The Impact of Contextual Information in Social Media Posts on Human Deepfake Detection Accuracy

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Cyber Security Engineering
June 2026
Sini Heikkilä

Supervisors:
Jouni Isoaho
Tahir Mohammad
Aisvarya Adeseye

UNIVERSITY OF TURKU
Department of Computing

Sini Heikkilä: The Impact of Contextual Information in Social Media Posts on Human Deepfake Detection Accuracy

Master of Science (Tech) Thesis, 100 p.
Cyber Security Engineering
June 2026

Deepfakes are form of synthetic media. They are generated using deep learning algorithms to create realistic but misleading representations of reality. Currently, the deepfake technology is advancing fast and humans are having difficulties separating deepfakes from authentic media, making deepfakes a serious threat. Based on existing research, human deepfake detection accuracy is approximately 60.6%. In addition to low human deepfake detection accuracy, technical deepfake detection solutions have challenges in real-world implementations. Deepfakes are commonly shared and encountered on social media platforms, making these platforms the primary environment for deepfakes. This means that there is a need for human deepfake detection improvement strategies especially on social media as current technical deepfake detection systems are not yet sufficient on these platforms. Existing deepfake detection research is mainly focusing on technical solutions or is not sufficiently considering real-world environment of deepfakes, making research on human deepfake detection in context of social media important.

This thesis examines the impact of contextual information in social media posts on human deepfake detection accuracy and proposes and tests a new improvement strategy called Intent Labeling. It first presents systematic literature review of 88 papers. A survey-based experimental study is then applied to collect quantitative and qualitative data from 73 cybersecurity students. The experimental study has three conditions testing impact of contextual information and Intent Labeling on deepfake detection accuracy. Within-subjects design is applied. In addition to detection accuracy, confidence and maliciousness ratings, contextual cues used in the detection task and opinions on Intent Labeling are collected and analyzed.

Results show that context has impact on detection accuracy. Quantitative data shows that giving suspicious context for image has an effect: detection accuracy increases when deepfake is given with suspicious context and decreases when authentic image is given with suspicious context. Students were more confident when context was given, even though higher confidence was not correlating with higher accuracy. Qualitative data provides insights into what cues students used in the detection task, including visual details, quality of the post, consistency, external knowledge, source, intent and interactions.

Keywords: cyber security, human deepfake detection, context, contextual information, social media, improvement strategy

Contents

List of Figures

List of Tables

1 Introduction

Deepfakes are a rapidly advancing technology that uses deep-learning to create synthetic media that looks highly realistic but is misleading representation of reality [1], [2]. Deepfakes can be in textual, image, audio, video or live format, of which the real-time live deepfake technology is now in interest of many developers [3]. Deepfakes have been ranked as the most serious artificial intelligence (AI) threat in 2020 [2] and as the deepfake manufacturing capacity is progressing faster than the deepfake detection methods [4], deepfakes remain as a serious threat in 2026.

Even though deepfake technology has its positive use cases, for example in the health sector by creating more training data for medical imaging [5], it also has many ways to threaten individual citizens and even entire societies. Deepfakes have been used to enhance social engineering attacks [6], to trick home voice assistants and biometric authentication systems [7], to create non-consensual porn causing psychological and reputational harm and to execute financial fraud schemes and impersonation scams [8]. All of these can directly affect individuals and most of the listed threats could be mitigated by improving human deepfake detection as they target individuals. Deepfakes are also being used for spreading misinformation and to manipulate election results, impacting politics [8]. It is also hypothesized that the use of deepfakes could lead to corruption of trust in society, as humans cannot separate real information from false information anymore [8]. However, the actual impact of these potential harms to society can only be seen in the future.

What makes deepfakes a more serious threat is the fact, that creating high quality deepfakes is currently possible even by amateurs, without technical expertise [9]. There are deepfake generation tools available that are easy and fast to use. Regulation of these tools and AI use in general is still insufficient even with the EU Artificial Intelligence Act being implemented in 2024 [10].

1.1 Problem statement

Technical solutions have been implemented to detect deepfakes and combat these threats but many of the solutions are insufficient in the real-world scenarios, which often means detecting and preventing the spread of deepfakes on social media platforms [11]. Technical methods still lack robustness and because they are not yet in general use, they are not yet effective as mitigation [12]. This means that non-technical measures could be used as a mitigation, at least until the technical detection methods become reliable [2]. However, even humans have difficulty distinguishing the high-quality deepfakes from real media, and human deepfake detection accuracy is approximately at chance level. In [13] authors state that the overall mean accuracy for human deepfake detection of different deepfake types have been 60.6% based on the existing research. Not only is better human deepfake detection important to prevent people from believing in fake media, it should also help people prevent identifying a real piece of media as a deepfake [14]. Therefore, it is important to research, if human deepfake detection accuracy can be improved.

Despite that, testing and improving of human deepfake detection are not much researched [2], [13], [15]. More research has been done to develop and improve technical deepfake detection methods, even though the importance of human deepfake detection is noticed in many existing studies. In [16] the author notices the possibility of an endless arms race if the focus is only on development of better technology and highlights the education of the social media users as an effective mitigation.

In [14] it is claimed that technical platform-level solutions have challenges with cropped, compressed and edited images and are not yet robust, making human detection ability and its enhancement important. In [2] it is noted that deepfakes can take analogue form in addition to their digital form, as the images can be printed out onto a paper or a card. In this case, there could be handheld deepfake scanners, but it could also be that human detection ability is needed in this situation. When thinking about the overall security, this kind of cases need also be considered [2]. Another reason, why research on human deepfake detection is needed, is that there are differences between how humans and technology models detect deepfakes and it cannot be said that the models always perform better than humans [15]. There are aspects that humans are currently better than models, including holistic facial processing and understanding of contextual information [17].

In existing research, context has been used in the technical solutions to help models detect misinformation more accurately, but this approach has been challenging [18]. However, as humans can process contextual information better, it could be used to improve human deepfake detection accuracy. Many works of human deepfake detection research have presented experiments of the detection accuracy in controlled lab environments, that are not corresponding to the real-world environment, where people would have at least some contextual information in addition to the actual deepfake. Research article [12] presents an experiment with conditions similar to online platform environmental conditions, but it excludes contextual information like captions, comments and the source, and sees it as a limitation. The potential of using contextual information is also mentioned in [7], observing that giving a bit of context for each clip in the experiment could have helped the participants and in [19], claiming that the participants were concerned about not having necessary context to properly evaluate the given images and videos. However, it is not researched if the contextual information will actually increase the detection ability. As is stated

in [20]: "context might help or hinder people in their detection abilities".

In conclusion, the problem statement consists of the following three main observations: "Humans are encountering deepfakes and therefore deepfake threats on social media platforms", "There is a need for improved human deepfake detection accuracy and situations where using non-technical detection is more meaningful" and "Existing research notices potential of using contextual information in deepfake detection but has not sufficiently examined whether contextual information on social media posts helps or harms human deepfake detection accuracy".

1.2 Research questions and objectives

In this thesis a new human deepfake detection improvement strategy is proposed and the impact of contextual information on human deepfake detection accuracy is being examined. The research questions are the following:

RQ1 What kind of contextual information can be acquired from a social media post?

RQ2 What human deepfake detection improvement strategies there are in existing literature and how effective they are?

RQ3 Does contextual information increase, decrease or have no impact on the deepfake detection accuracy of cybersecurity students?

RQ4 Can Intent Labels improve human deepfake detection accuracy?

Three research objectives are needed to answer the research questions.

Objective one: Classify different examples of contextual information mentioned in existing research into categories and examine what kind of contextual information is usually related to malicious intent and what kind of contextual information is usually related to trustworthiness.

Objective two: Gather existing human deepfake detection improvement strategies from literature and analyze strengths and challenges of each existing strategy.

Objective three: Gain quantitative and qualitative data of students' deepfake detection accuracy and contextual cues used in different conditions to analyze the impact of contextual cues and Intent Labeling improvement strategy.

1.3 Structure of the thesis

The rest of the thesis is organized as follows. First, chapter two is a systematic literature review covering state-of-art detection methods, findings from existing human deepfake detection research and improvement strategies as well as presenting how contextual information has been used in related fields. Different deepfake types are discussed and defined in more detail, to make it clear how the term "deepfake" is used in this thesis. Social media is also presented as an environment where deepfakes are usually encountered, and its impact is discussed. In addition to finding what kind of contextual information have been used in existing research, the literature review also tries to find what variables make humans perceive deepfakes as credible.

Chapter three explains the research methodology. It has methodology for both: systematic literature review (chapter two) as well as for the following survey-based experimental study (chapter four). This chapter justifies the methodology choices made in this thesis and explains how the research questions are answered.

Chapter four is explaining the design process of the survey-based experimental study. It summarizes the main findings from literature review and proposes a concept of a new improvement strategy. It explains the creation of the dataset, survey and experiment process.

Chapter five discusses and analyses the results. It also discusses limitations of the study. Finally, chapter six is the conclusion.

2 Systematic literature review

This chapter defines what is included in the term "deepfakes". It provides a literature review of the state of current deepfake detection methods and their challenges. Both technical and human deepfake detection are introduced. This chapter also presents findings related to contextual information and how it has been used in related fields in existing literature.

2.1 Deepfakes

The definition of deepfakes has varied over different studies, but there are definitions that are more cited than the others [1]. Characteristics often mentioned include realistic look that can deceive humans into believing the deepfake actually represents reality and how deepfakes can be used to represent people saying or doing something that never happened [1]. Mentioning deep learning has also been common, as the term "deepfake" includes "deep learning" in it, in addition to "fake", and because deepfakes are primarily created with deep learning algorithms to synthesize and modify images, sounds and videos [1], [4]. Use of AI in the creation process of a deepfake should be necessary, to separate them from traditional photo editing like computer-generated imagery (CGI). Traditionally edited media can also be seen as synthetic media, but they are not deepfakes. In addition to high-quality editing like CGI, traditional editing contains simpler methods like cropping and cutting. Terms "cheapfakes" and "shallowfakes" have been used to describe this kind of fake videos

that usually are of lower quality than deepfakes [1].

However, the line between legitimate processing and deepfakes is still problematic. Digital media is never representing reality exactly as it is as cameras must process captured raw data to store it as image or video file. Same can be said about microphones. Currently, some phone cameras are using AI tools for image processing by default, but it does not always make the image a deepfake [21]. In addition, nearly all social media posts are edited some way, for example by adjusting contrast or brightness or using blurring effects to make distorted media to the perfect one [22]. Therefore it should be noted that subtle editing does not make the authentic image a fake, as most of the images are being edited to some extent, and none of them perfectly represents reality.

