-Square test revealed no significant association between education level and the choice of verification sources ($\chi^2 = 32.187, df = 28, p = 0.267$). These findings suggest that education level does not play a significant role in shaping verification behavior. However, while the Chi-Square test did not reveal a statistically significant association, the data provides some directional insights into participants’ verification behaviors based on education level. Participants with higher education levels, such as those holding master’s degrees or doctorates, appeared more inclined to consult academic journals or scientific publications compared to those with lower education levels. For instance, 12 out of 49 participants with master’s degrees and 3 out of 10 with doctorates reported using academic journals, while only 18 out of 96 bachelor’s degree holders and 7 out of 66 high school diploma holders did so. This is likely due to the fact that master students and doctorates are more likely to have been exposed to a broader space of academic digital libraries and trained in their use, giving them greater familiarity and access to these resources. In contrast, participants across all education levels heavily relied on online searches, with 77% of the total sample citing this as their primary verification source. These directional trends suggest that while education level may subtly influence the choice of certain verification sources, the overall reliance on accessible methods like online searches remains consistent across all groups. This highlights an opportunity for platforms, such as requiring podcasts to provide links to reliable resources or fact-checking articles, making it easier for users to verify claims directly within the platform.

3.2.4. Incentives to Flag Misinformation in Podcasts

The survey revealed that 18% of participants find it acceptable to be stopped with a pause intervention when encountering misinformation, 43% are comfortable with being alerted, and 73% prefer voluntary interventions. To understand these preferences further, we explored whether participants’ trust in podcasts, both general trust and trust in podcasts they listen to regularly, along with their confidence in identifying misinformation, correlated with the incentives they found acceptable. However, the Chi-Square tests did not reveal statistically significant associations between these variables. There are directional trends suggesting that “Sense of Responsibility” and “Monetary Rewards” stood out as favored incentives among participants across varying levels of trust and confidence.

Furthermore, participants were primarily motivated by a sense of responsibility (65%), monetary rewards (62%), and altruistic reasons (55%). For non-monetary incentives, 88% of those who preferred them opted for ad-free content, while 79% preferred free or discounted streaming subscriptions. Additionally, 55% of participants answered that they would be very likely to participate in identifying misinformation in podcasts for monetary rewards, whereas 60% were only somewhat likely to participate given non-monetary rewards. Participants were also somewhat motivated by knowing that others would participate or that their efforts would help others.

An effective rewards system could offer points, ad-free content, or subscription discounts for engaging in the flagging process, addressing both monetary and non-monetary preferences. Social and community engagement also play a role, many users are motivated by knowing others are involved or that their actions contribute to a collaborative effort. To leverage this, the intervention could include community-based features, such as showing the number of users who flagged specific content or displaying trust scores for flagged segments. This could stimulate user participation in flagging misinformation by leveraging user motivation.

3.2.5. Limitations

This study has two limitations that should be considered when interpreting the results. First, the reliance on self-reported data introduces potential biases, such as overconfidence, availability bias, and salience bias, particularly when participants were asked about their verification behaviors or motivations. Second, while the use of multiple-choice questions streamlined data analysis, it may have limited the ability to articulate more nuanced perspectives, potentially reinforcing confirmation bias.

3.3. Conclusion

Study I explored how users respond to misinformation in podcasts, focusing on their trust in podcast content, confidence in identifying misinformation, and behavioral reactions. The findings revealed diverse user attitudes and behaviors toward misinformation, offering valuable insights into the factors that shape listener engagement. The results revealed that the majority of participants expressed trust in podcasts, especially those they regularly consume. Confidence in detecting misinformation was influenced (weakly) by demographic factors such as education level. Reactions to encountering misinformation varied widely. Many participants continued listening while remaining mindful of the misinformation, though a minority reduced or stopped their engagement with the podcast or series. Listening intent emerged as a key factor. Despite participants' willingness to research misinformation, only a small fraction reported taking active steps to flag it. This discrepancy highlights potential barriers, such as a lack of trust in reporting systems or unfamiliarity with the process. Finally, participants expressed a strong preference for voluntary over intrusive interventions, signaling the importance of user-focused design in addressing misinformation. The findings provide a foundational understanding of user attitudes and behaviors regarding misinformation in podcasts. They explore the need for interventions that balance user autonomy with the goal of encouraging critical engagement. These insights informed the design of Study II.

4

Study II: How can we optimize user engagement to flag misinformation?

According to Atkins, Wanick, and Wills [6], user engagement is defined as the frequency and duration of user interactions with an application. On social media platforms, engagement manifests in various forms depending on content format and platform design. For example, richer media formats like videos often encourage active engagement through comments, while simpler formats like photos drive passive interactions such as likes [10]. Similarly, personalized feeds in eHealth platforms [34] and social media demonstrate how tailoring content to individual preferences can foster both passive engagement, such as clicking on feed items, and active participation, like contributing posts or engaging in discussions. Moreover, platform-specific dynamics influence engagement types. Optimizing user engagement is essential because it improves the relevance and impact of interactions, creating connections and encouraging participation. For instance, Instagram's visually driven design creates emotionally engaging content, while Twitter's emphasis on real-time updates promotes fact-based interactions [23].

Furthermore, literature suggests that engaged users are more likely to flag inaccuracies, trust the platform, and advocate for its value [21]. Encouraging this transformation from passive consumers to active participants is particularly important for audio-based platforms like podcasts, where traditional forms of engagement, such as likes or comments, are less applicable. Designing interventions that empower users to critically evaluate content and flag misinformation requires an understanding of how engagement can be elicited and sustained without disrupting the listening experience.

This challenge leads to our central research question: **How can we optimize user engagement to flag misinformation?** By exploring engagement strategies, including feedback mechanisms and intervention designs, our study tests three different intervention approaches that encourage users to actively contribute to misinformation detection while maintaining a frustration free podcast experience. The findings will provide valuable insights into creating more effective and user-centered solutions for combating misinformation in podcast platforms.

Effective feedback can be elicited through push and pull strategies [40]. Pull feedback involves users proactively identifying and flagging misinformation when they perceive it, promoting autonomy and voluntary participation. Conversely, push feedback involves prompting users to respond at specific moments, such as when the system detects potentially misleading content. Evidence suggests consistent and immediate feedback is more effective than delayed responses in capturing user attention and prompting action [45]. Additionally, a simplified feedback process and well-timed requests are critical for increasing response rates and engagement without disrupting the primary experience like listening to a podcast.

Building on these principles of user engagement and effective feedback mechanisms, to answer our research question, this study explores three interventions designed to test different feedback mechanisms to promote user engagement. The design of the three interventions: PAUSE, ALERT, and VOLUNTARY test a range of feedback mechanisms and levels of disruption.

The PAUSE intervention incorporates immediate feedback by actively interrupting the user's listening experience to prompt a response. This design is rooted in the evidence that real-time feedback sustains attention and reinforces engagement by requiring users to evaluate the content at specific moments. By forcing users to interact, the PAUSE intervention mirrors the success of performance feedback strategies, such as those used in the CoastMaster intervention [67], where immediate prompts led to measurable behavioral changes. However, as this intervention style is the most intrusive, it is expected to cause the greatest feedback fatigue [40], particularly as it disrupts the listening experience once per podcast. Striking the right balance for an acceptable frequency of interruptions is challenging, and this intervention is anticipated to be perceived as more frustrating compared to the ALERT and VOLUNTARY interventions.

The ALERT intervention utilizes auditory cues as indirect feedback to draw user attention to potential misinformation. It draws inspiration from examples like Facebook's use of visual cues to highlight content [37], adapting the concept to the podcast context to draw user attention to potential misinformation. By incorporating auditory cues, the intervention aims to guide user behavior more subtly, minimizing disruptions to the listening experience. This hybrid method is designed to balance pushing the user to provide feedback while maintaining the flow of the podcast.

The VOLUNTARY intervention, in contrast, relies on passive feedback, allowing users to flag misinformation at their discretion. This intervention maintains a user's autonomy to act when they feel the need to flag misinformation. While less intrusive, it will serve as a baseline to compare how users engage naturally without external prompts.

4.1. Hypotheses

Studies on user engagement emphasize the impact of feedback strategies on user interaction and accuracy. Subtle nudges, such as auditory cues, have been suggested to encourage user action while preserving their sense of autonomy. This approach aligns with the rationale for the ALERT intervention, which aims to enhance engagement through such cues [37]. Similarly, real-time feedback mechanisms, like immediate prompts in the PAUSE intervention, compel users to focus and respond actively, potentially driving higher engagement through forced interaction [67]. Cues such as prompts and alerts can encouraging users to activate their System 2 thinking and critically evaluate content [56].

In contrast, VOLUNTARY engagement relies on intrinsic motivation and user initiative, which can influence participation. Factors such as curiosity, enjoyment, skill development, and personal growth are critical drivers of intrinsic motivation [36]. While a user may feel intrinsically motivated, the task has not been explicitly designed to stimulate such motivation—assuming it is even possible to do so effectively. Furthermore, Delaney and Royal [19] highlights that intrinsic motivation tends to diminish when extrinsic motivation increases, emphasizing the delicate balance between the two. In contrast, VOLUNTARY engagement relies on intrinsic motivation and user initiative, which can significantly influence participation. Factors such as curiosity, enjoyment, skill development, and personal growth are critical drivers of intrinsic motivation [36]. However, the task has not been explicitly designed to stimulate such motivation—assuming this is even possible to do so effectively. Furthermore, Delaney and Royal [19] highlights that intrinsic motivation tends to diminish as extrinsic motivation increases.

Studies on user engagement emphasize the critical role of feedback strategies in shaping user interaction and accuracy [37]. Real-time feedback mechanisms, such as immediate prompts in the PAUSE intervention, create a forced interaction which should driver higher engagement [67]. Furthermore, subtle nudges, such as auditory cues can trigger System 2 thinking, enabling users to critically evaluate potential misinformation [56], while still preserving a users sense of autonomy [37]. These points lay the foundation for the following hypotheses regarding user behavior and responses to different interventions. In contrast to externally driven engagement, VOLUNTARY engagement depends on intrinsic motivation and user initiative. Intrinsic motivators—such as curiosity, enjoyment, skill development, and personal growth—play a significant role in participation [36]. However, tasks designed without explicit stimulation of intrinsic motivation may struggle to elicit active engagement. Furthermore, Delaney and Royal [19] shows that intrinsic motivation tends to diminish when extrinsic motivators, such as monetary incentives or forced tasks, are introduced. These contrasting dynamics between intrinsic and extrinsic motivation inform the design and expected outcomes of the VOLUNTARY, PAUSE, and ALERT interventions.

- **H1:** The **combination of a prompt and an auditory signal** will result in higher engagement than prompts without an auditory signal.

- **H2:** The presence of an **auditory signal** will increase response rates compared to silent prompts, as it directs user attention toward potential misinformation more effectively.

- **H3:** The **PAUSE intervention** will have a higher cognitive load and be perceived as the most frustrating due to the forced interaction. Conversely, the **VOLUNTARY intervention** will have the lowest cognitive load.

- **H4:** Accuracy in the auditory-based interventions (PAUSE and ALERT), regardless of the presence of a prompt, will be higher than in the VOLUNTARY intervention.

To test these hypotheses, an experiment was designed to measure differences in engagement and cognitive load across three interventions: PAUSE, ALERT, and VOLUNTARY. The purpose of the experiment was to create a controlled environment that simulated a real-world podcast listening experience, exploring how each intervention type affected user responses in a task where users were asked to flag misinformation. The simulated scenario focused on participants who were solely listening to the podcast, without multitasking. From Study I, we observed that over 70% of participants reported listening to podcasts while multitasking. Therefore, the results of this study are not directly generalizable to multitasking scenarios and are more relevant to focused listening contexts. Participants were randomly assigned to one of the three interventions and observed as they interacted with six short labeled podcast segments. Each segment lasted up to two minutes, covered a distinct topic, and was categorized by one of six truth labels, with each segment containing a sentence explicitly labeled with its truth value. Following the experiment, participants completed a post-task survey assessing System Usability Scale (SUS), User Engagement Scale (UES), cognitive load, familiarity, trust, and other related factors. This approach aimed to evaluate the intervention's impact on user behavior and the broader factors influencing effectiveness.