Different studies have different viewpoints of what types of media are included in the term "deepfake" [1]. Original deepfake, meaning the first time the term was used in this purpose, was a video published on Reddit in 2017, so some research has defined deepfakes to be only videos [1]. In this thesis, deepfakes are seen more broadly. Deepfakes can include videos, images, audio and text, as well as the live deepfake that is more complex than a pre-created or modified video. Only focusing on videos does not capture the full range of synthetic, hyper-realistic media falsely representing reality that can be made with deep learning and large language models (LLMs) [1]. AI generated text has often been excluded from deepfakes in existing studies, but there are also literature that sees text as part of deepfakes. In this thesis, highly realistic text, generated for example with LLMs, is included as deepfake. Reasoning for the choice is that deepfake text can be important part of multimodal deepfakes, for example as caption of a post. As text is often part of the context on social media, it is important to include it in this thesis.

Deepfakes can be further divided into being partially or fully deepfakes. Both of these types can be encountered in real world scenarios. Partial deepfakes have

some part of the original authentic media, and combine it with a part of a deepfake [23]. For example, partial deepfake could have part of original video, then have one short deepfake part, and then continue as the original video. It can also be that for example in a video, its visual component is authentic but the audio component is deepfake. In this thesis, partial deepfakes are not used.

2.2 Technical deepfake detection

Current state-of-the-art deepfake detection capabilities achieve accuracy of 90-95% but have challenges with scalability, complexity and keeping up with the evolving deepfake techniques [24]. However, in real-world scenarios, the accuracy achieved drops to approximately 65% [25]. In this section, the most common approaches are presented and their strengths as well as the general challenges of technical deepfake detection techniques are discussed.

2.2.1 Deepfake detection techniques

There has been many different approaches to deepfake detection in the existing literature. These approaches include traditional image forensics, deep learning-based detection, physiological signal analysis and hybrid techniques. Traditional image forensics include use of handcrafted features to separate authentic from fake. Pixel-level analysis, compression artifact detection and metadata and error-level analysis are examples of these techniques [26]. Deep learning-based methods include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and transformers like Vision Transformers (ViTs), Interpretable Spatial-Temporal Video Transformer (ISTVT) and Convolutional Vision Transformers (CViT) [26]. These methods are used to detect occurring patterns that then help detecting deepfakes [26]. Different models can be combined to construct novel hybrid models [26]. Psy-

chological signals include for example detecting deepfakes based on the skin color changes that are caused by blood circulation [26]. Proactive approaches include watermarking that is based on Generative Adversarial Networks (GANs) [26]. Different approaches are further explored in the following subsections.

Face and body analysis techniques

These techniques try to identify anomalies and inconsistencies that are present in deepfakes like lack of blinking, facial deformities and erratic movement [3], [4]. These techniques include Facial Landmark Detection, Blink Analysis and Lip Synchronization Detection [3]. The first one tracks facial features and analyzes if they align correctly over time [3]. Blink analysis tries to track unnatural blinking patterns and Lip Synchronization Detection analyzes if mouth movement corresponds to the spoken words [3]. These techniques can help detecting deepfakes and manipulated content by detecting unnatural facial behaviors, but for the same reason the scope is limited only to deepfakes portraying humans. These techniques are also less effective when the deepfakes are of high-quality meaning they do not have these inconsistencies in them [3], or if the resolution or lighting is of low quality, as these methods focus on the small details.

Image and video analysis techniques

Similar to face and body analysis techniques, these techniques try to find inconsistencies from the content. The techniques include Inconsistencies in Resolution, Temporal Analysis and Metadata Examination [3]. Sudden changes in image quality within a video, inconsistent facial expressions or movements, and suspicious metadata like out-of-ordinary modification timestamps can help detecting a deepfake [3]. These methods could be used with other techniques to increase the detection accuracy, even though they can be ineffective to videos of low quality [3].

Audio analysis techniques

Inconsistencies can also be detected from audio modality. Techniques used in audio deepfake detection are Source Verification and Audio-Visual Synchronization. Source verification means analyzing the speaker's voice characteristics like pitch, tone, rhythm, and speech patterns and finding inconsistencies from them, and Audio-Visual Synchronization detects if the audio matches the visual modality [3]. Speech deepfake detection methods can also use the knowledge of human anatomy. It is possible to reconstruct anatomical features of the vocal tract by analyzing frequency patterns during certain phoneme transitions of the speaker, to be able to detect deepfake artifacts [23].

Deep-learning models

Deep learning models are effective in deepfake detection as they can automatically learn relevant features [9]. Techniques in this category include CNNs, RRNs and Siamese Networks [27]. Transfer learning approaches have also been researched in deepfake detection task [27]. CNNs can be used to recognize specific patterns from images and videos that have been left by generative models when the deepfake have been generated [27]. They are also effective for spatial analysis [27]. Combination of CNN and FastText have been used for deepfake text detection with promising results [28]. RRNs are better suitable for temporal analysis. With them it is possible to track the consistency of videos over time [3], [27]. Siamese networks are used to compare similarities between frames [3]. Transfer learning approaches have high accuracy in deepfake image detection. These methods include feature extraction, fine-tuning and hybrid approaches [29]. The strengths of transfer learning methods are improved accuracy even with limited samples, and robustness in detecting sophisticated deepfakes [29]. However, while deep learning-based models are effective, they too have challenges like needing large, diverse and good-quality datasets, as

well as high computational resources.

Generative models

GANs are used in deepfake creation, but they can also be used in deepfake detection in a different manner [3], [27]. Generative model techniques include Model Artifacts, Detection of GAN Noise [3] and Watermarking. GANs often leave specific artifacts into deepfakes that can be detected, and they also add specific noise patterns to deepfake images making detection possible [3]. However, as GANs evolve, also the deepfakes evolve and these methods could become unreliable [3]. Watermarking as a technique is proactive mitigation method. GANs could "add imperceptible signals to AI-created images so that they can be later identified as machine-generated by a detector" [30]. This approach could be effective, but it requires developers to actually implement the watermarking into their systems.

2.2.2 Challenges of technical deepfake detection

Traditional detection techniques have challenges with generalization, and these techniques are often affected by post processing, compression and adversarial perturbations [9]. Generalization across datasets and different deepfake types is also a challenge for detection models using machine learning [23], [27], [31]. Domain shift impacts model generalization [29] and models overfit to training artifacts making them unreliable [9]. Another challenge is the difficulty to adapt to new deepfake synthesis techniques, as the field is developing fast [9]. Practical implementations should adapt better to evolving deepfake generation techniques [29], meaning that models need to be updated continuously [32].

Many of the challenges with techniques utilizing AI models like deep learning, are caused by insufficient and biased datasets. These detection techniques are dependent on the datasets they are being trained and tested on, and these datasets should be

diverse and of high quality [32]. This is not always the case, and dataset bias limits reliability of the models [32], [27]. The problem with data scarcity and limited access to data [32], [29], is one of the reasons that detection models perform worse with real-world data. Using only benchmark datasets causes loss of variation of real-world deepfakes under compression, noise and varying lighting conditions, making the model less reliable [32]. Models can also have difficulties with beautification filters and super resolution, causing decreased performance with real-world data [33]. High compression and low resolution media are challenging for other detection techniques too, not only for techniques based on machine learning [32], [34]. Related to insufficient datasets and real-world data, there has been a trending scenario where images or videos have a box overlay that contains text inside [35] and challenges persist with obstructed or missing face data [34]. One study [36] presents new high-quality, Human-Indistinguishable Deepfake Dataset and claims that existing research is using deepfake datasets of low quality not corresponding to real world deepfakes. However, focusing only on high-quality is not considering compressed and low resolution deepfakes, so it also has these same challenges. With media including text or speech, there also exists language bias in the datasets, and most research has been done focusing on English [23]. There is not many datasets including much accent variation in other languages than English [23].

Many of the detection models have high computational complexity [32], [27], [31] and the deepfake patterns are also complex [32]. This complexity does not only cause computational overheads when deployed in large scale [29] and high computational costs making the models inapplicable for many resource-limited environments [27], [32], but also makes these techniques lack interpretability [9]. These models are difficult to understand as majority of models are black boxes, which can cause problems in legal and forensic applications [32]. Even though traditional techniques are more interpretable than techniques based on machine learning, these techniques can

also fail to provide explainability and traceability [9]. Additionally, not all research papers provide source code for replication, only about 10% of published papers do so [23]. There is also the threat of using third-party datasets, because they can inject malicious backdoors to be able to later manipulate the detector performance [37], [38]. Adversarial attacks are increasing threat [32], [27], [31], [38]. More research should be made about incident response plans and tools addressing deepfake attacks [6].

Future work in technical deepfake detection should "prioritize integrating multimodal data, refining algorithms for emerging techniques, enhancing real-time detection capabilities, expanding datasets, and addressing ethical considerations" [39]. Lack of research in detection of real-time deepfakes is mentioned also in [23]. Another critical topic that is not much researched is the problem of individual fairness [40]. This should be further researched to improve technical deepfake detection methods.

2.3 Human deepfake detection

In this section, studies of human deepfake detection are discussed. First, general findings are presented and then different deepfake types are discussed in more detail. Findings about factors impacting human deepfake detection accuracy from existing literature are gathered.

Based on a review of 56 papers [13], participants of human deepfake detection tasks performed better at detecting real stimuli, with 68.08% accuracy, than detecting deepfake stimuli, which achieved 55.54% rate. Different deepfake types have different detection accuracies, videos having the highest accuracy at 63.26%, then audio at 63.11%, images 58.04% and text being the lowest at 58.00%. Similar results are presented in [12], where audiovisual stimuli had highest accuracy, and visual stimuli performance was the worst. In [41] it has been noticed that using vi-

sual cues often leads to incorrect choice among participants from Gen Z. This could be because visual modality of deepfakes is the most advanced and seems the most realistic.

Study emulating typical online feed browsing achieved overall accuracy close to change-level 50%, with results that show decreased detection rates when content is synthetic compared to authentic, images of human faces instead of objects and single-modality stimuli compared to multimodal stimuli [12]. Survey covering 3002 participants from three countries (USA, Germany and China) supports the claim that current deepfakes are difficult to distinguish and majority of participants were guessing when asked if the content is authentic or machine-generated [20]. Performance varied also depending on the dataset used in the task, as well as the measurement methods of the accuracy [17]. On average, the detection accuracy was from 57.6% to 75.43% and the studies often use these accuracy ratings to calculate other metrics like AUC (Area Under the Curve), F1 score and recall that make it possible to compare human detection accuracy to machine learning model accuracy [17]. Higher quality deepfakes are more difficult to detect, and familiarity with the subject in the deepfake improves the accuracy [17], making the datasets in use important to consider. Different experiment setups have also impact on accuracy, for example either making the participants aware of the existence of deepfake content or not. In the study where participants were not aware of deepfakes [42], the success rate was only 3.2%.

It is also tested that people have worse accuracy identifying stimuli that uses foreign languages (languages that participant is not fluent in), people with older age perform worse than younger people and that prior knowledge of synthetic media do not seem to affect the detection performance [12]. Generalized trust, cognitive reflection, holistic thinking and political orientation had an effect on ability to differentiate authentic from fake [20]. Gender, education, income or profession do not

seem to affect deepfake detection [17]. Based on survey data from eight countries, people previously exposed to deepfakes believe in them more often, as well as people who rely on social media for news [43].