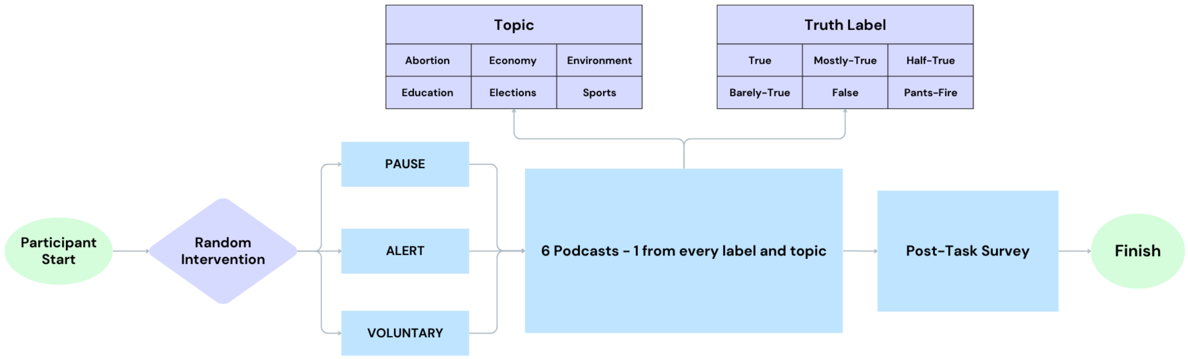


Figure 4.1: Experiment Flow

4.2. Experimental Setup

4.2.1. Independent Variables

We examine the differences in evaluation metrics across the baseline **VOLUNTARY** intervention, the **ALERT** intervention, and the **PAUSE** intervention.

1. **Pause Feedback**: The listening experience is paused, and users are prompted to respond, indicating whether they believe the last segment contained misinformation.
2. **Alert Feedback**: Users receive an alert encouraging them to provide feedback on potential misinformation, with the option to respond or ignore.
3. **Voluntary Feedback**: Users can voluntarily submit feedback whenever they suspect misinformation in the content.

Furthermore, participants are tasked with listening to **six randomly picked podcast** segments based on a filtered subset. This filtered subset orders the least selected podcasts to ensure a more balanced distribution across the dataset. The distribution can be found in Appendix E. Participants are encouraged to identify and flag any misinformation they encounter. Their performance is assessed based on the accuracy of the misinformation flagged.

4.2.2. Dependent Variables

We investigated the impact of user flagging through two key factors: **the number of user flags** and **the accuracy of those flags**. The number of user flags was determined by the frequency with which a user reports misinformation using the options "TRUE", "FALSE", or "I DON'T KNOW".

Furthermore, we gathered several post-task measures through a survey. Essential elements captured were system usability, user engagement, cognitive load, trust, user preferences, and participants' familiarity and interest across the different categories.

The System Usability Scale (SUS) is a widely recognized and validated tool which offers a quick yet effective measure of usability, capturing participants' perceptions of ease of use, learnability, and satisfaction [13]. The SUS measured whether our design supported an intuitive user experience and was suitably designed to empower users to handle misinformation.

Furthermore, to understand how participants interacted with the podcast content and interventions, **the User Engagement Scale (UES)** was measured. This section focused on aspects like System Usability, Perceived Usability, Novelty, Felt Involvement, and Endurability [48]. These dimensions allow us to gauge how engaging and immersive the experience was, which is critical in assessing to what extent the interventions disrupted participants' attention. Insights from these measures help identify areas where the interface design may support or limit engagement.

The NASA Task Load Index (NASA-TLX) was included to measure cognitive load, assessing the mental effort required from participants throughout the study. This tool evaluates multiple dimensions of workload, including mental and physical demand, effort, and frustration, providing insights into the cognitive demands imposed by the task and intervention [28]. By examining these factors, the study aims to understand whether cognitive load affects user engagement. A balanced cognitive load is hypothesized to support genuine engagement with the podcast content, resulting in overall higher engagement.

Additionally, we explored questions that were designed to assess participants' perceptions of and engagement with the flagging intervention. They explore whether the intervention prompted critical thinking and assisted in identifying misinformation, as well as its effects on participants' trust and engagement with the podcast. Lastly, questions address the intervention's impact on the flow of the listening experience and whether participants felt empowered or burdened by the task of flagging. By examining participants' willingness to engage and their perceived responsibility, these questions provide insights into how the intervention influenced user behavior and interaction with the content.

4.2.3. Podcast Player Interface

The podcast player interface was intentionally designed to resemble familiar podcast listening platforms like Spotify. This design choice aims to provide a user-friendly experience that participants can navigate intuitively, minimizing the learning curve and reducing cognitive load. By creating a familiar environment, the interface allows participants to focus on the primary task without unnecessary distractions or potential misuse. This approach helps maintain control over the interaction flow, ensuring that the study's results are not influenced by unfamiliarity with the player interface.

To ensure this, the player includes traditional icons for essential functions (see Figure G.1), such as play, pause, and rewind. The play and pause buttons, represented by a triangle and two vertical bars respectively, are intuitive. Additionally, a rewind button allows users to go back 15 seconds in the podcast, a common feature in podcast players that supports re-listening to specific sections, which can be valuable for listening to segments that may be dubious for listeners. The volume slider, another standard feature, allows participants to adjust audio levels to their preference.

An "Upcoming Episode" box is also displayed within the interface, mirroring elements commonly found in podcast players to add to the authentic experience. Although non-functional, this feature serves to create an immersive podcast listening environment without distracting from the primary task. Collectively, these design elements give participants a familiar environment that encourages listeners to immerse themselves in the task. The design decision was made to exclude a seek bar that would allow users to skip forward in the podcast. This restriction was necessary to prevent participants from skipping through the task. By limiting forward navigation, the study ensures that all participants engage with every segment of the task.

Similarly, the inclusion of a podcast cover image was initially considered. However, to avoid introducing potential biases linked to the content or source, it was decided to exclude a cover image. This choice helps ensure that participants focus solely on the audio content, free from visual elements that might inadvertently influence their judgment or engagement with the task.

The design choices reflect a careful balance between providing a familiar and functional user interface and maintaining control over the user experience to ensure the integrity of the study. The study aims to engage participants effectively without introducing unnecessary complexity. At the same time, restrictions such as the absence of a seek bar and the exclusion of the podcast cover image are deliberate choices to prevent bias and ensure that the study results are reliable and valid. At the beginning of the task, participants were presented with a task description that emphasized the study's goal (See

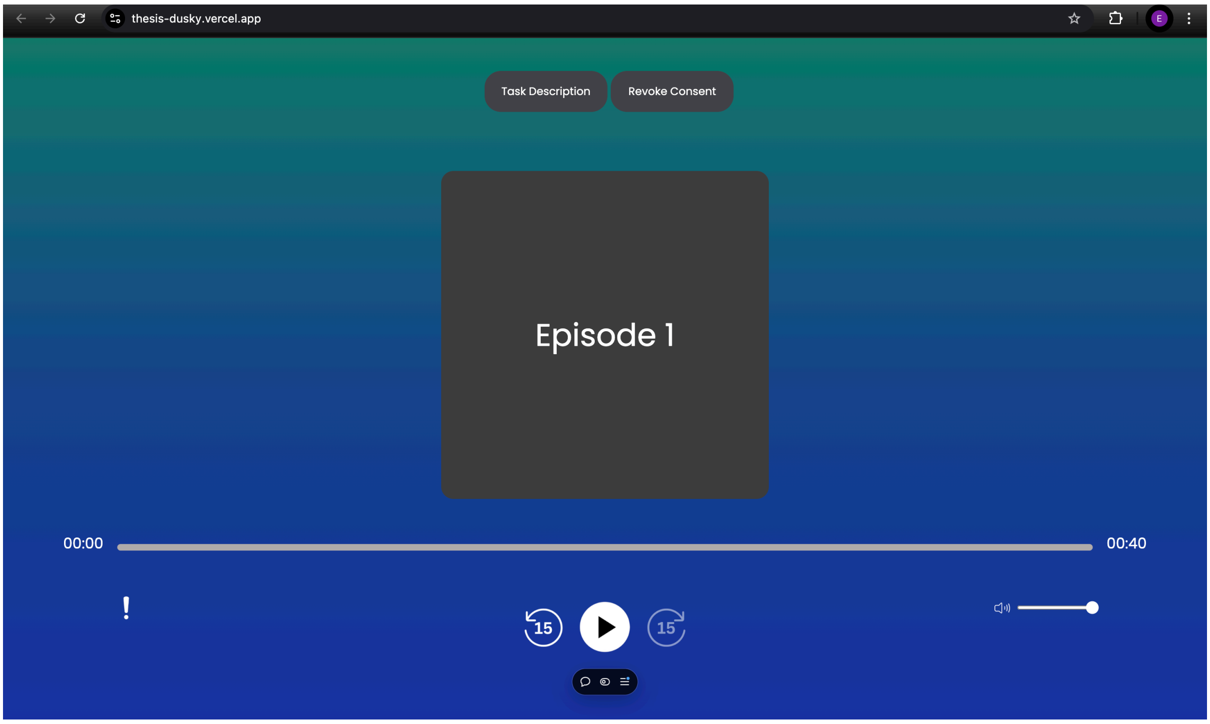


Figure 4.2: Podcast Player Web Application

Appendix G) and provided a detailed breakdown of each button in the player. Participants could access this task description at any time via a dedicated button located at the top of the interface. This feature allowed participants to conveniently refer back to the task guidelines without, helping them with the primary task if necessary. Additionally, a second button was incorporated to enhance user control and transparency, allowing participants to revoke their consent and exit the study at any time. This button is in line with good ethical practices, giving participants the freedom to withdraw and no longer participate. By including these options, the interface makes sure to allow for participant autonomy and reinforces the integrity of the study.

4.2.4. Cognitive Bias Checklist

We extend the approach outlined in Study I by applying the structured cognitive bias checklist proposed by Draws et al. [20] to this task. We apply the checklist post-hoc which gives us insights into the different potential biases that may have affected results. Several factors were identified that could have played a role in influencing results. Self-interest bias was potentially present, as participants may have prioritized completing the task quickly due to monetary incentives, potentially under-flagging misinformation. Similarly, there is a risk of confirmation and availability bias, with participants more likely to flag content aligning with their beliefs or focusing disproportionately on familiar topics. To measure potential effects of these biases, the post-task survey includes questions regarding topic familiarity and interest Further, there is a risk of Salience bias due to the inclusion of specific podcast names in the attention check questions, which could have disproportionately influenced participants' responses by triggering personal experiences with these podcasts. Additionally, participants may have experienced the Sunk Cost Fallacy, where participants might feel compelled to complete the task once started, even if they lose interest, potentially impacting the quality of the measured data. To address this, participants were given the option to revoke consent and exit the survey at any point. However, for future iterations, it would be beneficial to explicitly inform participants that they will be compensated for their time regardless of whether they complete the survey. This clarification could further reduce the influence of this bias and improve the reliability of the collected data. The breakdown of the checklist can be found below in Table 4.1.

Cognitive Bias	Identified? (Yes/No)	Explanation/Impact
Self-interest Bias	Yes	The survey compensates participants only upon completion, which could incentivize rushed or inattentive responses to complete the survey quickly.
Affect Heuristic	No	N/A
Groupthink or Bandwagon Effect	No	N/A
Salience Bias	Yes	Prominent examples, like specific podcast names in the attention check question (Q25), could disproportionately affect responses.
Confirmation Bias	Yes	Talking about misinformation may induce bias that leads participants to believe that they have come across misinformation more.
Availability Bias	Yes	Prominent examples, like specific podcast names in the attention check question (Q25), could disproportionately affect responses.
Anchoring Effect	No	N/A
Halo Effect	No	N/A
Sunk Cost Fallacy	Yes	Participants might feel compelled to complete the task once started to ensure compensation, even if they lose interest.
Overconfidence or Optimism Bias	No	N/A
Disaster Neglect	No	N/A
Loss Aversion	No	N/A

Table 4.1: Results of applying the Cognitive Bias Checklist to Study I post-hoc.