2.3.1 Image detection

Online survey of 280 participants [2] researching human face image deepfake detection accuracy achieved overall accuracy only just above chance. What was more concerning was the fact that participants' confidence was not related to their accuracy. This means that people might be overconfident in their deepfake detection ability. Three experimental groups: one with familiarization, one with one-time advice and one with advice and reminders, did not perform significantly differently than the control group [2]. Still, it was noted that specific advice for example advice about asymmetry of earrings, strange clothing fabric, asymmetric glasses and color bleeds from background resulted in correct detection more often [2] emphasizing the importance of good quality advice.

Even though advancement of GANs decreases the detection accuracy in the future as deepfakes will be of higher quality [44], many models still leave human-identifiable artifacts in generated images [14]. Participants of study [44] were suspicious of images especially because of regions containing hair, eyes, and ears. Scene complexity of an image, artifact types within an image, display time of an image and human curation of generated deepfake images impact human image deepfake detection accuracy [14]. It is easier to detect complex scenes with multiple people than simple portraits, as portraits tend to blur the background details [14]. Portraits also often have simpler poses limiting the number of noticeable mistakes, and because authentic portraits often include retouching and editing more than images with complex scenes in the real-world photography, complex scenes are easier to distinguish as authentic or fake [14]. Therefore, images with more details and more complex

settings have more elements that could be erroneous and human-distinguishable [14]. Noticeable errors include anatomical implausibilities, stylistic artifacts, functional implausibilities, violations of physics and sociocultural implausibilities [14].

It can be said that there are different difficulty levels of deepfake image detection, depending on the generation or manipulation method chosen [45]. In study [2], detection accuracy across images ranged between 30% and 85%, supporting the claim that other images are more difficult to distinguish than others [2]. Other factors that can affect the results include age and hyperrealism bias. Age affects image deepfake detection accuracy, as participants in age range of 46 to 65 had worse accuracy [46]. Hyperrealism bias affects human image deepfake detection, as people have higher tendency to pick images of faces made with generative artificial intelligence (GenAI) as authentic because they look realistic visually and humans often believe in what they see [47]. There is not proven to be significant differences between genders in human image deepfake detection accuracy [46].

It is also researched that professional experience do not increase the detection accuracy of human face image deepfakes, and in some cases, specialized professionals were worse detecting deepfakes than people with less or no experience [45]. Another study [48] tests deepfake image detection ability of young and educated college students that could have advantage when detecting deepfakes, but it also achieved low average detection accuracy. These findings can be explained by small sample size and the possibility that professionals are using some kind of internal criteria in addition to basic face perception [45]. However, there are individual participants, who have accuracy reaching 96.30%, but it was not clear what could have caused them to perform exceptionally well [45]. Another study researching face image deepfake detection ability tried to experiment if Super-Recognizers could have better detection accuracy, but no relation was found [49].

One study finds that human detection abilities for image deepfakes beyond hu-

man faces are not much researched [50]. This study focuses on landscapes, architecture and interior, gaining an overall accuracy of 63.7%. According to this study, participants performed better when they were confident most of the time but not always, suggesting that people with more AI experience may have overconfidence bias. Participants used well-documented visual artifacts, like issues with texture, lighting and geometry, even though these cues can be sometimes unreliable, as the images are advancing in quality like stated previously. Participants could still use "too perfect to be real" as a visual cue when doing the detection task.

2.3.2 Audio detection

When researching human audio deepfake detection accuracy, also blind and low vision participants can be included. Two studies [7] and [51] take this approach. Paper [7] claims that blind and low vision individuals have similar detection accuracy than sighted individuals. The other paper, [51], supports this claim by noticing that both blind and sighted individuals struggle with distinguishing authentic audio, with accuracy less than 65%. It also explains differences and similarities between cues used in the detection task: blind people compared traits from screen readers to given samples and sighted humans compared real human voices to the samples [51]. Heuristic cues that can be used when detecting authentic audio from deepfake include speaker's accents, vocal infections, the presence of breathing sounds, lip movement sounds, irregular pausing patterns, audio quality and perceived emotions [51]. Audio recording settings like the room size and recording distance have impact on accuracy with specific samples [51]. Cognitive strain might have also affected the results [51]. In [7] it is stated that participants relied more on cues like inflection, speech imperfections, subjective impression of the speech intensity and speaker's identity, emotion, enunciation and fluency in expression and articulation, than on cues like recording quality.

According to [15], humans can achieve detection accuracy of 73% on average and perform better on detecting authentic samples than fake samples. Most of the participants tried to find some kind of fault in the voice and if they didn't, they believed it was authentic [15]. This means, humans seem to take the reasoning from their experiences and base the detection on cues like prosody, speaking style, speed, disfluency codes of the voice speaking or liveliness and quality of sounds outside the speech [15]. Some participants also relied on their intuition and heard some samples as more human-like and others as more robotic. However, relying on intuition led participants into misclassifying fake samples with authentic features and authentic samples that they thought sounded robotic [15]. Humans could perform well when the deepfake samples included sentence mistakes, odd speed or quality issues [15].

One interesting study [42] experiments human audio deepfake detection accuracy with a cover story. Participants were told that they are testing the user-friendliness of voice messages, but they were actually listening to deepfake audio [42]. None of the participants reacted to the deepfake message and only one admitted noticing something when asked afterwards [42]. After revealing the true experiment, the deepfake voice message was correctly identified by 83.9% of participants [42]. This experiment showed that even though the deepfake was human-distinguishable, not one person reacted to it, emphasizing the fact that deepfake detection in unexpected situations is even more difficult than in lab settings or in situations where people know there can be deepfakes. The study [42] states that "The human ability to detect deepfakes is largely influenced by the fact that people don't think about the voice they are listening to, are used to poor-quality audio conversations, and focus primarily on the content of the message".

2.3.3 Video detection

Humans can use different deepfake identification strategies when they encounter video deepfakes. Study [52] categorizes them into three groups: media-based, knowledge-based and search-based strategies. Media-based strategies include detecting graphical and behavioral anomalies, analyzing production quality, voice inflection and sound quality [52]. Knowledge-based categories include evaluating the message of the video, using their existing knowledge about the actor or the topic or general knowledge on deepfakes [52]. Search-based categories include general search such as Google search and asking others' opinions of the video [52]. Many participants of this study used multiple strategies in detection task [52]. Many of the participants noticed, that while media-based strategies are simpler and faster for them to use, the other strategies that need more cognitive processing are also required to improve their deepfake detection accuracy [52].

For deepfake video detection, humans use combination of facial cues including micro-expression eye movements and lip synchronization and contextual information including background scenery, lighting conditions and audio quality [53]. Cognitive load may also influence individual's detection ability [53]. Being able to correctly identify emotional states of others based on facial expressions improves deepfake video detection accuracy [54]. It was tested if people with autism spectrum disorder (ASD) could therefore perform differently than neurotypical people but no significant differences were found, even though people with ASD were slightly more confident in their judgments [54].

Humans are also better at distinguishing authentic videos from deepfakes, if they are familiar with the person in it [55]. In addition, more time spend on social media and more conspiracy thinking correlate with higher detection accuracy [55]. Analytical thinking and political interest are associated with identifying political deepfake videos [56]. Humans can use video-audio features, like artifacts in lipsync-

ing [57]. Technological glitches can help them detecting fake videos from authentic [56]. Humans could also analyze what the deepfake video is portraying [57]. This means that domain-specific knowledge can enhance the ability to detect deepfake from that specific domain [57]. The language spoken in deepfake video and similar ethnicity of the person in the video as the viewer affected the detection accuracy [53]. It improved, when the viewer was fluent in the language the video used and when the actor has similar ethnic background [53].

Human video deepfake detection ability has also been tested on videos transmitted through noisy channels. This is relevant in cases like video surveillance [58]. The findings show that humans are outperformed by automatic detection methods and the gap is even wider when the videos are distorted [58].

In the future, Electroencephalography (EEG) analysis could be further used in deepfake detection research as a biomarker for deepfake classification [53]. In addition, there has not been much research that includes deepfake videos actually shared on social media [55], meaning research on real-world videos should be made. However, it should be considered that participants could have already seen these videos before, if existing or popular deepfakes are used in research [56].

2.3.4 Text detection

Human deepfake text detection has also been researched, and the accuracy has been compared to automatic detection methods. Internet users with no prior training have difficulties detecting ChatGPT generated text and LLMs seem to be more effective in this task [59].

Text is also important part of a social media profile. Fake profile detection can include generated text, and to determine if the profile is authentic or fake, humans can use cues like grammar errors or intentions behind the text [60]. These clues can also lead humans to incorrectly think that an authentic profile is fake, because

currently state-of-art LLMs are not making as many grammatical errors as human writers [60].

2.3.5 Improvement strategies

Review [13] lists strategies that have been used to improve the human deepfake detection ability. These strategies include: AI support, attentional strategies, feedback training, financial incentives, human collaboration, creating deepfake caricatures and raising awareness. In following sections different strategies are discussed in more detail.

Developing assistive tools

Human deepfake detection accuracy could be improved with developing new assistive technology and tools. One study which discusses possible improvement strategies with older adults [61], presents the concept of Deepfake Detection Glasses, which could detect deepfakes in real-time with the use of AI technology. However, concerns about accessibility, inconvenience and authority were raised [61].

More common approach have been the development of online deepfake detection tools, and they have received positive feedback [62]. Similarly, authors of paper [17] see predictions generated by AI-based models a promising idea to help humans distinguish deepfakes. However, when using these kinds of tools, users might start to rely too much on them even though the tools might not always be trustworthy [62]. When deepfake detection tools are developed, they should be made as reliable as possible and the users should also be trained to use them correctly [62]. It could also help if there could be more explanation of the results the tools give instead of only getting the result claiming the media is authentic or deepfake [62].

One idea of a manual assistive tool is presented in [61]: Deepfake Feature Checklist. This checklist could be an assistant app, that would give features that are

related to the deepfake content and helps identifying deepfake-related features that way [61]. In [5] idea of integrating more advanced deep learning models for example GANs and Autoencoders for video and audio deepfake detection, real-time analysis and multi-platform deployment in the future is presented.

Training

Concept of a Deepfake Identification Practice App is presented in [61]. It could include theoretical explanations, case analyses and simulation practice to teach the user the principles and identification of deepfakes. A training app could embed learning into daily activities more easily than a standalone educational page [61]. Another way to make training more engaging is to use gamification. Study [62] shows how scenario-based role-play could be used to train journalists on deepfakes and improve their verification skills.

One study [46] finds that training on artifact detection to detect false images successfully improved the accuracy rate. However, training participants should be made aware of the capabilities of current deepfake generators, as using human-like features are causing false negatives in deepfake detection experiments [15]. Participants have thought that use of accents and external background noise are only signs of authentic videos, so relying on these cues have led to incorrect classifications [15]. This means also that the training material should be updated continuously and be of high-quality that actually corresponds reality. Similar limitation is noted in [55], where the effectiveness of training involving deepfake exposure, detailed examples of common deepfake video inconsistencies and identification strategies for long term is questioned.