4.3. Labeled Podcast Dataset

In order to test our interventions, we required a podcast dataset with annotated truth labels for two key purposes: first, to determine whether a podcast segment contained misinformation, and second, to verify whether participants could accurately flag misinformation at the appropriate moments. However, despite the availability of different datasets with truth labels for misinformation detection, such as the LIAR dataset [72] FakeNewsNet dataset [63] and an extensive COVID-19 Fake News [52], there is a notable gap for datasets specifically designed for audio content like podcasts. Existing datasets primarily focus on textual data and do not cater to audio formats.

Textual datasets, while valuable for misinformation detection in written formats, do not capture several key dimensions of audio content. First, podcasts often convey information through spoken language, where tone, inflection, and emphasis can alter meaning or convey intent beyond what is transcribed. For example, sarcasm or subtle doubt expressed in a speaker's tone may not be obvious in a written transcript. Secondly, podcasts often rely on non-verbal elements such as pauses or vocal emphasis, which can influence the listener's perception of credibility and intent which are missing in textual datasets. This lack of an annotated podcast dataset created the need to develop one, which is a contribution of this study and can be used for future research.

The selection of the LIAR dataset [72] as the foundation for this work was motivated by its diverse range of topics and its labeling system, spanning a wide spectrum of truth values. Such diversity enabled the creation of podcast segments providing a basis for examining whether domain familiarity or interest influences user engagement and accuracy. While the LIAR dataset contains a variety of topics, a limitation is that they are based on US contexts, which makes them more likely to be familiar to US participants, limiting the generalizability of our podcast dataset.

To create a set of podcasts, six topics were selected from the LIAR set: "Abortion and Healthcare," "Economy and Taxes," "Energy and Environment," "Children and Education," "Elections and Legal Issues," and "Sports." From each topic, one statement was chosen to represent each of the six truth labels: "True," "Mostly True," "Half True," "Barely True," "False," and "Pants on Fire," combining to make a dataset of 36 statements. To transform these statements into engaging podcast segments, we prompted

ChatGPT-4 [49] to generate contextualized content around each statement. The prompts instructed ChatGPT-4 to place the labeled statement at varying points within the segment, either at the beginning, middle, or end, to include variation in the listening experience. This process required several iterations to fine-tune the segments. Initial attempts were overly formal and did not align with the intended conversational tone. Adjustments to the prompts specifying the style preference were made to achieve a more casual style and to ensure contextual content did not reveal or overly hint at the labeled statement. See Appendix D for a detailed example of the prompt structure.

The segments were then converted to audio format, using Google's Text-to-Speech (TTS) function [16]. Specifically, we used the 'en-US-Journey-O, Female' voice, which offered a natural-sounding output. The voice was chosen for its realistic delivery, including subtle pauses and breathing sounds, which replicated real-like speech. This method allowed the creation of a 36-segment audio dataset with truth labels embedded at specific timestamps.

The advantage of this approach is its feasibility; it enabled us to efficiently create a labeled dataset without the extensive time and resources required to label existing podcast content manually. A key limitation is that these synthetic segments may lack the natural flow and context of real podcasts.

4.4. Participant Recruitment

For the experiment, to determine the number of participants for our experiment, we conducted a G-power analysis [24]. For a medium effect size of 0.25 [17], we needed a sample size of 224 participants for the results to be statistically significant. This was split into 3 different batches, PAUSE, ALERT and VOLUNTARY.

The target audience was crowdsourced on Prolific. The participants were paid on average £9 an hour. As the podcast segments were derived from the LIAR dataset [72], which focuses on topics in the U.S., selecting U.S. participants ensured that the crowd possessed the necessary cultural and contextual knowledge to accurately interpret and evaluate the content. The decision to restrict participants to those from the United States is supported by the findings of [10], which emphasize the importance of task suitability and specific expertise in crowdsourced tasks. This approach aligns with the recommendation to tailor tasks to the expertise of participants, minimizing misunderstandings.

From the target audience, 12 participants either failed the attention check or revoked their consent, resulting in incomplete entries. The participants were found to be distributed as follows in 4.2 after filtering out the 12 participants.

Intervention	**Unique Participants**
Voluntary	73
Alert	74
Pause	65
Total	**212**

Table 4.2: Number of Unique Participants by Intervention

4.5. Results and Discussions

4.5.1. Demographics

The participant pool showcased a diverse demographic profile, spanning a big age range from 18 to 76, with a median age of 32. Gender representation included 58.0% identifying as female, 39.2% as male, 2.3% as non-binary, and 0.5% opting not to disclose. Educational backgrounds were equally varied, with the majority holding a Bachelor's degree (41.5%) or a high school diploma or equivalent (35.9%). While, 16.5% had achieved a Master's degree, 3.3% held doctorates, and 2.8% prefered not to say. This diverse cohort brought a big mix of perspectives to this study.

4.5.2. Intervention User Engagement

H1 *proposed that the presence of an auditory signal would increase response rates compared to silent prompts by directing user attention toward potential misinformation more effectively.*

H2: *proposed that the combination of a prompt and an auditory signal would result in higher engagement* *than prompts without an auditory signal.*

To explore this, we conducted an Analysis of Variance (ANOVA) to compare the levels of engagement across the different intervention types. ANOVA was chosen as it allows for the evaluation of mean differences among multiple groups while accounting for the variance within and between these groups. This method is effective in experimental setups like ours, where the aim is to determine whether the different type of intervention has a statistically significant impact on engagement [30]. Our analysis revealed a statistically significant difference in engagement among the intervention groups, (F(2, 105) = 5.25, p = 0.007, ω^2 = 0.073). Post hoc Tukey analyses indicated that the significant difference was specifically between the PAUSE (M = 1.21, SD = 0.49) and VOLUNTARY (M = 0.91, SD = 0.38) groups (p = 0.005). Therefore, H1 cannot be rejected, as the PAUSE intervention showed higher engagement compared to the VOLUNTARY condition. However, no significant differences were observed between the VOLUNTARY and ALERT groups (p = 0.284) or the PAUSE and ALERT groups (p = 0.204). These findings suggest that while the PAUSE intervention led to higher engagement than the VOLUNTARY intervention, the presence of an auditory signal in the ALERT intervention did not significantly increase response rates compared to the silent prompts in the VOLUNTARY condition, we therefore reject H2.

The average engagement scores supported these findings: PAUSE showed the highest engagement with 1.212 clicks per podcast, followed by ALERT with 1.049 clicks per podcast, and VOLUNTARY with 0.905 clicks per podcast. The higher engagement in the PAUSE intervention aligns with expectations, as participants were actively prompted to flag misinformation during the intervention. It is important to note that the baseline number of clicks for the PAUSE group was 1, due to the intervention design where participants were forced to answer, whereas the baseline for both ALERT and VOLUNTARY groups was 0.

Furthermore, the post-survey results showed valuable insights into participants' perceptions of their engagement with flagging misinformation. Here we explore 4 questions in particular: ”I actively chose to engage”, ”I felt no need to engage”, ”I felt empowered to flag misinformation”, and ”I forgot to engage”. Each dimension was assessed using a Likert scale from Strongly Disagree to Strongly Agree. We explored these dimensions and how different intervention groups shaped user responses.

The results (see Appendix I) across several survey questions suggest that ALERT may serve as a promising middle ground for fostering engagement in flagging misinformation. In this analysis, responses for “Agree” and “Strongly Agree” are aggregated as agreements, while “Disagree” and “Strongly Disagree” are aggregated as disagreements to provide a clearer comparison across interventions. Participants most agreed with the ALERT intervention (82.43%), closely followed by VOLUNTARY (80.56%), while PAUSE lagged behind (70.77%). Furthermore, we also see that ALERT intervention had the lowest (1.35%) of disagreement compared to the PAUSE (10.77%) and VOLUNTARY (6.94%). For the statement “I forgot to engage,” participants disagreed most strongly with the ALERT intervention (77.03%) compared to PAUSE (66.15%) and VOLUNTARY (73.61%). This highlights ALERT's potential to effectively encourage active participation, indicating that ALERT may be more effective in addressing forgetfulness and helping participants stay engaged.

Moreover, when asked whether participants agreed with the statement, “I did not feel the need to engage,” the ALERT condition again stood out, with the highest percentage of participants disagreeing (90.54%), compared to PAUSE (76.92%) and VOLUNTARY (76.39%). Similarly, for the statement, “I felt empowered to flag misinformation,” ALERT participants showed the highest percentage of agreement (59.46%), surpassing PAUSE (33.85%) and VOLUNTARY (40.28%). These results further support that the ALERT intervention was the most effective in stimulating and empowering participants to engage actively in flagging misinformation.

Despite these findings showing a directional trend where ALERT interventions could potentially offer a balanced approach to fostering active engagement. A conducted Chi-Square test, shown in Table 4.3 revealed no statistically significant differences among the intervention types across all four questions.

The results in Figure 4.3 illustrate the distribution of average flags per podcast across various combinations of topics and truth labels. The visualization shows a relatively consistent pattern, with no significant differences observed between the topics or truth labels that influenced user engagement.

To further explore whether user interaction was impacted by the topic, an analysis was conducted on self-reported familiarity and interest in each topic. By examining participants' engagement in relation to their perceived familiarity and interest in the podcast topics, we aimed to identify any potential patterns or differences in flagging behavior. The analysis of user-reported familiarity and interest scores reveals no significant relationship between these factors and user engagement, as measured by total clicks. The Pearson's correlation coefficient for interest score and total clicks was 0.029 (p = 0.956), in-

Test	X² Value	p-value
Actively Chose to Engage	9.575	0.296
Forgot to Engage	7.588	0.475
Did Not Feel the Need to Engage	8.114	0.422
Felt Empowered to Flag	12.749	0.121

Table 4.3: Chi-Square test results for engagement responses across interventions (Pause, Alert, Voluntary) with a sample size $N = 211$ and degrees of freedom $df = 8$.

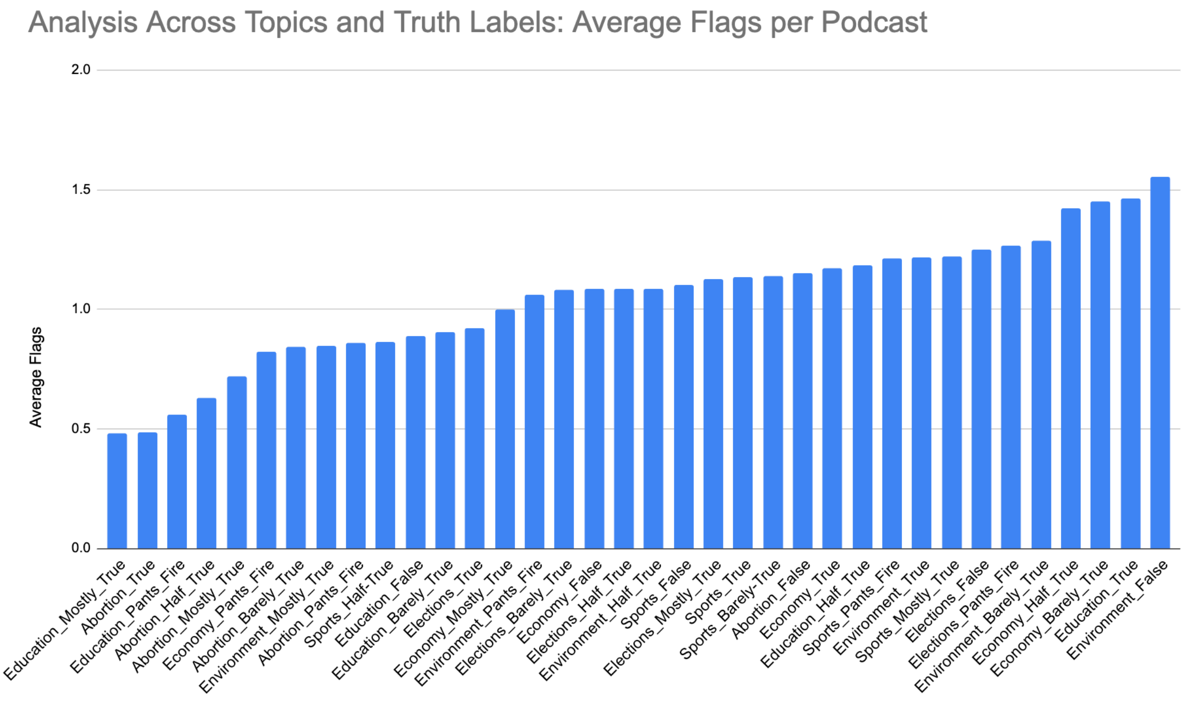


Figure 4.3: The average number of user flagging interactions during a podcast episode across every podcast in the dataset.

dicating no meaningful association. Similarly, familiarity scores were weakly negatively correlated with total clicks (Pearson's r = -0.301, p = 0.563), suggesting that higher familiarity did not translate into higher engagement. These findings indicate that neither familiarity with the podcast topics nor participants' self-reported interest levels significantly influenced their engagement in flagging misinformation. This further supports the observation that user interaction with the flagging process remained consistent across varying levels of personal relevance, suggesting that the intervention, rather than topic familiarity or interest, played a central role in driving engagement.