Training the information verification skills could also help with deepfake detection. Many free resources are available online, including video tutorials [63]. Educating students on verification skills could help them see the value of it and make it

more comfortable and less time-consuming [63]. Incorporated and interactive training seemed to be most effective according to [17], and the use of example videos and feedback is emphasized.

Training can help mitigating hyperrealism bias, but it can also lead to people over-identifying deepfakes, meaning humans could start seeing deepfakes in places where in reality everything is authentic [47]. The goal of training is also to develop critical thinking, not only to "create infallible human detectors" [47]. However, study [16] claims that a three-month education period and exposure to content warnings did not have significant effect to prevent unintended re-posting of misinformation, meaning education on critical thinking might not change human behavior on specific scenarios such as during crises. Despite that, future training could include also more guided practice with feedback as well as develop broader contextual analysis skills extending beyond visual features [47].

Raising awareness

One paper [64] explains that after interviewing 36 diverse social media users from rural and urban India, it was found out that not everyone knew that videos can be edited and knew nothing about deepfakes. Some of the interviewed people thought despite knowing of deepfakes that the quality would be low and easily distinguishable [64]. These findings highlight the importance of spreading the awareness about deepfakes as a way to improve human deepfake detection accuracy.

To raise awareness, workshops, games, public events, web-based campaigns and quizzes, satire and visual resources such as infographics and videos [65] as well as critical art [66] could be used to share awareness. Sharing information on specific deepfakes identified for example via mass media and on fact-checker websites could increase awareness [56]. Generally, people should be informed about deepfakes, likely motivations of their creators and known sources sharing deepfakes [56].

Content labels and warnings

As deepfakes are often encountered on social media where they are being shared, one aspect of improving human deepfake detection is to include deepfake labels or warnings on social media. There is research about labels in existing literature, and the presence of labels have had a significant effect on users' beliefs that the post contains AI-generated content, is a deepfake or is edited by AI [67]. Labels could be important aid in detecting deepfakes in the wild, and Meta platforms like Facebook, Instagram, and Threads as well as TikTok are already using AI labels. YouTube is also requiring creators to notify the use of altered or synthetically generated media that seems realistic [67]. However, law does not require for clearly labeling AI-generated content, making it unclear who, developers, users or platforms, is responsible for possible harmful consequences caused by deepfake misinformation [10]. Because of this, malicious actors would probably not label their media as a deepfake. If technical methods are used to detect deepfakes to label them, challenges previously discussed in section 2.2.2 are probably encountered.

Type of a label can impact its effectiveness. Study [68] researches two content labeling strategies, one aimed at clarifying how the content was made including labels such as "AI-generated" and the other aimed at presenting content's potential to mislead including labels such as "Misleading". Results of this study include decreased belief for posts containing labels, but claims that labels did not have an impact on users engagement behaviors [68]. It is also noticed in [68] that large portion of content may evade detection and therefore would not be labeled correctly. At the same time, these posts could gain more credibility, as they are not being labeled as misleading, if people start to trust content labels [68].

Even though users trusted the label, their behaviors did not change much in the experiment in [67]. In another paper, [69], it is found that giving technique tips is superior to giving content warnings, and technique tips are better at improving

human deepfake detection accuracy. This might be due to complexity of deepfakes, and technique tips can better address the complexity [69]. This finding suggests that humans are not able to distinguish authentic content from deepfakes even when they had the motivation to do so and could see that the content is labeled to possibly be a deepfake. That is why, to make labeling effective, the labels should be extremely trustworthy.

Similar to labeling, a standard for audio and visual markers is presented in [70]. This standard defines specific markers that could assist humans in distinguishing communication with a human, a machine, or a combination of both. Study [71] researches misinformation warnings on videos and agrees that the accuracy of the identification and risk assessment of misinformation needs to be improved in the future to make misinformation warnings more effective. It also suggest the use of comments as an alternative way for the users to detect misinformation from videos [71]. Authors of [19] think that relevant context information, such as verification of the source, should be given in addition to misinformation warning.

2.4 Deepfakes and social media

Social media platforms are primary environment where deepfakes are shared and encountered [57]. Therefore, social media as environment is also affecting the human deepfake detection accuracy. As explained previously, many studies have researched deepfake detection accuracy in lab environment, and the results could be different in more realistic environment. In this section the impact of social media and social media platforms are discussed.

As social media has developed, communication has changed. Source has become more difficult to identify as there are multiple layers of message transmission [72]. In addition, social media users have access to system-generated cues, like number of views and re-posts, which have not been used in traditional communication en-

vironments [72]. These are challenges to the field of communication effects [72] and these could also impact deepfake detection as deepfakes are mainly encountered on social media.

In [12] it is hypothesized that human deepfake detection accuracy could be more constrained under social media news feed conditions. This could be explained by online platform environmental factors like divided attention and exposure length [12]. On social media, there are also lot of challenges to information credibility assessment as the online environment is complex [73]. These challenges include anonymity, lack of information quality standards, multiple sources and ambiguous context as well as ease of manipulation and alteration of the content [73]. It has been stated that social media platforms are dynamic and noisy which challenges traditional detection methods [57]. On social media platforms, users need to process large amounts of information at the same time, and information can be irregular and conflicting, which makes determining its credibility difficult [72].

People encounter deepfakes through entertainment and media like music videos, educational videos, or just for entertainment on platforms like YouTube and TikTok [8]. Study [8] finds out that people often have similar reactions to seeing deepfakes including: amazement about advancement of technology, noticing that seeing is no longer believing, predicting dystopian threats of deepfakes, noticing how deepfakes enhance the video experience, causing strong emotional reactions and treating deepfakes as a toy for play. It was also noted that even though deepfakes would be intended to entertain on social media, they still can cause confusion, doubt, and uncertainty depending on individuals awareness of deepfake technology [8].

Study [72] states that "The contextual communication situation is a key differentiator of social media from traditional mass media, thus social factors should be given greater consideration when it comes to evaluating online information". It means that when researching human deepfake detection accuracy, the social context

needs to be considered. This context includes system-generated information such as number of re-posts and likes, but also qualitative cues such as comments and threads of discussion.

2.5 Contextual information

This section presents findings related to contextual information from existing literature. First, it focuses on the various intents of posting deepfakes. Then, cues affecting credibility are discussed and finally it is presented how contextual information have been used in related fields of research.

2.5.1 Intentions behind posting deepfakes

Examples of malicious uses of deepfakes are listed in Chapter 1. On social media, malicious deepfakes are often created and posted by fake profiles, spamming profiles, bot profiles, compromised profiles or cloned profiles [73]. Deepfake posts are then shared using fake followers, direct spam, fake reviews, fake news, fake posts and clickbaits, and can lead to creator's wanted outcome, that can be for example: financial fraud, biasing people perception, overloading negative sentiments, click frauds, malware propagation or spamvertising products [73]. Humans can also share harmful deepfakes by accident or because of ignorance. Paper [67] separates terms misinformation and disinformation. Misinformation refers to spreading inaccurate information without the intent to deceive or manipulate and disinformation has malicious intent behind it [67]. It makes a difference knowing the intent of the deepfake's creator or propagator when encountering a deepfake, as it can help to decide how it should be addressed.

Deepfakes can also be created and posted with good intentions. Film industry is using deepfakes for dubbing, post-processing, making up for a lost voice or for

having an actor that has passed away still star in a movie [4]. Business is also utilizing deepfakes in marketing and advertising. It is possible to create highly personalized ads with deepfake technology, and for example establish an online shop where customers can virtually try on the clothes to see how they look on them [4]. In addition to creating more training data for medical imaging, deepfakes could be used in aesthetic medicine to help patients visualize surgical results before surgeries [74].

Individual social media users can use deepfakes for better privacy, as deepfake obfuscation techniques are seen more effective as traditional methods like blurring, pixelating, masking, and avatars [75]. Deepfakes and AI art can be created for leisure, curiosity or self-expression [76]. In addition, they are used to create design artifacts for work and personal projects [76]. Using deepfakes for self-expression in creative way is, however, a controversial topic. Better approaches that are not seen as problematic could include using only AI tools during the process instead of generating the art piece fully with AI, or making hybrid artistic applications by combining human imagination with AI-generated inspiration [2]. Still, it is not illegal or malicious to create deepfakes for these purposes. Deepfakes can be created for entertainment and humor as well, and people can try to influence others for example with satire, which is different from malicious manipulation.

One interesting use case presented in [77] discusses deepfakes' effects on human memories. People can animate old photographs of deceased family members [77], and it impacts the way how they are remembered. Deepfakes could be used for therapy and to help people process their traumatic experiences by enhancing a photo from difficult period to be more calming [77]. The alteration of memories is possible, because memories are being stored digitally in a form of photos and videos [77]. AI editing tools might be applied automatically to some photos without the user even noticing it, altering users' memories of that captured moment. Similarly,

deepfakes could potentially influence collective memory and historical narratives by manipulating media [77]. Animated photos and deepfakes can also be used in education and research, especially in history, to make reenactments of historical events [34].

Humans can deduce intentions of the creator by combining different contextual information and analyzing them, but there are also studies trying to automate that with AI solutions [78]. This approach has challenges similar to technical detection methods and it is noted that defining the intent is a complex task. Content-based intent analysis that aims to identify the behavioral intention of users has also noticed the complexity and importance of classifying intents of social media posts [79].

2.5.2 Credibility of deepfakes

In this section factors affecting credibility of deepfakes are presented. It is discussed, what makes digital content seem more trustworthy, or more suspicious. This section is important because results gotten from the experiment of this thesis can then be compared with existing literature more easily, and similarities and differences with results can be identified.

Study [80] claims that deepfakes are not effective in making people believe that anyone could say anything, because people can evaluate credibility based on their knowledge of the person presented in the video. If they know what views that person has had before, it lowers the credibility if the same person says something completely opposite in the deepfake. Hyper-realistic deepfakes with plausible content manipulations are seen more credible than deepfakes with implausible content manipulations, but implausible deepfakes have stronger delegitimizing effect [80]. Analytical thinking might enhance resilience toward deepfakes [80]. Study [81] compares how Indian social media users assess credibility of text, audio and video modalities, and what are their intentions for sharing misinformation on WhatsApp. Results show that

video is seen more credible than audio and text, and is more often shared, linking sharing behavior to perceived credibility [81]. It is also found that raising awareness and belief that others would find the shared content interesting were the main reasons for the participants to share the false story [81].

On the contrary, another study claims that "audiovisual disinformation is not perceived as more credible or believable than the same disinformation in textual format" [82]. This study focuses on political cheapfakes and deepfakes, and observers that cheapfakes are perceived more credible than sophisticated deepfakes [82]. Videos chosen in the experiment and its limitations could have affected these results, and it is noticed in [82] that visual disinformation could be perceived more credible if it portrays less known and less politicized topics.

When humans valuate media content, they use cognitive heuristics [83]. Humans need to rely on heuristic cues because of the information overload on social media [72]. Visual modalities are processed more heuristically than text or audio which leads to greater trust and sharing of visual content [83]. This is called principle of "seeing is believing" and means that people think visual content that seems realistic is credible by default. Paper [83] presents new heuristic called synthetic heuristic, that can arise when people see realistic looking content, but then become suspicious of it. This heuristic has developed after awareness of deepfakes and other AI-generated media has been spreading further [83]. However, humans can still fail to activate synthetic heuristic, and be more susceptible to believing in misinformation [83].

Bandwagon heuristic have been confirmed in disinformation research, where findings show that people follow collective opinion of crowds when determining credibility online [72]. When number of followers, re-posts, likes and replies is higher, the perceived believability is higher [72]. Popularity of the video has significant positive effect on the perceived credibility of deepfakes [72]. Having large number

of followers can make people perceive social media influencers as more attractive and trustworthy than those with lower number of followers [72]. High following can suggest high quality content, which is seen more credible than low quality content [72]. Similarly, better technical quality can be considered as effort towards the post, and it has positive impact on its credibility [72], [84]. If the post has lower frame rate or poor information quality, it is seen as less credible [72]. Expertise heuristic could also have an affect on credibility, as people might trust statements from experts more easily [81].

Media richness theory suggests that better description and more information could reduce users' uncertainty and improve information credibility, however, [72] finds that adding titles and descriptions does not actually increase credibility. Overloading content can sometimes decrease the credibility, when cognitive burden for users increases [72]. Study [84] shows that familiarity of the content as a heuristic cue increases trust and perceived credibility, but it did not have significant effect on detection accuracy. Length of a video has an effect in a way, that edited deepfakes are seen as less credible when they are short videos, but the editing has no significant effect in long videos, as suspicious editing traces are not detected as easily in them [72].

Study [85] researches students' trust on deepfakes and factors behind this perceived credibility. Findings of this study show that number of positive comments, discussion between users and explanations of evidence do not affect credibility [85]. Instead, user knowledge, platform and uploader have significant influence on student's trust [85]. If the uploader have many followers and good reputation, students see them as more credible and trustworthy [85]. This could explain why some people might want to trust deepfake post even with comments claiming it is fake, if the uploader is perceived credible and trustworthy. However, this study uses only survey asking about hypothetical situations, and do not implement experiment where

it could be tested if any of these cues affect students' deepfake detection accuracy.

2.5.3 Use of contextual information in related fields of research

This section presents how contextual information has been used already in existing literature. Even though not many works research deepfakes and contextual information, many studies from related fields notice importance of context and make findings that could apply to human deepfake detection as well. These findings are next presented.

Rumor detection

Study [86] is about rumor detection, but is similar to deepfake detection as deepfakes can also be multimodal and spread on social media. Often, social media posts contain text, images and videos. It requires evaluation of each modality and the combination of them to assess the credibility [86]. Understanding the combination of different modalities can be hard for current technical solutions, but natural for humans. This observation could suggest that comparing information from these different modalities of a social media post could help humans in detecting deepfakes as well. Rumor detection research also shows that rumors and non-rumors spread differently on social media, with different patterns [86]. Study [86] lists propagation-based evidence including post statistics such as publication date, hashtag and URL, user demographics such as age, gender, location, and education, network structure, and user reactions such as number of re-posts or likes. Comments including re-posts and replies could also be used [86]. There could be different patterns based on these indicators in authentic and deepfake media as well. Paper [86] notes that it is important to consider also the state of the world when trying to identify misinformation, and that metadata should be complementary to textual or visual information as it

is not sufficient to use it alone. In addition, it must be remembered that previously mentioned indicators can change over time, meaning they can be different just after creation of the post and after the post has been gaining more popularity.

Journalism

Journalists face misinformation and their work can be affected by deepfakes. There are contextual verification steps that journalists usually take [62]. These steps include regular steps: contacting the sender, searching Google for context, looking for other news reports, analyzing original poster's account, contacting the original author, analyzing re-posters' accounts and searching social media for context. There are also technical steps: checking metadata, using deepfake detection tools, manually analyzing videos and requesting biometric analysis. In an experiment in [62] some participants told they did not need to use any deepfake detection tools, because they already had come to a conclusion from various contextual steps. These steps could be useful also in human deepfake detection, because deepfakes are always encountered in some kind of context. One journalist states that there needs to be various layers of evidence to get diverse view on the content that is being inspected [62].

Some findings from the field of journalism could be applied to deepfake detection research, as they can both be related to misinformation and multimodal media [63]. Even though fake videos are becoming more realistic and harder to detect by human eyes, it can be said that focusing on the provenance or origin of the video can instead be used as a means of verification [63]. Determining the source of the deepfake, and questioning its motivations and incentives should be topics taught to students [63] and should be considered when designing training for other groups as well.

Social media moderation

Another field where misinformation and deepfakes need to be detected is social media content moderation. It often relies on machine learning to detect deepfakes automatically, but also manual and moderator-based approaches are used. Study [60] observes that many variables affected the decision when humans were told to classify a social media account as authentic or fake. These variables include expectations of the content usually posted on the specific platform, career expectations, identity-based stereotypes, personal experiences, thinking that a profile is too good to be true or too bad to be true, anthropomorphized views of AI such as thinking bland and emotionless profiles are more likely fake, thinking higher status profiles as well as vague profiles are more suspicious and that "phishy" or "scammy" behaviors are suspicious [60]. It was also noticed how participants tend to think that as machine learning algorithms are struggling with representing certain identities due to bias in their datasets, it will make profiles of people in these groups poorly [60].

Social media analysis

A work on social media analysis [87] notices the value of contextual information, especially networked context when deciding check-worthiness of a tweet. There is contextual information embedded in user reactions and discussion, that could include comments, shares, reactions and discussion threads [87]. Findings of this paper suggest that qualitative context information such as comments improves detection accuracy for check-worthy tweets, possibly because it is more informative than quantitative engagement metrics such as number of likes [87].

Fake X (previously Twitter) profiles have been a topic for research as well, including the use of AI-generated image as profile pictures. Deepfakes can be used as profile pictures too, so exploring typical contextual cues related to them is of interest. Paper [11] researches this topic. It is said that accounts with fake images often

have low social engagement and small number of followers and followed accounts, but there can also be accounts that are very active, possibly because of spamming [11]. Accounts using fake images have approximately 393 followers on X while the authentic profiles have average of 5086 followers. Many fake accounts had under 9 followers and many of them had exactly 0 followers [11]. Fake accounts follow approximately 283 accounts and authentic accounts follow 760 other accounts [11]. Very few fake accounts follow zero accounts, but many of them follow exactly two accounts [11]. This finding could suggest that fake accounts can be used for coordinated behavior [11]. Fake-image accounts are posting less than authentic accounts on average, 0.95 vs. 3.68 tweets per day, but there are fake accounts that are posting more than 50 tweets per day which is exceptionally many compared to authentic accounts [11].

Another important detail is the account creation time, as many fake accounts that are used for malicious purposes might be created in bulk, and are often quite new [11]. Many fake accounts have been created in the current year and they have a high probability to be suspended in the near future. Many of the accounts were spamming content similar to each other, often about topics like giveaways, cryptocurrencies, and pornography as well as political topics like the war in Ukraine, debates on COVID-19 and vaccinations, and election-related discourse [11].

Digital forensics

Digital forensics can be used when gathering metadata of a social media post. More contextual information can be gathered that way, in addition to technical and social network information that is visible on the platform. Image metadata that could be found with digital forensics include GPS coordinates, city, state, country, sublocation, GPS timezone offset, GPS timestamp, GPS date stamp, exposure time, title, image caption, headline, image description and content description [88]. However,

it should be remembered that often social media platforms remove public access to this kind of metadata [88] for users' privacy.

Fact-checking

Fact checkers need to work with fake news and distinguish misinformation from real [18] so they need to be able to also distinguish deepfakes in their work. Their verification tools include reverse image search, metadata analysis, fact-checking databases, and a novel verified news processor that extracts spatial, temporal, attribution, and motivational context [18]. Contextual information used are captions, descriptions, social media posts, news articles, and available metadata [18]. The purpose of the tool is to extract four source details: spatial context, temporal context, attribution context identifying published users and reporters in news, and motivational context [18]. In this way, enough background information is gathered. These approaches could also work in human deepfake detection.

3 Research methodology

Systematic literature review is chosen to identify, select and synthesize relevant literature transparently and reproducibly and create a strong base for the contextual cue framework, the Intent Labeling concept and the design of the survey-based experimental study. In addition to systematic literature review, a survey-based experimental study will be conducted. Quantitative data of deepfake detection accuracy of the students as well as qualitative data of the cues and strategies used are gathered. Proposed concept of Intent Labeling will be evaluated with the survey.

3.1 Systematic literature review

Sources are gathered from IEEE Xplore, ACM Digital Library, Oxford Academic, Sage journals, DOAJ: Directory of Open Access Journals, ASIS&T Digital Library, ScienceDirect and Taylor & Francis Online which are accessible through the University of Turku institutional access. Keywords "deep fake", deepfake, "synthetic media" combined with detection or "detection accuracy" are used to find sources for the current state of the deepfake detection, both technical and human detection. Search query "(deepfake* OR "deep fake*" OR “synthetic media”) AND (purpose* OR motiv* OR intent*) AND "social media"" is used to gather data about contextual information. Search queries were modified depending on the academic database, so each query used to find sources for this thesis are presented in Table 3.1. Details given in Table 3.1 ensure reproducibility of this systematic literature review.

Table 3.1: Search query specifications

Database name	**Search query**	**Search date**	**Restrictions**	**Number of results**
IEEE Xplore	(deepfake OR "deep fake" OR "synthetic media") AND "detection accuracy" AND review	1.12.2025	years 2023-2025	23
IEEE Xplore	(deepfake OR "deep fake" OR "synthetic media") AND human	1.12.2025	years 2023-2025	544
IEEE Xplore	(deepfake OR "deep fake" OR “synthetic media”) AND (purpose* OR motiv* OR intent*) AND "social media"	1.12.2025	none	118
ACM Digital Library	("deep fake" OR deepfake OR "synthetic media") AND detection AND review	3.12.2025	years 2023-2025	806
ACM Digital Library	("deep fake" OR deepfake OR "synthetic media") AND human	3.12.2025	years 2023-2025	1149
ACM Digital Library	(deepfake OR "deep fake" OR “synthetic media”) AND (purpose* OR motiv* OR intent*) AND "social media"	3.12.2025	none	790
Oxford Academic	("deep fake" OR deepfake)	7.12.2025	none	1002
DOAJ: Directory of Open Access Journals	deepfake AND review	7.12.2025	years 2023-2025, “technology”	21
DOAJ: Directory of Open Access Journals	"deep fake" AND review	7.12.2025	“technology”	2
ASIS&T Digital Library	deepfake OR "deep fake" OR "synthetic media"	7.12.2025	none	42
ScienceDirect	(deepfake OR "deep fake" OR "synthetic media") AND review	7.12.2025	years 2023-2026, “review articles”	527

Database name	**Search query**	**Search date**	**Restrictions**	**Number of results**
ScienceDirect	(deepfake OR "deep fake" OR "synthetic media") AND "human detection"	7.12.2025	years 2023-2026	16
ScienceDirect	(deepfake OR "deep fake" OR “synthetic media”) AND (purpose OR motivation OR intent) AND "social media"	7.12.2025	none	599
Sage journals	("deep fake" OR deepfake OR "synthetic media") AND detection AND review	8.12.2025	years 2023-2026	350
Sage journals	(deepfake* OR "deep fake*" OR “synthetic media”) AND (purpose* OR motiv* OR intent*) AND "social media"	8.12.2025	none	449
Taylor & Francis Online	deepfake OR "deep fake" OR "synthetic media"	22.12.2025	none	813

Total of 7251 papers were found and 187 advanced to the next phase based on the title and abstract. Papers that were out of scope, were duplicates or had too similar findings were excluded during this phase. After reading the full papers, 88 papers were included in this thesis as seen in Figure 3.1. Exclusion criteria for the second phase included: paper did not provide new relevant information to the thesis, the paper presented too detailed and specific area of the topic, the paper was not in the scope of the thesis, the information was not relevant for the thesis, the study was incomplete or the access to the full paper was blocked. Many articles that were out of scope, were focusing too much on a specific topic area that was not relevant for this thesis, such as fake news or one specific technical deepfake detection method.