One participant mentioned that they did not "know how I was supposed to know what was misinformation, as the audio you had me listen to didn't discuss anything I am intimately familiar with." This comment highlights the possible role that topic familiarity may play in participants' ability to identify misinformation, even though the overall analysis did not reveal a significant correlation between familiarity scores and user engagement. It suggests that while familiarity and interest might not directly impact engagement levels, they could influence the confidence or accuracy with which participants assess and flag misinformation. This insight points to an area for further exploration, particularly in understanding how topic-specific expertise might shape users' flagging behaviors and effectiveness in identifying misinformation.

4.5.3. Intervention User Experience

H3: *proposes that the PAUSE intervention would result in a higher cognitive load and be perceived as the most frustrating due to its forced interaction, while the VOLUNTARY intervention would result in the lowest cognitive load.*

It is noteworthy that in Study I, participants were introduced to the idea of three interventions: VOLUNTARY, ALERT, and PAUSE. The results showed that the only 18.18% found the PAUSE intervention acceptable, 51.82% the ALERT intervention and 73.64% found the VOLUNTARY intervention, with 3.64% indicating that none of the interventions were acceptable. These findings helped to rationalize H3, reinforcing the hypothesis that the PAUSE intervention would be perceived as the least acceptable due to its higher cognitive load and forced interaction. However, the ANOVA test results for SUS, UES, and NASA-TLX revealed no statistically significant differences across the PAUSE, ALERT, and VOLUNTARY intervention types. For SUS, the mean scores were similar across interventions (PAUSE: 70.81, ALERT: 71.52, VOLUNTARY: 72.06), with a p-value of 0.907, indicating no meaningful differences in perceived usability between interventions. Similarly, for UES, the mean engagement scores were nearly identical (PAUSE: 3.36, ALERT: 3.33, VOLUNTARY: 3.31), with a p-value of 0.908, suggesting that engagement levels were consistent across intervention types. Finally, for NASA-TLX, while PAUSE showed the highest mean score (4.70) compared to ALERT (4.58) and VOLUNTARY (4.37), the differences were not statistically significant (p = 0.269), in particularly one of the questions in the NASA-TLX was “The flagging tool used to capture your input was acceptable”, where VOLUNTARY scored a (7.06) as most acceptable, followed by ALERT (7.05) and PAUSE (6.98), which are much closer than the expected hypothesis. These findings indicate that within this setting all three intervention types were perceived similarly in terms of usability, engagement, and cognitive load, suggesting that no single intervention had a clear advantage in these measures, therefore, we reject H3, as this suggests that the type of intervention we designed may not have a strong impact on perceived disruption, cognitive load or usability.

However, these results may not fully capture how these interventions would function in other contexts, such as those where users multitask while listening to podcasts. From Study I, we observed that 70.00% of participants reported multitasking during podcast listening, and the post-survey results from this study similarly indicated that 65.57% of participants engage in multitasking. Users that are simultaneously engaged in other activities, such as commuting, working, or exercising, interventions like PAUSE, which interrupt the podcast flow, might impose additional cognitive demands. This disruption could require users to frequently shift attention between the podcast and their primary task, potentially leading to higher cognitive load or increased frustration. Such conditions might not only affect the usability and acceptance of the intervention but could also influence its effectiveness in helping users identify misinformation. Understanding these dynamics in multitasking contexts would be an interesting follow-up to this study.

Conversely, in casual listening settings where users are not explicitly instructed to perform a task like flagging misinformation, the dynamics might differ, potentially shifting their perceptions of interventions. Ideally, the most effective way to test this would be to implement the interventions on an existing podcast platform and observe real-world user behavior over time. However, this approach presents its own limitations such as the potential resistance from users. An alternative approach could involve designing experiments with longer podcast segments that allow users to become more immersed in the content. This setup would better mimic casual listening conditions and help researchers evaluate how interventions are perceived when users are focused on enjoying the content rather than completing a specific task.

Measure	Intervention	N	Mean	SD	F	p
SUS	PAUSE	65	70.808	13.779	0.098	0.907
	ALERT	74	71.520	16.137		
	VOLUNTARY	73	72.055	19.004		
UES	PAUSE	65	3.358	0.544	0.097	0.908
	ALERT	74	3.333	0.613		
	VOLUNTARY	73	3.312	0.671		
NASA-TLX	PAUSE	65	4.704	1.489	1.321	0.269
	ALERT	74	4.576	1.116		
	VOLUNTARY	73	4.365	1.117		

Table 4.4: ANOVA Results for SUS, UES, and NASA-TLX across intervention types.

Other interesting findings from the post-survey results revealed no statistically significant differences between interventions for statements such as “I felt it was a burden to flag potential misinformation” (X^2 = 8.32, df = 8, p = 0.403), and “I felt responsible to flag misinformation” (X^2 = 5.297, df = 8, p = 0.725). Most participants did not perceive flagging as burdensome (VOLUNTARY: 70.36%, PAUSE: 89.95%, ALERT: 81.08%), and the majority agreed or strongly agreed that they felt responsible for flagging misinformation (VOLUNTARY: 56.16%, PAUSE: 63.08%, ALERT: 58.11%). These findings suggest that implementing a flagging intervention does not introduce a feeling of burden to flag misinformation and may foster a sense of responsibility. These findings align closely with results from Study I, where 65% of participants indicated they would be motivated by a sense of responsibility and 55% by altruistic reasons. This reinforces the potential for such interventions to leverage intrinsic motivations for engagement. However, further research is needed to confirm these trends and explore their broader applicability.

4.5.4. Flagging Accuracy

H4: *proposed that accuracy in the auditory-based interventions (PAUSE and ALERT), regardless of the presence of a prompt, will be higher than in the VOLUNTARY intervention.*

Due to an unforeseen technical issue with recording timestamps, we were unable to verify the accuracy of flags in the VOLUNTARY and ALERT interventions. However, we were able to analyze the PAUSE intervention, as the forced nature of the flags allowed us to infer the responses without relying on timestamps, ’FLAG’ clicks in this group did not occur when participants were prompted to evaluate the content. Accuracy was calculated by grouping the six truth labels into two categories: TRUE and FALSE. Labels such as “TRUE,” “MOSTLY-TRUE,” and “HALF-TRUE” were categorized as TRUE, while “BARELY-TRUE,” “FALSE,” and “PANTS-FIRE” were categorized as FALSE. The confusion matrix for the PAUSE intervention in Table 4.5 offers insights into how users identify and classify misinformation during the task. Users correctly identified 65 cases of true flags (true positives) yet also labeled 60 false flags as true (false positives), suggesting that users often misjudge whether the information is true or false. Additionally, users missed 35 true flags (false negatives), failing to recognize certain instances of misinformation. Performance metrics reflect these challenges, with an overall accuracy of 43.61%, indicating that less than half of the predictions were correct. Precision scores for TRUE (43.33%) and FALSE (43.97%) show that users struggled to reliably distinguish between true and false flags, while the recall for TRUE cases (65%) suggests users were more likely to identify misinformation compared to correctly identifying accurate information (FALSE recall at 54.05%). Using an ANOVA test, we further examine whether the accuracy differed among the different topics, which revealed no statistical significance between accuracy and topics (p = 0.160), indicating that participants faced consistent challenges in identifying misinformation regardless of the topics.

In Study I, participants reported a higher average confidence score of 3.745 out of 5, with 70% scoring 4 or above, prior to engaging in the task of identifying misinformation. In contrast, in Study II, after completing the task, participants in the PAUSE intervention cohort reported a lower average confidence score of 3.338 out of 5. This decline in confidence suggests that identifying misinformation is a challenging task that may have prompted participants to reevaluate their abilities, resulting in greater awareness of their limitations.

Despite this reduction in confidence, participants in Study II still exhibited relative confidence in their ability to identify misinformation. However, the accuracy metrics show difficulty in accurately distinguishing true from false information. This gap between perceived and actual ability could be explained by the Dunning-Kruger effect [53], which observes that individuals often overestimate their abilities when they lack sufficient knowledge or expertise to accurately assess their competence. Before attempting the task, participants may have had inflated confidence due to a lack of experience with trying to identify misinformation. However, once they engaged with the task, their confidence may have diminished, depicting more realistic capabilities. In this case, it would have been more generalizable if participants had been asked about their confidence levels both pre-task and post-task. Such an approach would provide a clearer picture of how the task itself influenced changes in confidence.

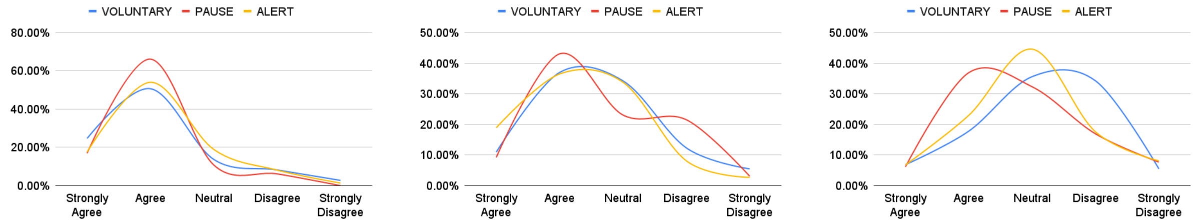
In addition to the forced flags in the PAUSE intervention, participants across all interventions flagged misinformation 153 times. These flags were distributed as follows: 68 were labeled as false, 70 as true, and 15 were skipped. However, since each podcast segment contained at most one false statement, it can be inferred that the majority of these 68 flags were incorrectly labeled as false. It would be valuable to investigate the placement of these flags within the segments to determine whether they coincided with the false statement or occurred elsewhere, potentially in the surrounding context. Without this additional analysis, we cannot conclusively determine the accuracy of these flags.

We further examined participants’ perceptions of the interventions and their impact on misinformation

Table 4.5: PAUSE intervention: confusion Matrix on flags

	Predicted TRUE	**Predicted FALSE**
Actual TRUE	65	35
Actual FALSE	60	51
Actual SKIP	25	30

detection through three statements: ”The intervention helped me to critically think about the content,” ”The intervention helped me to identify misinformation,” and ”Following the flagging intervention, I had more trust in the podcast.” As shown in Figures 4.4, a few clear trends emerge across the interventions. For ”critically think about the content,” the PAUSE intervention had the highest level of agreement, suggesting that its structured, forced engagement effectively encouraged participants to critically evaluate the content. What stands out is that across all three interventions, participants generally perceived that simply having an intervention in place prompted them to think more critically. Similarly, for ”helped me to identify misinformation,” the PAUSE intervention again had the most agreement, reinforcing its role in fostering active evaluation and engagement. The VOLUNTARY and ALERT interventions followed similar trends but were perceived as less impactful, likely due to their more passive or optional nature. Interestingly, when it came to trust in the podcast following the intervention, the ALERT intervention showed the highest level of agreement, suggesting that its less intrusive and more subtle nature may have been effective in maintaining participants’ trust in the content. In contrast, the PAUSE intervention, despite being effective in fostering critical thinking and identifying misinformation, had lower trust scores, likely due to its perceived disruptiveness. The VOLUNTARY intervention exhibited a broader distribution of responses, with participants appearing more neutral, reflecting its reliance on intrinsic motivation and less guided engagement. These findings show the perceived strengths and limitations of each intervention. The PAUSE intervention is effective for encouraging critical thinking and identifying misinformation but may reduce trust. The ALERT intervention balances engagement and trust. Meanwhile, the VOLUNTARY intervention appears to be the least impactful.



(a) The intervention helped participants critically think about the content.

(b) The intervention helped participants identify misinformation.

(c) Following the intervention, participants had more trust in the podcast.

Figure 4.4: Participant perceptions of the interventions: critical thinking, misinformation identification, and trust in the podcast.

Ideally, understanding how the PAUSE intervention compares to the other interventions in terms of accuracy would have been valuable. The PAUSE intervention, by forcing users to make a decision, may trigger their System 2 thinking [45], which involves the deliberate processing of information. Exploring whether the ALERT and VOLUNTARY intervention leads to better accuracy compared to the PAUSE interventions could offer insights into how prompting deeper cognitive engagement affects users’ ability to identify misinformation.

An important question to address is whether users generally perform poorly in flagging misinformation within this context. The challenges observed in accurately distinguishing true from false information could be influenced by factors such as the auditory format of podcasts, the lack of visual aids, or the inherent lack of ability to evaluate what is credible information in real-time. Another critical consideration is the extent to which interventions affect accuracy. If interventions do impact performance, it becomes essential to define what an acceptable baseline accuracy might be for real-world applications. For instance, a 43.61% accuracy rate is likely insufficient for practical use, as it falls below a threshold that would meaningfully aid in identifying misinformation. Such a low accuracy level might lead to an overwhelming number of false positives and negatives, ruining trust in the system. These questions point to a broader issue: whether intervention can be used to help empower users to handle misinformation.

While these interventions may not be able to deal with the challenge of misinformation on its own, it could serve as a layer of filtering, flagging potential misinformation for further review or verifying flagged misinformation. However, ensuring that this process is both effective and scalable requires more research to decide on benchmarks for accuracy and user engagement.

4.6. Future works

4.6.1. labeled podcast dataset

The creation of a labeled podcast dataset lays a foundation for more research in podcast-based misinformation detection. However, future work in audio datasets could benefit from capturing the conversational nuances of podcasts, which often feature multiple speakers and dynamic interactions. These complexities, including speaker roles, turn-taking, and contextual cues, present unique challenges for interpretation. While current tools may fall short in addressing these intricacies, emerging technologies like NotebookLM [26] offers promising potential, which hopefully can offer even more realistic environments for testing podcast misinformation. Alternatively, a more time-intensive but potentially better approach would involve annotating transcripts of existing podcasts. This method could produce longer, more authentic segments that better encapsulate conversational nuances and provide a more accurate representation of misinformation dynamics in real-world audio content.

4.6.2. Application Design

While this study provides valuable insights into the general user experience, several technical and functional issues were identified that need consideration for future work. A couple of participants reported challenges with the flagging button, including unresponsive clicks, the need to click in specific areas near the icon, and issues caused by the tooltip partially covering the icon, which led to misclicks. Additionally, delays in flagging responses and alerts prompting immediately after a flag were noted as areas of frustration. Some participants expressed confusion about the interface functionality, with one mistaking the auditory prompt as part of the podcast content, highlighting the need for improved task-description, such as a demonstration of the alert sound in the instructions. Suggestions for improvement included increasing the size and accessibility of the flagging icon, adding closed captions, and enabling playback at adjustable speeds to accommodate user needs. Making sure these concerns are addressed in future iterations will be crucial for enhancing the usability and accessibility of the flagging system.

Furthermore, the design choices for the flagging icon and alert sound were chosen to be intuitive. The exclamation mark was intentionally chosen for its recognition as a symbol of attention, aiming to provide an intuitive experience. Similarly, the alert sound in the ALERT and PAUSE interventions was selected to subtlety guide participants' attention without causing unnecessary disruption to the podcast's flow. While these design elements were considered, there is room for more deliberate decision-making in design choices that could further help enhance usability and influence user engagement.

4.6.3. Flagging Interventions

Many of these findings need further exploration into how these interventions perform in real-world settings, particularly in multitasking or casual listening contexts where engagement dynamics may shift. Multitasking could impose additional cognitive demands, particularly for interventions like PAUSE, which interrupt the flow of content. Understanding how these interventions impact engagement and usability in such contexts could inform their design for broader applicability.

4.7. Conclusion

This study aimed to optimize user engagement in flagging misinformation within podcasts through the design and evaluation of three intervention types: PAUSE, ALERT, and VOLUNTARY. By exploring these interventions, we aimed to understand how different feedback mechanisms influence user engagement as the primary measure, while also examining cognitive load and accuracy in identifying misinformation as secondary measures. The findings offer valuable insights into the effectiveness of these interventions and contribute to the broader understanding of user behavior in combating misinformation in podcasts. This study provides 2 contributions. The first is a the development of a labeled podcast dataset derived from the LIAR dataset. This contribution addresses the gap in misinformation detection datasets for podcasts, providing a foundation for future research into user engagement and the detection of audio-based misinformation. The second contribution is evaluating the user engagement of three intervention designs to empower user to flag misinformation.

These findings on each of our interventions provide interesting insights. The PAUSE intervention demonstrated significantly higher engagement compared to the VOLUNTARY intervention. Whilst, not having performed worse in the user experience metrics.

The ALERT intervention emerged as an interesting middle ground, showing directional trends in user perceptions on user engagement. These findings suggest that the ALERT intervention may effectively nudge users toward active participation while maintaining a more natural listening experience compared to PAUSE. However, the lack of significant scores across interventions suggests that more research needs to explore this. The ALERT intervention emerged as a compelling middle ground, showing directional trends in fostering engagement. These findings suggest that the ALERT intervention may effectively nudge users toward active participation while offering a more natural listening experience compared to PAUSE. However, the lack of statistically significant differences across interventions highlights the need for further research to better understand its impact. While the VOLUNTARY intervention showed the lowest engagement overall, its performance in the survey questions indicates that it still holds value by preserving the autonomy of listeners to flag misinformation. This intervention provides the autonomy for users who may be intrinsically motivated.

5 Conclusion

Misinformation in podcasts presents a unique challenge due to the inherently passive nature of audio content and the absence of interactive tools commonly found in other media formats, such as comments or likes. As podcasts continue to grow in popularity and influence, there is an urgent need to address this issue. This study aimed to fill this gap by investigating how we can empower users to handle misinformation. In Study I, we examined how listeners respond to misinformation in podcasts, gathering insights into their behaviors and perceptions. Building on these findings, Study II focused on designing and testing interventions aimed at optimizing user engagement in flagging misinformation. Together, these studies provide a foundation for understanding user behavior and lay the groundwork for exploring effective interventions to empower users to handle misinformation in podcasts.

Study I provided valuable insights into podcast listeners' listening habits, trust, confidence, and reactions to misinformation. It also explored the incentives participants found most motivating and their preferences for initial flagging interventions. Participants generally trusted podcasts they listened to regularly (78%) and expressed high confidence in identifying misinformation (70%), with a weak positive correlation between education level and confidence. Verification behaviors showed minimal variance by education level, although higher-educated participants were more inclined to consult academic sources (36%), while the majority (77%) relied on online searches. Participants preferred the initial voluntary feedback interventions (73%) over alert (43%) and pause (18%) interventions, with motivations driven by a sense of responsibility (65%) and altruistic reasons (55%).

Building on the findings of Study I, Study II evaluated three intervention, PAUSE, ALERT, and VOLUNTARY, designed to encourage user engagement in flagging misinformation. Each intervention followed a different approach to eliciting feedback, ranging from forced interaction (PAUSE) to auditory nudges (ALERT) and fully autonomous engagement (VOLUNTARY). The study measured engagement, cognitive load, usability, and accuracy to assess the effectiveness of these interventions. The results revealed notable differences in user engagement. The PAUSE intervention demonstrated the highest engagement due to its forced nature. Conversely, the VOLUNTARY intervention, while the least disruptive, showed the lowest engagement rates, reflecting the challenges of relying solely on user initiative. The ALERT intervention emerged as a promising middle ground, effectively balancing engagement and autonomy through subtle auditory cues that minimized disruption. Although, we see that interventions could have the potential to positively empower users, we need to be weary of the accuracy of participants. Accuracy results from the PAUSE intervention revealed challenges in identifying misinformation, with participants frequently misjudging false statements as true or vice versa. Intrusive interventions have shown potential in empowering users to identify misinformation in podcasts. However, their effectiveness in solving the misinformation challenge remains uncertain, requiring further investigation. This thesis highlights the importance of addressing misinformation in podcasts through user-focused interventions, providing a foundation for future research in this critical area.

A

Study I: Consent Form

Participant Consent Form

You are being invited to participate in a research study titled Exploring How Listeners Handle Misinformation in Podcasts. This study is being conducted by researchers from TU Delft.

The purpose of this research study is to understand the responses of listeners when encountering misinformation in podcasts. Four main points are considered; Demographics and Listening Habits, Trust and Concern about Misinformation in Podcasts, Reaction to Misinformation in Podcasts and Incentives to flag Misinformation in Podcasts. The survey should take you approximately 6 minutes to complete.

The data collected will be used to understand how listeners respond to misinformation in podcasts and identify effective incentives. The results will be aggregated and may be used for publication.

Participants will be asked to complete various types of questions, including Multiple-choice questions, Likert scale questions and Multi-select questions.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by keeping this survey anonymous, which means that IP addresses, location data, and contact information will not be collected. Additionally, we will minimize the collection of personal data. The collected data will be securely stored on OneDrive.

You may stop the survey and withdraw at any time. However, your participation in this study will **only** be compensated after you complete the survey and your responses are verified. Once payment is made, data withdrawal is no longer possible, as the data will be anonymized and can no longer be traced back to an individual.

If you have any questions or concerns, you may contact: **extantudelft.nl**.

B

Study I: How do users respond to misinformation Survey

Section	**Question/Description**	**Answer Options**
Consent Form	Do you consent to participate in the study?	I consent / I do not consent
Demographics	What is your Prolific ID?	Free text
	What is your age?	Free text
	What is your gender?	Male, Female, Non-binary, Prefer not to disclose, Prefer to self-describe
	What is the highest level of education you have completed?	High school, Bachelor's, Master's, Doctorate, Prefer not to say
Listening Habits	How often do you listen to podcasts?	Never, Rarely, Weekly, 2-4 times a week, Every day
	How many hours of podcasts do you typically listen to per week?	Less than 1 hour, 1-3 hours, 4-6 hours, 7-10 hours, 11-15 hours, More than 15 hours
	What types of podcasts do you typically listen to?	News, Educational, Entertainment, Political, Business, Sports, Other
	Why do you listen to podcasts?	Entertainment, Educational, Staying informed, Relaxation, Convenience, Other
	How do you discover new podcasts to listen to?	Recommendations, Social media, Platform suggestions, Public figures, Advertisements, Other
	What activities do you typically engage in while listening to podcasts?	Just listen, Housework, Walking, Exercise, Traveling, Other
Trust and Concern about Misinformation	Rate your agreement with the following statements:	Likert scale (Strongly Disagree to Strongly Agree)
	How often do you encounter misinformation in podcasts?	Frequently, Occasionally, Rarely, Never, I don't know

Section	Question/Description	Answer Options
	What factors most influence your trust in a podcast?	Host credibility, Source credibility, References, Production quality, Popularity, Other
Reaction to Misinformation	How does encountering misinformation affect your listening behavior?	Stop listening, Reduce listening, Continue listening, I do not care
	When encountering potential misinformation, I usually:	Ignore it, Research, Stop listening, Report, Other
	What sources do you use to verify information?	Academic journals, Online search, News outlets, Social media, Personal network, Other
Incentives to Flag Misinformation	Which methods are acceptable for flagging misinformation?	Pausing for feedback, Alert for feedback, Voluntary feedback, None, Other
	Which type of incentive motivates you?	Monetary, Non-monetary, Reciprocity, Responsibility, Altruistic reasons, None
	How likely are you to flag misinformation based on incentives?	Likert scale (Very likely to Very unlikely)
Feedback Section	Would you like to provide further information on your interaction with misinformation?	Free text
End of Survey	Final message thanking participants and redirecting them to submit their responses.	N/A