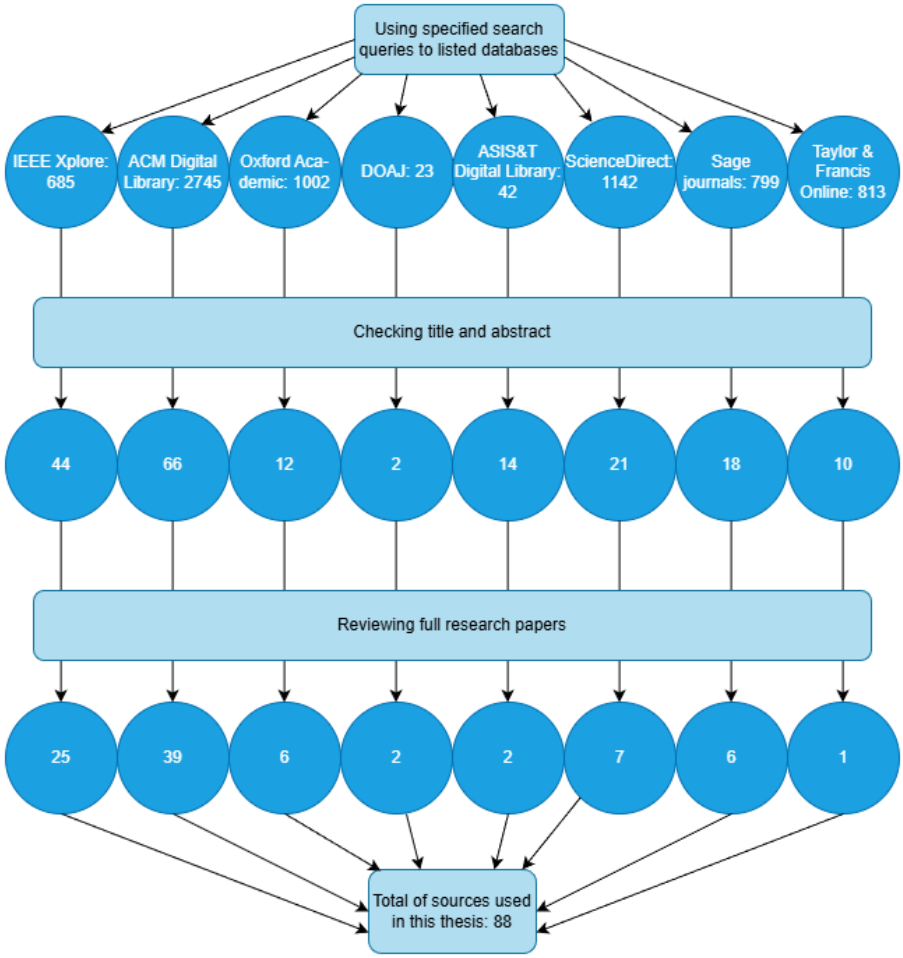

Figure 3.1: Simplified systematic literature review process of this thesis

Selected papers are used to explain the background, current state-of-art deepfake detection methods, findings from human deepfake detection, social media as an environment for deepfake detection and contextual information. Further content analysis will be made to answer research questions one and two.

3.2 Survey-based experimental study

The experiment uses a mixed methods approach. Pragmatic mixed methods approach is adapted and it is tested if contextual information has an impact on human deepfake detection accuracy. Images are chosen for the experimental study as the image detection accuracy is currently worse than videos and because of limited resources of the researcher to gain enough high quality deepfake videos for the study. Online survey is chosen as the research method to get quantitative data from the participants with image deepfake detection task. This data includes number of correctly/incorrectly labeled images, confidence rating, interpreted maliciousness rating of an image and the overall time taken to answer the survey.

Descriptive analysis is applied to quantitative data. Quantitative results of the survey-based experimental study are analyzed with repeated measures ANOVA as the design is within-subjects design and includes measuring accuracy, confidence and maliciousness in different conditions and for different contextual information groups. For analyzing accuracy, independent variable is the condition (images without any context, images with social media context and images with social media context and Intent Label) and dependent variable is deepfake detection accuracy. This analysis will show if there are significant differences between these conditions while considering that measured accuracies are from same individuals in different conditions. Accuracy is also analyzed when independent variable is the context group (authentic images with trustworthy context, deepfakes with trustworthy context, deepfakes with suspicious context and authentic images with suspicious context),

and dependent variable is deepfake detection accuracy to see if there are significant differences between different context groups. Repeated measures ANOVA is then used to test interactions between these two previously mentioned independent variables and deepfake detection accuracy as a dependent variable. Confidence is analyzed similarly with repeated measures ANOVA in different conditions, condition as independent variable and confidence rating as dependent variable. Correlation is tested between perceived confidence and deepfake detection accuracy with Pearson correlation coefficient and Spearman's rank correlation coefficient.

Impact includes increasing or decreasing accuracy but it will be also important finding if it could be observed that contextual information could lead to faster or easier detection for the participants. It would be valuable to know what cues and strategies the students are using and what cues have made them change their choices, to further analyze the impact of contextual cues. That is why also qualitative data will be gathered with open questions asking what cues were used in the detection task. Reasoning is asked only after showing each version of one image and giving it is optional for the participant to reduce cognitive overload. Reasoning for answers could give new ideas or support findings made in other research papers. At the end of the survey it will be important to get insights of how students felt about Intent Labeling. Similar study [48] did quantitative analysis across some dimensions and a qualitative discussion evaluating other topics in addition, as understanding the though process while making decisions is of interest. In this thesis, understanding the thought process of the students could lead to important findings. Thematic analysis of qualitative data is conducted and both quantitative analysis and thematic analysis are used to answer research questions three and four.

Non-probability sampling is chosen in this thesis. Students taking course DTEK2029 Human Element in Information Security are chosen as participants of the experiment because of convenience. Sample size is N = 73. In [85] it is stated that "Students

have a good literacy level regarding deepfakes and often use social media, so they feel they are familiar with deepfake videos on social media" meaning that students are suitable for this study acquiring also qualitative data from the participants. Even though students are part of people that possibly use social media and therefore are part of the population of this experiment, the findings will not be generalizable. The goal of this study is further explore the topic area of human deepfake detection, and for that students are suitable as participants as they have the knowledge to analyze their answers further than any participant with no experience or previous knowledge on deepfakes. Experiment uses within-subjects design, so each student will fill the survey that is testing different conditions to limit the influence of human differences.

As mentioned, online survey will be the data collection method of the experimental study. Survey is chosen for convenience, as it will be easier for participants to do the detection task in their own time rather than having them to take the survey in class or specific experiment setup. However, this means that their answers do not have validation from outside as they are not being in controlled or monitored environment during the experiment. To encourage students to not use AI or other tools when doing the survey, they are given instructions stating that the amount of correct answers is not affecting their grade, but they should do their best in evaluating the images without any tools. The chosen survey platform should be convenient and easy to use for the students so they can focus on doing their best in the experiment to ensure reliable results.

This survey-based experimental study informs participants of its purpose, ensures anonymity of the participants and their answers and safe handling of participant data. Personally identifiable information, name of participant, is only collected for the purpose of the course the survey is implemented in, and all personally identifiable information is removed before the results are forwarded to be analyzed in this thesis work.

4 Survey-based experimental study design

This chapter summarizes relevant findings from systematic literature review that are used in both dataset and survey creation. Categorization of contextual cues, analysis of existing human deepfake detection improvement strategies and a new improvement strategy concept are presented. Dataset creation process for the survey and the survey design are explained.

4.1 Background for the survey-based experimental study design

This section analyzes the main findings from the literature review. First, different contextual cues are collected and categorized. Then, effectiveness of different human deepfake detection strategies are analyzed. Finally, a new human deepfake detection improvement strategy, that could combine the use of contextual information and ideas from already existing improvement strategies, is proposed.

4.1.1 Categorizing contextual information

In this thesis, contextual information is seen as broadly as possible but limited to social media environment as deepfake images are commonly encountered on social media. Contextual information considered in this thesis includes context of what the images are portraying, when, where and why they are posted as well as interactions and network context of the original source. This section answers research question one. Based on the systematic literature review, different contextual cues were gathered. All examples of different contextual cues mentioned were first listed and grouped together based on which part of a social media post they are related to (post itself, platform, poster, interactions, viewer or combination of these aspects), leading to seven unique categories: perceived quality, external knowledge, source and cross-references, intent, networked context, platform context and metadata. These categories are tightly related to each other and are explained in more detail in Table 4.1.