C

Study I: Chi Square Test - Genre/Listening Intent and Misinformation Reaction

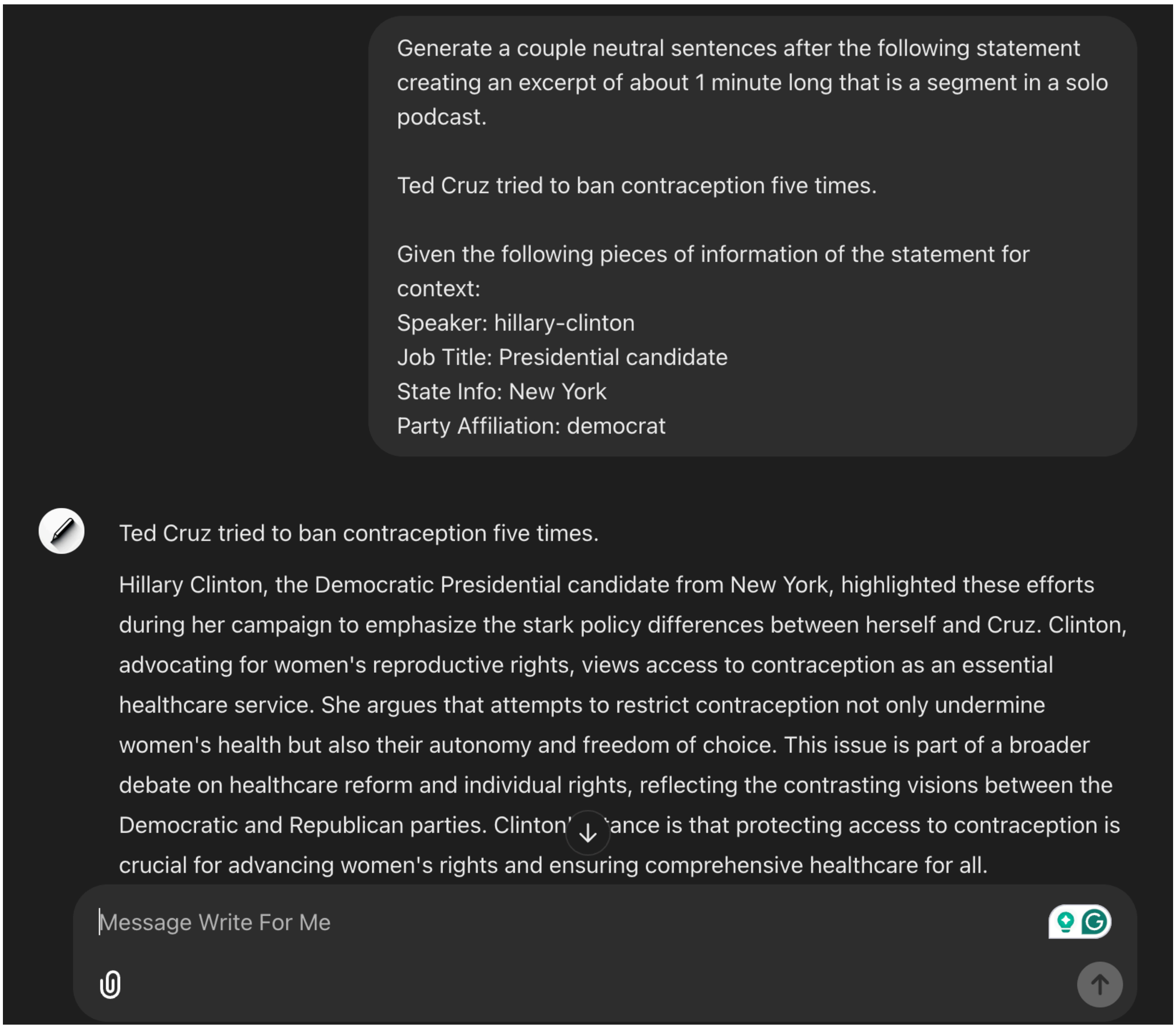
Table C.1: Chi-Squre Test: Reaction to Podcast Episodes by Genre (N = 262)

Reaction Type	Entertainment	Relaxation	Staying Informed	Educational	Convenience	Other	Total
I stop listening	19 (0.841)	10 (0.19)	7 (-1.357)	12 (0.182)	3 (-0.235)	1 (0.589)	52
I reduce listening	19 (0.408)	10 (-0.101)	9 (-0.873)	13 (0.219)	5 (0.836)	0 (-0.908)	56
I keep it in mind but continue listening	39 (-0.412)	24 (0.176)	30 (1.264)	27 (-0.398)	7 (-0.655)	1 (-0.541)	128
I somewhat reduce listening	6 (-0.738)	4 (-0.22)	6 (0.61)	6 (0.354)	2 (0.385)	0 (-0.553)	24
I do not care	0 (-0.967)	0 (-0.672)	1 (1.052)	0 (-0.757)	0 (-0.374)	1 (6.519)	2
Total	83	48	53	58	17	3	262

Note: The first number in each cell represents the count of participants, and the value in parentheses is the standardized residual for that cell. A residual value greater than 2 or less than -2 indicates a statistically significant deviation from expected frequencies.

D

Podcast Data creation prompts



Generate a couple neutral sentences after the following statement creating an excerpt of about 1 minute long that is a segment in a solo podcast.

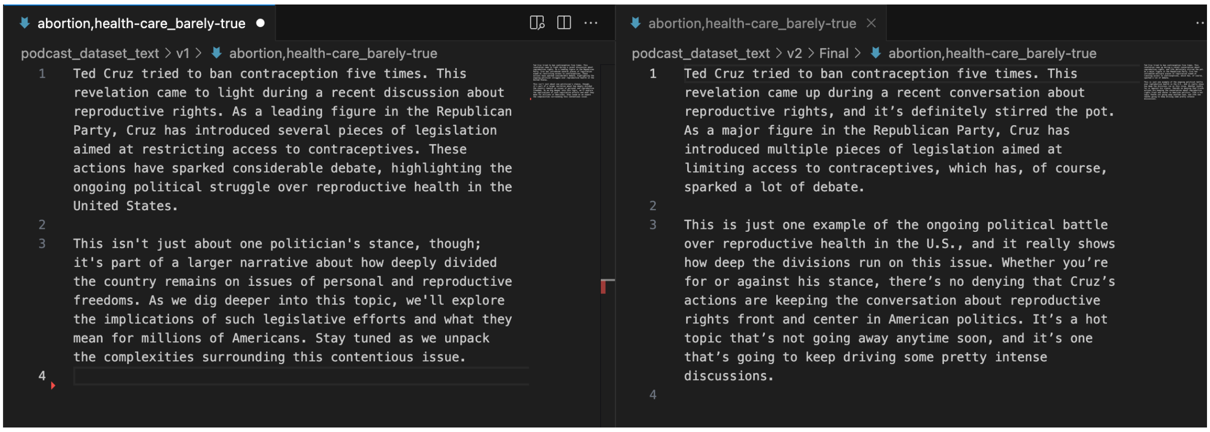
Ted Cruz tried to ban contraception five times.

Given the following pieces of information of the statement for context:
Speaker: hillary-clinton
Job Title: Presidential candidate
State Info: New York
Party Affiliation: democrat

Ted Cruz tried to ban contraception five times.

Hillary Clinton, the Democratic Presidential candidate from New York, highlighted these efforts during her campaign to emphasize the stark policy differences between herself and Cruz. Clinton, advocating for women's reproductive rights, views access to contraception as an essential healthcare service. She argues that attempts to restrict contraception not only undermine women's health but also their autonomy and freedom of choice. This issue is part of a broader debate on healthcare reform and individual rights, reflecting the contrasting visions between the Democratic and Republican parties. Clinton' ↓ 'ance is that protecting access to contraception is crucial for advancing women's rights and ensuring comprehensive healthcare for all.

Figure D.1: Initial Prompt Example


abortion,health-care_barely-true
podcast_dataset_text > v1 > abortion,health-care_barely-true

Ted Cruz tried to ban contraception five times. This revelation came to light during a recent discussion about reproductive rights. As a leading figure in the Republican Party, Cruz has introduced several pieces of legislation aimed at restricting access to contraceptives. These actions have sparked considerable debate, highlighting the ongoing political struggle over reproductive health in the United States.

This isn't just about one politician's stance, though; it's part of a larger narrative about how deeply divided the country remains on issues of personal and reproductive freedoms. As we dig deeper into this topic, we'll explore the implications of such legislative efforts and what they mean for millions of Americans. Stay tuned as we unpack the complexities surrounding this contentious issue.

abortion,health-care_barely-true
podcast_dataset_text > v2 > Final > abortion,health-care_barely-true

Ted Cruz tried to ban contraception five times. This revelation came up during a recent conversation about reproductive rights, and it’s definitely stirred the pot. As a major figure in the Republican Party, Cruz has introduced multiple pieces of legislation aimed at limiting access to contraceptives, which has, of course, sparked a lot of debate.

This is just one example of the ongoing political battle over reproductive health in the U.S., and it really shows how deep the divisions run on this issue. Whether you’re for or against his stance, there’s no denying that Cruz’s actions are keeping the conversation about reproductive rights front and center in American politics. It’s a hot topic that’s not going away anytime soon, and it’s one that’s going to keep driving some pretty intense discussions.

Figure D.2: Example comparison between v1 and v2 of podcast segment creation

E

Podcast Counts by Topic and Truth Label

Table E.1: Podcast Counts by Topic and Truth Label

Topic	Truth Label	Count
Abortion and Healthcare	Barely True	32
	False	40
	Half True	27
	Mostly True	39
	Pants Fire	36
	True	37
Children and Education	Barely True	32
	False	36
	Half True	44
	Mostly True	25
	Pants Fire	34
	True	41
Economy and Taxes	Barely True	40
	False	35
	Half True	33
	Mostly True	34
	Pants Fire	39
	True	29
Elections and Legal Issues	Barely True	37
	False	32
	Half True	35
	Mostly True	40
	Pants Fire	30
	True	37
Energy and Environment	Barely True	35
	False	38
	Half True	35
	Mostly True	33
	Pants Fire	34

Topic	**Truth Label**	**Count**
	True	37
Sports	Barely True	36
	False	29
	Half True	37
	Mostly True	41
	Pants Fire	38
	True	30

F

Study II: Consent Form

Participant Consent Form

Please read the following terms and conditions carefully before proceeding.

You are being invited to participate in a research study titled Empowering Users to handle Misinformation in Podcasts. This study is being done by researchers from the TU Delft.

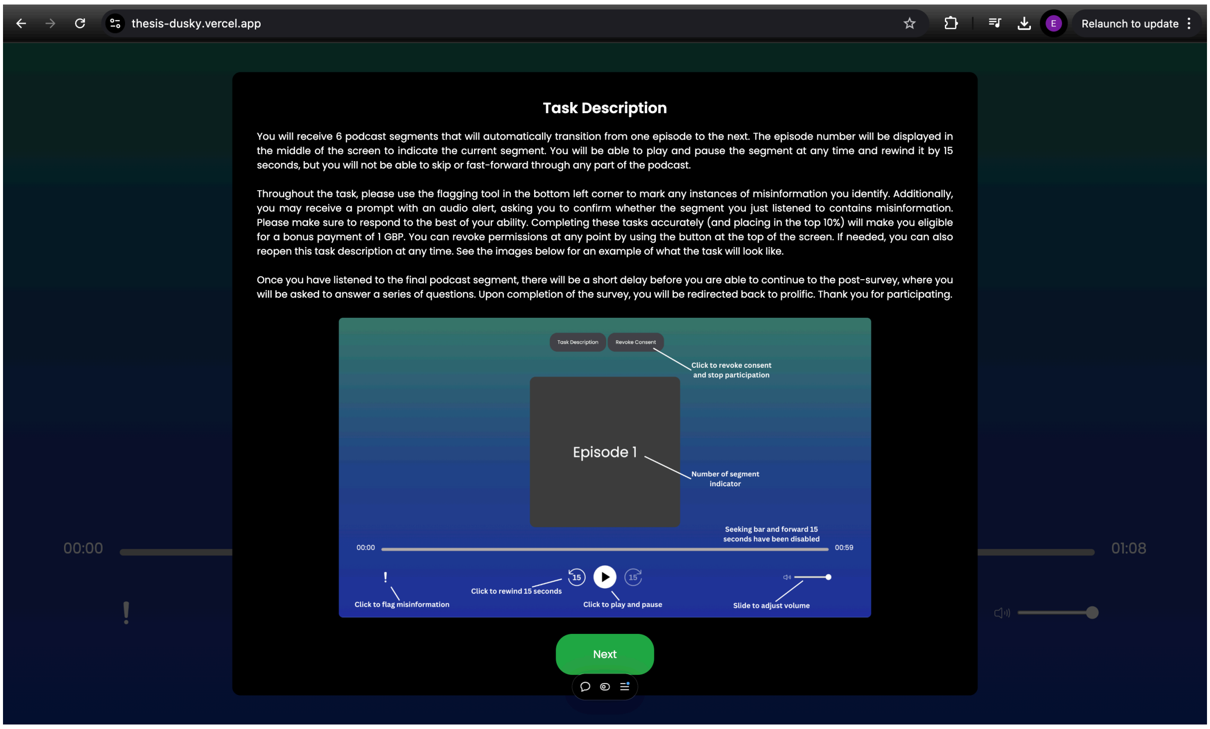
The purpose of this research study is to explore interventions to empower listeners to handle misinformation in podcasts. Users will be asked to listen to 6 short podcast segments and to flag misinformation to the best of their ability. Following this, a short survey will follow asking about the task experience. The entire task should take approximately 15 minutes to complete.

The data collected will be used to understand the engagement and experience of podcast listeners in the given setting. The results will be aggregated and may be used for publication. The survey will contain Multiple-choice questions, Likert scale questions and Multi-select questions.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by keeping the task results anonymous, which means that IP Address, location data, and contact info will not be collected. Additionally, we will minimize the collection of personal data. The collected data will be securely stored on OneDrive.

G

Study II: Podcast Player Task Description



Task Description

You will receive 6 podcast segments that will automatically transition from one episode to the next. The episode number will be displayed in the middle of the screen to indicate the current segment. You will be able to play and pause the segment at any time and rewind it by 15 seconds, but you will not be able to skip or fast-forward through any part of the podcast.

Throughout the task, please use the flagging tool in the bottom left corner to mark any instances of misinformation you identify. Additionally, you may receive a prompt with an audio alert, asking you to confirm whether the segment you just listened to contains misinformation. Please make sure to respond to the best of your ability. Completing these tasks accurately (and placing in the top 10%) will make you eligible for a bonus payment of 1 GBP. You can revoke permissions at any point by using the button at the top of the screen. If needed, you can also reopen this task description at any time. See the images below for an example of what the task will look like.

Once you have listened to the final podcast segment, there will be a short delay before you are able to continue to the post-survey, where you will be asked to answer a series of questions. Upon completion of the survey, you will be redirected back to prolific. Thank you for participating.

Task Description
Revoke Consent
Click to revoke consent and stop participation
Episode 1
Number of segment indicator
Seeking bar and forward 15 seconds have been disabled
00:00
00:59
!
15
15
Click to flag misinformation
Click to rewind 15 seconds
Click to play and pause
Slide to adjust volume
Next

Figure G.1: Podcast Player Web Application

H

Study II: Post-task survey

Section	**Question/Description**	**Answer Options**
Consent Form	Do you consent to participate in the study?	I consent / I do not consent
Demographics	What is your Prolific ID?	Free text
	What is your age?	Free text
	What is your gender?	Male, Female, Non-binary, Prefer not to disclose, Prefer to self-describe
	What is the highest level of education you have completed?	High school diploma or equivalent, Bachelor’s degree, Master’s degree, Doctorate, Prefer not to say
Usability, Perceived Trust, and Intervention Experience	Please select ”The Joe Rogan Experience” and ”Huberman Lab” (Attention check).	Multiple choice: Lex Fridman Podcast, The Joe Rogan Experience, TED Talks Daily, Huberman Lab, The Daily (NYT), Other
	Rate your agreement with the following statements about the flagging process:	Likert scale (Strongly Disagree to Strongly Agree): - Helped me critically think about the content. - Broke the flow of the podcast. - Helped me identify misinformation. - Increased my trust in the podcast. - Increased my engagement with the podcast.
Podcast Listening Habits	What activities do you typically engage in while listening to podcasts?	Multiple select: No other activities, Housework, Walking, Exercise, Traveling, Other
	Rate your agreement with the following statements:	Likert scale (Strongly Disagree to Strongly Agree): - I trust the information in podcasts. - I am confident in identifying misinformation in podcasts. - I encounter misinformation in podcasts.

Section	Question/Description	Answer Options
User Engagement and Cognitive Load	Rate your agreement on the flagging process used to capture input:	Likert scale (Strongly Disagree to Strongly Agree): - I lost myself in this experience. - I felt frustrated using this intervention. - It was taxing to use. - The intervention was aesthetically appealing. - My experience was rewarding. - I felt interested in this experience.
	Rate your mental and physical effort during the intervention:	Scale from 0 (low) to 10 (high): - Mental and perceptual activity required. - Physical activity required. - Time pressure felt. - Satisfaction with performance. - Effort required to achieve performance. - Stress and irritation experienced.
Feedback Mechanism and Preferences	How frequently would an alert to capture input be acceptable?	Multiple choice: As many as necessary, Once every 5 mins, Once every 10 mins, Once every 20 mins, None.
	How familiar and interested are you in the following topics?	Familiarity/Interest scale (Not at all to Extremely): Abortion, Education, Economy, Elections, Environment, Sports.
Open-Ended Feedback	Would you like to provide further information on your interaction with misinformation?	Free text
End of Survey	Final message thanking participants and redirecting to Prolific for submission.	N/A

Study II: Survey results - Intervention Engagement

Actively Chose to Engage

Response	PAUSE (%)	ALERT (%)	VOLUNTARY (%)
Strongly Agree	21.54	21.62	20.83
Agree	49.23	60.81	59.72
Neutral	18.46	16.22	12.50
Disagree	7.69	1.35	6.94
Strongly Disagree	3.08	0.00	0.00

Forgot to Engage

Response	PAUSE (%)	ALERT (%)	VOLUNTARY (%)
Strongly Agree	3.08	5.41	1.39
Agree	9.23	9.46	9.72
Neutral	21.54	8.11	15.28
Disagree	40.00	39.19	38.89
Strongly Disagree	26.15	37.84	34.72

Did Not Feel the Need to Engage

Response	PAUSE (%)	ALERT (%)	VOLUNTARY (%)
Strongly Agree	6.15	1.35	1.39
Agree	13.85	9.46	8.33
Neutral	16.92	17.57	27.78
Disagree	41.54	44.59	37.50
Strongly Disagree	21.54	27.03	25.00

Felt Empowered to Flag

Response	PAUSE (%)	ALERT (%)	VOLUNTARY (%)
Strongly Disagree	7.69	1.35	6.94
Disagree	15.38	8.11	16.67
Neutral	32.31	24.32	26.39
Agree	33.85	59.46	40.28
Strongly Agree	10.77	6.76	9.72

Bibliography

[1] Bill Adair. *The Lessons of Squash, Our Groundbreaking Automated Fact-Checking Platform*. Accessed: 2024-11-12. 2024. URL: `https://reporterslab.org/the-lessons-of-squash-our-groundbreaking-automated-fact-checking-platform/`.

[2] Zachary Adams et al. "Why Is Misinformation a Problem?" In: *Perspectives on Psychological Science* 18.6 (Nov. 2023), pp. 1436–1463. DOI: `10.1177/17456916221141344`. eprint: `Epub2023Feb16`.

[3] Hunt Allcott and Matthew Gentzkow. "Social media and fake news in the 2016 election". In: *Journal of economic perspectives* 31.2 (2017), pp. 211–236.

[4] Riku Arakawa and Hiromu Yakura. "Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: `10.1145/3411764.3445339`. URL: `https://doi.org/10.1145/3411764.3445339`.

[5] K. Arin, Deni Mazrekaj, and Marcel Thum. "Ability of detecting and willingness to share fake news". In: *Scientific Reports* 13 (May 2023). DOI: `10.1038/s41598-023-34402-6`.

[6] Adam Atkins, Vanissa Wanick, and Gary Wills. "Metrics Feedback Cycle: measuring and improving user engagement in gamified eLearning systems". In: *International Journal of Serious Games* 4 (Dec. 2017). DOI: `10.17083/ijsg.v4i4.192`.

[7] Chen Avin, Hadassa Daltrophe, and Zvi Lotker. "On the impossibility of breaking the echo chamber effect in social media using regulation". In: *Scientific Reports* 14.1 (2024), p. 1107. ISSN: 2045-2322. DOI: `10.1038/s41598-023-50850-6`. URL: `https://doi.org/10.1038/s41598-023-50850-6`.

[8] Catherine Beauvais. "Fake news: Why do we believe it?" In: *Joint Bone Spine* 89 (Mar. 2022), p. 105371. DOI: `10.1016/j.jbspin.2022.105371`.

[9] John Bellettiere et al. "Developing and Selecting Auditory Warnings for a Real-Time Behavioral Intervention". In: *American Journal of Public Health Research* 2 (Jan. 2014), pp. 232–238. DOI: `10.12691/ajphr-2-6-3`.

[10] Md Momen Bhuiyan et al. "Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW2 (Oct. 2020). DOI: `10.1145/3415164`. URL: `https://doi.org/10.1145/3415164`.

[11] Leticia Bode and Emily K Vraga. "Correction experiences on social media during COVID-19". In: *Social Media+ Society* 7.2 (2021), p. 20563051211008829.

[12] Elena Broda and Jesper Strömbäck. "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review". In: *Annals of the International Communication Association* 48.2 (2024), pp. 139–166. DOI: `10.1080/23808985.2024.2323736`. eprint: `https://doi.org/10.1080/23808985.2024.2323736`. URL: `https://doi.org/10.1080/23808985.2024.2323736`.

[13] John Brooke. "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189 (Nov. 1995).

[14] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. "Does fake news affect voting behaviour?" In: *Research Policy* 52.1 (2023), p. 104628. ISSN: 0048-7333. DOI: `https://doi.org/10.1016/j.respol.2022.104628`. URL: `https://www.sciencedirect.com/science/article/pii/S0048733322001494`.

[15] Caleb T. Carr and Rebecca A. Hayes. "Social Media: Defining, Developing, and Divining". In: *Atlantic Journal of Communication* 23.1 (2015), pp. 46–65. DOI: `10.1080/15456870.2015.972282`. eprint: `https://doi.org/10.1080/15456870.2015.972282`. URL: `https://doi.org/10.1080/15456870.2015.972282`.

[16] Google Cloud. *Cloud Text-to-Speech*. `https://cloud.google.com/text-to-speech/?hl=en`. Accessed: 2024-11-20. 2024.

[17] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[18] Livio Cricelli, Michele Grimaldi, and Silvia Vermicelli. “Crowdsourcing and open innovation: a systematic literature review, an integrated framework and a research agenda”. In: *Review of Managerial Science* 16.5 (2022), pp. 1269–1310. ISSN: 1863-6691. DOI: 10.1007/s11846-021-00482-9. URL: https://doi.org/10.1007/s11846-021-00482-9.

[19] Molly L. Delaney and Mark A. Royal. “Breaking Engagement Apart: The Role of Intrinsic and Extrinsic Motivation in Engagement Strategies”. In: *Industrial and Organizational Psychology* 10.1 (2017), pp. 127–140. DOI: 10.1017/iop.2017.2.

[20] Tim Draws et al. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Oct. 2021), pp. 48–59. DOI: 10.1609/hcomp.v9i1.18939. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/18939.

[21] Pejman Ebrahimi et al. “User Engagement in Social Network Platforms: What Key Strategic Factors Determine Online Consumer Purchase Behaviour?” In: *Economic Research-Ekonomska Istraživanja* 36.1 (2023), p. 2106264. DOI: 10.1080/1331677X.2022.2106264. eprint: https://doi.org/10.1080/1331677X.2022.2106264. URL: https://doi.org/10.1080/1331677X.2022.2106264.

[22] Ullrich KH Ecker et al. “The psychological drivers of misinformation belief and its resistance to correction”. In: *Nature Reviews Psychology* 1.1 (2022), pp. 13–29.

[23] Seyed Pouyan Eslami, Maryam Ghasemaghaei, and Khaled Hassanein. “Understanding consumer engagement in social media: The role of product lifecycle”. In: *Decision Support Systems* 162 (2022). Business and Government Applications of Text Mining & Natural Language Processing (NLP) for Societal Benefit, p. 113707. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2021.113707. URL: https://www.sciencedirect.com/science/article/pii/S0167923621002177.