Table 4.1: Contextual cues found in the systematic literature review

Contextual cue category	**Explanation**	**Papers mentioning cues in this category**
Perceived quality	Observing post's visual, audio and/or textual modalities could help detecting deepfakes. Analyzing visual and audio artifacts, grammar and language used in the post could help assessing the credibility of a post when compared to viewer's expectations, experiences and knowledge.	[2], [7], [8], [11], [12], [13], [14], [15], [16], [17], [36], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [55], [56], [57], [59], [60], [61], [62], [64], [65], [68], [69], [72], [73], [80], [81], [82], [83], [84], [85], [87], [88]

Contextual cue category	**Explanation**	**Papers mentioning cues in this category**
External knowledge	Knowledge and familiarity with the topic of a post as well as knowledge on deepfakes could increase detection accuracy. Knowing about current state of world, history and cultural norms can help analyzing credibility.	[2], [7], [8], [12], [13], [14], [15], [17], [18], [19], [41], [42], [43], [45], [46], [48], [49], [51], [52], [55], [56], [57], [59], [60], [61], [62], [63], [64], [65], [67], [68], [69], [71], [73], [80], [81], [82], [83], [84], [85], [88]
Source and cross-references	Knowing who is the original creator of a post and information presented in it could help analyzing its credibility. Further analyzing creator's and re-posters' accounts information such as followers and followed accounts and their numbers, account activity, number of posts, posts' topics, account's social engagement and account creation time could improve detection accuracy.	[11], [12], [16], [17], [18], [19], [52], [57], [60], [62], [63], [64], [65], [67], [71], [72], [73], [80], [82], [84], [85], [86], [87], [88]
Intent	Knowing why specific content was posted or shared could impact detection accuracy. It includes analyzing content of the message, motivation and incentives of a post.	[7], [8], [11], [16], [18], [42], [51], [52], [57], [60], [61], [63], [64], [70], [71], [73], [78], [85]

Contextual cue category	**Explanation**	**Papers mentioning cues in this category**
Platform context	Platform expectations and familiarity with the environment could impact detection accuracy. System generated cues such as number of views, re-posts, likes and replies as well as publication date could help assessing credibility of a post. Labels could help detecting deepfakes. Knowing context of how a post was encountered, for example when browsing social media feed or from direct messages, could also impact detection accuracy.	[8], [11], [12], [17], [42], [43], [50], [52], [55], [57], [60], [61], [62], [63], [64], [67], [68], [69], [70], [71], [72], [73], [81], [82], [83], [85], [86], [87]
Networked context	User reactions, discussion threads, comments and shares could help assessing the credibility of a post. These cues help assessing account's social engagement and activity.	[11], [12], [57], [60], [64], [67], [72], [73], [85], [86], [87], [88]
Metadata	Location details of a post such as GPS coordinates, city, state, country and sublocation could be helpful when combined with other contextual information. Similarly, timing related cues such as GPS timezone offset, GPS timestamp, GPS date stamp and exposure time could be helpful when combined with other contextual information. Title, caption, headline and content description could help assessing credibility of a post.	[11], [12], [18], [57], [62], [63], [70], [72], [73], [86], [88]

The following two lists summarize findings from the literature review about features that can make contextual information look either trustworthy or malicious. First list has contextual cues that are often perceived as trustworthy. The second list includes suspicious contextual cues that are often related to malicious intent.

Contextual information related to trustworthiness

- Content corresponds to viewer's expectations.
- Content has no human distinguishable visual or audio artifacts and the technical quality is high.
- Content is familiar to viewer.
- Content corresponds to viewer's previous knowledge.
- The uploader of the content is popular, has high number of followers and has good reputation.
- The uploader is posting about topics that are less known and less politicized.
- The post has high number of likes, re-posts and replies.
- Content corresponds to viewer's expectations of the platform.

Contextual information related to malicious intent

- Content do not correspond to viewer's expectations.
- The content has low audio quality and/or low resolution.
- Content do not correspond to viewer's previous knowledge and makes it seem implausible.
- The uploader has low number of followers.
- The uploader has exceptionally high account activity and posting frequency.

- The uploader is posting about topics such as giveaways, cryptocurrencies, pornography, the war in Ukraine, debates on COVID-19 and vaccinations, and election-related discourse.
- Account is newly created.
- The post has low number of likes, re-posts and replies.
- Content do not correspond to viewer's expectations of the platform.

4.1.2 Effectiveness analysis of improvement strategies

This section analyzes the effectiveness of improvement strategies presented in 2.3.5. Therefore, it answers research question two. Analyzing strengths and challenges of existing strategies is critical when developing a new improvement strategy.

Assistive tools can be very effective, but they rely on robustness and reliability of the underlying technology. As technical deepfake detection solutions are not yet performing in sufficient accuracy in real-world environments, people should not fully trust the assistive tools, which lowers their effectiveness. These tools should also be updated continuously as deepfakes are also constantly developing, which could be seen as a limitation. This strategy is therefore promising, but not yet the most effective.

Training and raising awareness are similar and training should also raise awareness to be more effective, even though awareness can be raised in other ways too. Training has similar limitation as assistive tools; it needs to be continuously updated. Training is only effective if it is of good quality and up to date. It is also effective only if it is engaging and users have motivation for it. It is difficult to evaluate effectiveness for training, as it highly depends on what material is used and what methods are used. Raising awareness has similar limits, but it has potential to make people think and question more what they see when they use social media,

which in itself can be already quite effective. Training has many possibilities how it could be used to improve human deepfake detection.

Content labels and warnings can help reduce cognitive load on social media when there is a lot of information available. The idea of content labels is very effective, but the implementation relies on existing deepfake detection methods, if the creators are not being trustworthy. This is most probably not the case, as the malicious actors do not want to follow content labeling rules anyway if they want to use deepfakes for malicious purposes.

Regarding all improvement strategies discussed, limited research has been made on how older adults detect deepfakes [61]. Challenges like cognitive decline in memory, decision-making, and vision [61] should be also considered more when designing improvement strategies. Assistive tools, content label designs and contents of training should be designed in a way that they are accessible and suitable for everyone. These strategies could also be combined to make them more effective, for example developing a deepfake detection tool and giving its users training of how they can use also their own judgment in addition to just trusting the result given by the tool.

4.1.3 A concept for new improvement strategy using contextual information

In this section, a new improvement strategy is proposed. Based on the systematic literature review, multiple important findings can be concluded. Firstly, current technical deepfake detection methods do not yet achieve sufficient accuracy when they are used on social media. Secondly, human deepfake detection accuracy is at level of chance when no contextual information is used, highlighting the fact that deepfakes are currently very realistic and difficult to distinguish from authentic media. Thirdly, it has been noted that cognitive strain might have affected the detection ability and humans need to process large amounts of information especially

on social media. Fourthly, human deepfake detection accuracy becomes even lower when deepfakes are encountered unexpectedly, and humans tend to focus more on what message the content wants to convey instead of how it is presented. Fifthly, spread of misinformation and disinformation is one of the main threats of deepfakes, and the goal of the creator and propagators affects how the deepfake should be addressed.

Based on these findings, it is noticed that the intent behind the deepfake matters. If humans can't distinguish between authentic media and deepfake by only observing visual, audio or audio-visual artifacts, the focus could be shifted to observing what is the purpose of the post. If humans could focus critically only on posts that possibly have malicious intent, it would be safe to pass deepfakes made with good intentions and avoid cognitive overload. However, to know the intent, humans need to use their cognitive resources and analyze a lot of information, including broad contextual information previously discussed. This leads to the idea of this new improvement strategy, which could solve this problem: Intent labeling.

Similarly to content labeling strategy, Intent Labeling could be made with technical solutions, especially with AI like is explored in [78]. It could be possible to label the intent for example into categories like Humor/Satire, Art, or Misinformation in the future, but current architectures and datasets are not enough for reliable intent-aware classification [78]. Intent can also be seen as in the field of intent analysis, in domain of natural language understanding, where intent means purpose for action and can be inferred from the text [79]. This research also sees the limited amount of labeled datasets as a limitation, and states that contextual knowledge is needed for it [79].

Intent categories for Intent Labeling

For this section, intentions of posting to social media are gathered by observing social media content as well as using information gathered in section 2.5.1. After observing posts on various social media platforms, it was noticed that most of the posts were posted for gaining visibility, financial gain, connecting with others via sharing lifestyle, ideas or interests or to inform other users and share awareness and news. Malicious intents include financial fraud, biasing people perception, overloading negative sentiments, click frauds, malware propagation or spamvertising products as stated previously. These malicious intents therefore have similar goals as posting in general: advertising, getting visibility and influencing people. Finally, ten Intent Label categories are identified in this thesis: **Creativity, Entertainment, Humor, Lifestyle, Promotion, Advertisement, Informing, News, Connecting and Networking**. Creativity is for wanting to share art, creative hobbies or creative work and expressing oneself. Entertainment includes sharing for example sport or celebrity videos for the entertainment of other users. Humor includes memes and posts created to make people laugh and can also include satire. Lifestyle is for sharing own life with others and can include topics like travel, wellness and home interior. Promotion is for users that want to gain visibility for something, and that visibility could lead to purchasing some product or service or start a new hobby for example. Advertisement is for selling something specific. Informing is to share awareness and start discussions influencing opinions of other people. News is to share what is currently happening in some part of the world. Connecting is to share updates of own life for followers, friends and family. Networking is to promote oneself and gain new connections with new opportunities in mind. Each social media post should be able to be put into one of these categories based on the intent of the post's creator.

Strengths and challenges of Intent Labeling concept

To avoid challenges that technical solutions have, Intent Labeling could be integrated into the social media platforms. If it was mandatory for the creator of the post to label the post into some of the predefined categories, it could be shown to users when they are viewing the post. This would not be much work in addition to writing captions and creating the post for users with good intentions, but it could be more challenging for people posting deepfakes with malicious intent, because they now would need to target specific post category. If deepfake targeted for spreading misinformation is labeled as Entertainment, people could choose to not take the information it presents seriously. If it would instead be posted to News category, people would already have more critical attitude towards posts of this category, and they could be able to detect the deepfake more easily. Similarly, if a fake product was posted on Creativity and advertised, it would be suspicious because Advertising is its own category. However, if the same post was posted on Advertising, people could again view these posts more carefully. Specific intent categories could make humans more aware of possibility to encounter malicious deepfakes and help to activate synthetic heuristic in situations when not correctly detecting deepfake can lead to harmful outcomes. In [71] one participant explained ignoring content labels because "so many videos are flagged with it, even if it's not news". This highlights the need to separate different social media post categories, so the users could use their cognitive resources only for posts that need the most critical thinking.

However, it can be questioned if this approach is only trying to dodge the problem of human deepfake detection and not improving it. That is why it should be experimented how humans are using contextual information such as intent behind the post to make their decisions, and what information leads to correct classification. Misleading contextual information should also be explored. The survey-based experimental study will test Intent Labeling and examine if knowing the intent

could have an impact on detection accuracy or if students could make their decisions faster with help of these labels, reducing the cognitive overload and thereby improving their detection accuracy.

4.2 Dataset creation

Deepfake images are chosen as the stimuli for the experiment, because prior research reports lower human detection accuracy for images and because image-based, posts of visual modality, are common on social media. Existing studies have shown worse human deepfake detection accuracy for images and text than for audio and video, which means researching whether contextual information could improve image deepfake detection is meaningful.

When gathering the dataset for this experiment, it must be remembered that even the authentic images do not fully represent reality [21]. It is common that authentic images posted on social media are also edited to some extent, for example with adjusting contrast, brightness and enhancing blurry images [22]. When choosing the images to include in experiment, they should be as realistic as possible as well as diverse: "In real-world scenarios, deepfakes are often more sophisticated, featuring occlusions, side profiles, background noise, varied lighting conditions, cut scene changes, multiple people, and adversarial manipulations designed to evade detection systems" [57]. If the samples used in the experiment are not of high-quality, they are not corresponding to real world [36] and are not meaningful for the experiment. This also means that people might have to use something else in addition to the deepfake image to detect them, if the images are visually human-indistinguishable.