[24] Franz Faul et al. “G*Power 3: A flexible statistical power analysis program for the social, Behavioral, and Biomedical Sciences”. In: *Behavior Research Methods* 39.2 (2007), pp. 175–191. DOI: 10.3758/bf03193146.

[25] Renee Garett et al. “A literature review: website design and user engagement”. In: *Online journal of communication and media technologies* 6.3 (2016), p. 1.

[26] Google AI Blog. *NotebookLM: Your AI-powered notebook, grounded in the information you choose*. Accessed: 2024-11-02. July 12, 2023. URL: https://blog.google/technology/ai/notebooklm-google-ai/.

[27] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. “A Survey on Automated Fact-Checking”. In: *Transactions of the Association for Computational Linguistics* 10 (Feb. 2022), pp. 178–206. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00454. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00454.

[28] Sandra G. Hart and Lowell E. Staveland. “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research”. In: *Human Mental Workload*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Advances in Psychology. North-Holland, 1988, pp. 139–183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869.

[29] Farnaz Jahanbakhsh et al. “Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449092. URL: https://doi.org/10.1145/3449092.

[30] Hae-Young Kim. “Analysis of variance (ANOVA) comparing means of more than two groups”. In: *Restorative dentistry & endodontics* 39 (Feb. 2014), pp. 74–7. DOI: 10.5395/rde.2014.39.1.74.

[31] Alex Koo et al. “What Motivates People to Correct Misinformation? Examining the Effects of Third-person Perceptions and Perceived Norms”. In: *Journal of Broadcasting & Electronic Media* 65 (Apr. 2021), pp. 1–24. DOI: 10.1080/08838151.2021.1903896.

[32] Evangelos Lamprou and Nikos Antonopoulos. “Crowdsourcing as a tool against misinformation: The role of social media and user-generated content in overturning misinformation during the Greek Covid-19 pandemic”. In: Dec. 2021.

[33] Sian Lee et al. ““Fact-checking” fact checkers: A data-driven approach”. English (US). In: *Harvard Kennedy School Misinformation Review* 4.5 (2023). Publisher Copyright: © 2023, Harvard Kennedy School. All rights reserved. ISSN: 2766-1652. DOI: `10.37016/mr-2020-126`.

[34] Janette Lehmann et al. “Models of User Engagement”. In: *User Modeling, Adaptation, and Personalization*. Ed. by Judith Masthoff et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 164–175. ISBN: 978-3-642-31454-4.

[35] Jialun Li and Xiaoyi Chang. “Combating Misinformation by Sharing the Truth: a Study on the Spread of Fact-Checks on Social Media”. In: *Information Systems Frontiers* 25 (2023), pp. 1479–1493. DOI: `10.1007/s10796-022-10296-z`.

[36] Huigang Liang et al. “How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: A mediated moderation model”. In: *Computers in Human Behavior* 81 (2018), pp. 168–176. ISSN: 0747-5632. DOI: `https://doi.org/10.1016/j.chb.2017.11.040`. URL: `https://www.sciencedirect.com/science/article/pii/S0747563217306787`.

[37] Mathias-Felipe de-Lima-Santos and Wilson Ceron. “Mind the Gap: Facebook’s measures against information disorder do not go far enough”. In: *Media, Culture & Society* 46.5 (2024), pp. 1075–1090. DOI: `10.1177/01634437241237936`. eprint: `https://doi.org/10.1177/01634437241237936`. URL: `https://doi.org/10.1177/01634437241237936`.

[38] Listen Notes. *Podcast Stats: Facts and Trends*. Accessed: 2024-11-12. 2024. URL: `https://www.listennotes.com/podcast-stats/?srsltid=AfmBOoo-k3c4dBZO2BrCfXD_o7KWYiRZKR0SdwtEHzfcLLmX0dxyQ1O9`.

[39] Xingyu Liu et al. “Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness”. In: *Communication Research* 0.0 (0), p. 00936502231206419. DOI: `10.1177/00936502231206419`. eprint: `https://doi.org/10.1177/00936502231206419`. URL: `https://doi.org/10.1177/00936502231206419`.

[40] Walid Maalej, Hans-Jörg Happel, and Asarnusch Rashid. “When users become collaborators: towards continuous and context-aware user input”. In: *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*. 2009, pp. 981–990.

[41] Cameron Martel, Gordon Pennycook, and David G. Rand. “Reliance on emotion promotes belief in fake news”. In: *Cognitive Research: Principles and Implications* 5.1 (2020), p. 47. ISSN: 2365-7464. DOI: `10.1186/s41235-020-00252-3`. URL: `https://doi.org/10.1186/s41235-020-00252-3`.

[42] Cameron Martel and David G. Rand. “Misinformation warning labels are widely effective: A review of warning effects and their moderating features”. In: *Current Opinion in Psychology* 54 (2023), p. 101710. ISSN: 2352-250X. DOI: `https://doi.org/10.1016/j.copsyc.2023.101710`. URL: `https://www.sciencedirect.com/science/article/pii/S2352250X23001550`.

[43] Cameron Martel et al. “Crowds Can Effectively Identify Misinformation at Scale”. In: (Oct. 2022). DOI: `10.31234/osf.io/2tjk7`.

[44] Mary McHugh. “The Chi-square test of independence”. In: *Biochemia medica* 23 (June 2013), pp. 143–9. DOI: `10.11613/BM.2013.018`.

[45] Patricia L. Moravec, Antino Kim, and Alan R. Dennis. “Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media”. In: *Information Systems Research* 31.3 (2020), pp. 987–1006. DOI: `10.1287/isre.2020.0927`.

[46] Preslav Nakov et al. “Automated Fact-Checking for Assisting Human Fact-Checkers”. In: *CoRR* abs/2103.07769 (2021). arXiv: `2103.07769`. URL: `https://arxiv.org/abs/2103.07769`.

[47] Krishnadas Nanath and Liting Olney. “An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud”. In: *International Journal of Information Management Data Insights* 3.1 (2023), p. 100167. ISSN: 2667-0968. DOI: `https://doi.org/10.1016/j.jjimei.2023.100167`. URL: `https://www.sciencedirect.com/science/article/pii/S2667096823000149`.

[48] Heather L. O'Brien, Paul Cairns, and Mark Hall. "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form". In: *International Journal of Human-Computer Studies* 112 (2018), pp. 28–39. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2018.01.004. URL: https://www.sciencedirect.com/science/article/pii/S1071581918300041.

[49] OpenAI. *GPT-4*. https://openai.com/index/gpt-4/. Accessed: 2024-11-19. 2024.

[50] Stefan Palan and Christian Schitter. "Prolific.ac—A subject pool for online experiments". In: *Journal of Behavioral and Experimental Finance* 17 (2018), pp. 22–27. ISSN: 2214-6350. DOI: https://doi.org/10.1016/j.jbef.2017.12.004. URL: https://www.sciencedirect.com/science/article/pii/S2214635017300989.

[51] Folco Panizza et al. "Lateral reading and monetary incentives to spot disinformation about science". In: *Scientific Reports* 12 (Apr. 2022). DOI: 10.1038/s41598-022-09168-y.

[52] Parth Patwa et al. "Fighting an Infodemic: COVID-19 Fake News Dataset". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Ed. by Tanmoy Chakraborty et al. Cham: Springer International Publishing, 2021, pp. 21–29. ISBN: 978-3-030-73696-5.

[53] Gordon Pennycook et al. "Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence". In: *Psychonomic Bulletin & Review* 24 (Feb. 2017). DOI: 10.3758/s13423-017-1242-7.

[54] PolitiFact Staff. *Who is Robert Malone? Joe Rogan's guest was a vaccine scientist, became an anti-vaccine darling*. Accessed: 2024-11-02. Jan. 6, 2022. URL: https://www.politifact.com/article/2022/jan/06/who-robert-malone-joe-rogans-guest-was-vaccine-sci/.

[55] Ben Rein. "Harnessing social media to challenge scientific misinformation". In: *Cell* 185.17 (Aug. 2022), pp. 3059–3065. ISSN: 0092-8674. DOI: 10.1016/j.cell.2022.07.001. URL: https://doi.org/10.1016/j.cell.2022.07.001.

[56] Congjing Ren. "A comparative user study of misinformation intervention techniques on search engines". PhD thesis. University of British Columbia, 2021. DOI: http://dx.doi.org/10.14288/1.0401775. URL: https://open.library.ubc.ca/collections/ubctheses/24/items/1.0401775.

[57] Piero Ronzani et al. "How different incentives reduce scientific misinformation online". In: *Harvard Kennedy School Misinformation Review* (Jan. 2024). DOI: 10.37016/mr-2020-131.

[58] Carlos Diaz Ruiz and Tomas Nilsson. "Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies". In: *Journal of Public Policy & Marketing* 42.1 (2023), pp. 18–35. DOI: 10.1177/07439156221103852. eprint: https://doi.org/10.1177/07439156221103852. URL: https://doi.org/10.1177/07439156221103852.

[59] John Rula et al. "No "one-size fits all": Towards a principled approach for incentives in mobile crowdsourcing". In: Feb. 2014. DOI: 10.1145/2565585.2565603.

[60] Patrick Schober, Christa Boer, and Lothar Schwarte. "Correlation Coefficients: Appropriate Use and Interpretation". In: *Anesthesia & Analgesia* 126 (Feb. 2018), p. 1. DOI: 10.1213/ANE.0000000000002864.

[61] Ricky J. Sethi. "Crowdsourcing the Verification of Fake News and Alternative Facts". In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT '17. Prague, Czech Republic: Association for Computing Machinery, 2017, pp. 315–316. ISBN: 9781450347082. DOI: 10.1145/3078714.3078746. URL: https://doi.org/10.1145/3078714.3078746.

[62] Hamidreza Shahbaznezhad, Rebecca Dolan, and Mona Rashidirad. "The Role of Social Media Content Format and Platform in Users' Engagement Behavior". In: *Journal of Interactive Marketing* 53 (2021), pp. 47–65. ISSN: 1094-9968. DOI: https://doi.org/10.1016/j.intmar.2020.05.001. URL: https://www.sciencedirect.com/science/article/pii/S1094996820300992.

[63] Kai Shu et al. "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media". In: *CoRR* abs/1809.01286 (2018). arXiv: 1809.01286. URL: http://arxiv.org/abs/1809.01286.

[64] Spotify Newsroom. *International Podcast Day: Transcripts, Chapters, Show Pages, and More*. Accessed: 2024-11-02. Sept. 28, 2023. URL: https://newsroom.spotify.com/2023-09-28/international-podcast-day-transcripts-chapters-show-pages-global/.

[65] Statista. *Number of podcast listeners worldwide from 2019 to 2024*. Accessed: 2024-06-20. 2024. URL: https://www.statista.com/statistics/1291360/podcast-listeners-worldwide/.

[66] Statista. *Number of podcasts and podcast episodes worldwide in 2023*. Accessed: 2024-06-20. 2024. URL: https://www.statista.com/statistics/1418185/podcasts-and-podcasts-episode-worldwide/.

[67] Fabius Steinberger, Ronald Schroeter, and Christopher N. Watling. “From road distraction to safe driving: Evaluating the effects of boredom and gamification on driving behaviour, physiological arousal, and subjective experience”. In: *Computers in Human Behavior* 75 (2017), pp. 714–726. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2017.06.019. URL: https://www.sciencedirect.com/science/article/pii/S0747563217303904.

[68] Stephanie Tobin and Rosanna Guadagno. “Why people listen: Motivations and outcomes of podcast listening”. In: *PLOS ONE* 17 (Apr. 2022), e0265806. DOI: 10.1371/journal.pone.0265806.

[69] Sebastian Tschiatschek et al. “Detecting Fake News in Social Networks via Crowdsourcing”. In: (Nov. 2017).

[70] Ward van Zoonen, Vilma Luoma-aho, and Matias Lievonen. “Trust but verify? Examining the role of trust in institutions in the spread of unverified information on social media”. In: *Computers in Human Behavior* 150 (2024), p. 107992. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2023.107992. URL: https://www.sciencedirect.com/science/article/pii/S0747563223003436.

[71] Valentina Vellani et al. “The illusory truth effect leads to the spread of misinformation”. In: *Cognition* 236 (2023), p. 105421. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2023.105421. URL: https://www.sciencedirect.com/science/article/pii/S0010027723000550.

[72] William Yang Wang. ““Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: https://aclanthology.org/P17-2067.