4.2.1 Collecting authentic images and deepfakes

To gather authentic and deepfake images corresponding to these conditions, OpenFake dataset [89] is used. It is easily accessible dataset, with 3 million real images and 963000 synthetic deepfake images relevant for real-world social media context [89]. Deepfakes in this dataset are generated with state-of-the-art open-source and proprietary models [89]. From this dataset it is possible to manually gather a subset of authentic and fake images for this experiment, and gain realistic context for them from the given descriptions or prompts as well as with reverse image search. The dataset labels each image as real or fake, gives the model name, creation date and the description or prompt for each image. OpenFake [89] has CC BY-NC-SA 4.0 license, and it is based on LAION-400M dataset [1] which states: "We distribute the metadata dataset (the parquet files) under the most open Creative Common CC-BY 4.0 license, which poses no particular restriction. The images are under their copyright".

Manual inspection should be used for generated images to ensure the high quality [36]. First one thousand images from OpenFake are inspected manually to construct the subset for the experiment, and authentic images as well as deepfakes are categorized into each intent category (Creativity, Entertainment, Humor, Lifestyle, Promotion, Advertisement, Informing, News, Connecting and Networking). If an image could correspond to multiple categories the category is selected randomly. In the inspection phase, exclusion criteria for authentic images includes: image has too low resolution for meaningful detection, image contains clear copyrighted logo, image has watermark, image contains sensitive or improper material or the description of the image is not matching to the image. Exclusion criteria for the deepfake images includes: image has clear artifacts, image has too low resolution for meaningful detection, image contains clear copyrighted logo, image contains sensitive or

[1] https://laion.ai/blog/laion-400-open-dataset/

improper material or the prompt is not corresponding to generated image. After going through first 1000 images in OpenFake dataset manually, 235 authentic images and 106 deepfakes were chosen. Because OpenFake is a diverse dataset and it includes images with varying resolution, complex and simple scenes, images portraying objects, landscapes and humans of different ages, genders and ethnicity, also the subset should be sufficiently diverse and corresponding to real world. Selecting images randomly from that pool of images will then result in diverse selection of images.

4.2.2 Image quality assessment

Before adding contextual information to the images, they are processed to make the dataset more coherent and ensure the high quality. Watermarks generated with AI, for example Grok's watermarks, were cropped out because it would be also easy to do for people posting deepfakes with malicious intent. All images were then cropped into 1:1 ratio to imitate social media environment where the images often need to be of specific aspect ratio, for example on Instagram or TikTok. Cropping was made to preserve as much of the image as possible and include the most meaningful part of the image for the intent. To assess the quality of fake images further, a survey was made asking assessment of the chosen deepfake images. The task was to rate the quality from 1 to 5 for each image, with no context. Instructions given in the survey were: "Assess the quality of each image. Quality means how believable and credible the image is, it is not about the technical resolution. Do not compare the images only to actual reality, but to human made content that can be seen on social media, as there are creative images included in these deepfakes. 1 = very bad quality/very clearly AI, 5 = very good quality/hard to tell apart from authentic." This survey was sent to people with experience working with AI and AI generated images through connections, and two people answered the survey. The average

quality score of all deepfakes was 3.06. Images that achieved score under 3 and answers were not widely split, were excluded to ensure high quality deepfake images in the final dataset. In addition, images that had widely split opinions were again inspected by the researcher, and if clear artifacts were found, also these images were excluded from the final dataset. From 106 deepfakes, 33 were excluded after this quality assessment.

4.2.3 Selecting images for the survey

In the next phase, 20 images were selected from this pool of 235 authentic and 73 deepfake images. In the final selection, one authentic and one deepfake were selected from each intent category to ensure variety of images. First, each intent category were randomly allocated into one of these conditions: authentic image has trustworthy context and deepfake has suspicious context or authentic image has suspicious context and deepfake has trustworthy context. Half of the intent categories were allocated to the first group and half were allocated to second group, so it can be analyzed whether giving context has any impact to detection accuracy: positive or negative. Then, images were selected randomly from each intent category until image corresponding to the given criteria was found. Because perceived quality is also context, and familiarity, topic and expectations of the viewer affect perceived trustworthiness there exists criteria for selecting images. Images for trustworthy context should be of high resolution (over size 700 x 700 pixels), not have political topic and correspond to expectations of the intent category. Images with suspicious context should therefore be of low resolution, can have political topic and are not corresponding to expectations of the intent category.

4.2.4 Creating context for the selected images

To add contextual information, template seen in Figure 4.1 is used. In the template, every contextual information category from Section 4.1.1 is included. The image itself with the description allows the use of perceived quality and external knowledge. To analyze source, numbers of followers, followed, posts and account creation date are given. Intent can be deducted from the image, description and Intent Label. Intent Label was designed to seem like part of contextual information and not as part of the image and to be big enough for people to notice it. Number of likes, comments and shares are given for platform context. Example comments are given as networked context. Image publication date is given as metadata. Account names, profile pictures and biographies were not used to not accidentally make them impersonate any real account.

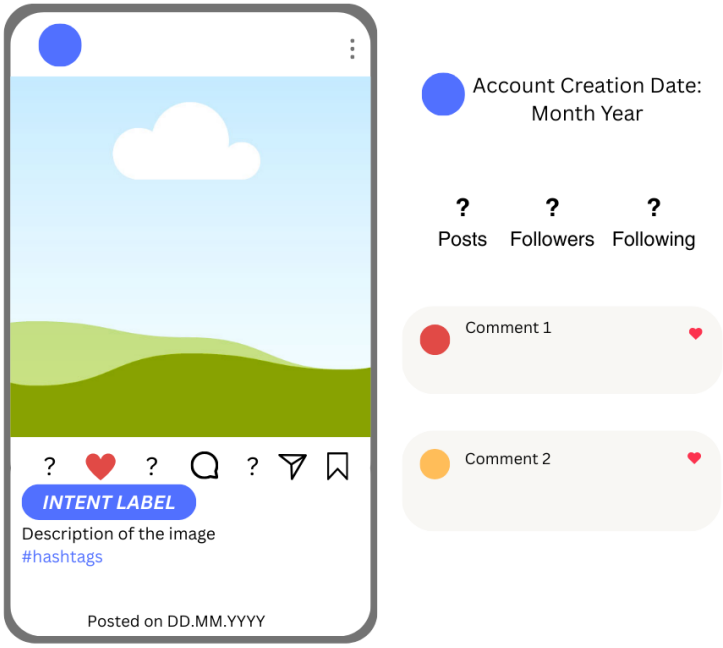

Figure 4.1: Template for the image with contextual information and Intent Label

Findings from Section 4.1.1 were used to fill the template for each image. InstaFake dataset[2] was used for follower, followed, total posts, like, comment and

[2]https://github.com/fcakyon/instafake-dataset

hashtag numbers. Real social media accounts and posts related to similar images and topics were also analyzed on Instagram. Descriptions and comments were written with the help of ChatGPT. Outputs were manually checked and curated. Base prompt that was used for every image was: "Write a social media post description for image described as: [description/prompt from the OpenFake dataset]. The meaning of the post is [explain the intent category/hint for malicious intent]. Write it in [trustworthy/kind/professional etc. manner]. Make it short. Include [number] hashtags." After this, more detailed instructions for example "Make it shorter" were given to achieve better result. For comments, prompt "Write comments for this post. The comments should seem like they are written by [humans/bots]" was used. After that, instructions such as "Include emojis" were given for more realistic comments. Real descriptions and comments on Instagram were also analyzed to get contextual information corresponding to real world social media posts. The final dataset of images used for the study is publicly available in Gitlab[3].

4.3 Survey design

After creating the dataset, a survey was made on Webropol. Webropol was chosen because it is familiar for students and has diverse settings for making a survey. Using familiar platform with clear instructions should not add to students' cognitive overload despite having a total of 228 questions in the experiment survey. Questions were chosen to be mostly multiple choice questions instead of open questions and open questions were given as optional, to reduce the cognitive load as much as possible but still getting enough data to analyze and get results.

Instructions were given at the start of the survey, stating: "The purpose of this survey is to gather data about human deepfake detection accuracy. The experiment consists of 20 sets of three images. You need to assess each image as authentic

[3]https://gitlab.utu.fi/siahei/deepfake-image-detection-experiment-dataset

(human made, human edited) or deepfake (AI generated). When assessing an image that comes with social media layout, only the authenticity of the image is asked. You can't go back to previous questions but you can change your opinion of whether the image is authentic or deepfake when you see the same image again with additional information. Do not use any tools or internet when doing this survey. The amount of correct answers does not affect the grading of the assignment, but it is important that you try your best when answering the questions." Similarly as in [58], instructions and questions are designed to be very clear and to not give any "helper information" to the participants.

After the general instructions, the survey has the detection task of 20 unique images: without context, with context and with context and Intent Label totaling to 60 images. Each image is on its own page with three questions. Student is forbidden from going back to previous questions after moving on to the next one, to prevent them from changing their answers afterwards. Question one is multiple choice question "Is the image authentic or deepfake?", with two answer options "Authentic" and "Deepfake". Only two options are given and the question is made mandatory to answer to be able to analyze the detection accuracy of the students. As is explained in [58]: "If "I do not know" is given as a choice for classification task, the results become more of experiment towards measuring a person's self-confidence than detection ability". Study [17] also claims that deepfake detection studies should avoid Likert scales/continuous measures for categorizing stimuli as fake or real because the accuracy is then difficult to calculate. Instead of adding a choice of uncertainty or having a Likert scale to answer, confidence is asked as the second question "How confident are you in your previous answer?" for each image with a scale of 1 to 5 where 1 = not confident and 5 = very confident. Question three "How likely do you think the image is posted with malicious intent?" is also asking for students rating from a scale of 1 to 5 where 1 = very unlikely and 5 =

very likely. With these questions, it will be possible to analyze whether the context can affect students' confidence or how malicious the post seems if it does not affect the actual detection accuracy. After each set of three images, there are two optional open-ended questions: "What cues or features did you use when making your choice? Describe how you made your decision for this set of images." and "Did your answer change at any point during the previous questions? If yes, what made you change your decision?". These questions could give deeper understanding of how students are making their choices. These questions are only asked for each set of images to reduce the cognitive overload and to make it more meaningful for the students to answer only once for each unique image. That is why this survey is designed to have the same image with different conditions in a row instead of in random order. The sets of three images are given in randomized order, but the order is same for every participant taking the survey because of platform limitations.

After the detection task, the survey has questions about use of contextual cues and opinions on Intent Label with instructions: "This is the last part of the survey, asking about your use of context in previous questions of this survey.". The first question "Select all cue categories you used when doing the detection tasks (choosing authentic or deepfake)" has selection of cues from every contextual cue category presented in 4.1 as well as option "I used only my intuition". Then, Intent Label is more explained: "In the survey, the last image of each set had an Intent Label, showing the intent of the post's creator, for example "Advertisement" and "Humor". Answer whether you agree or disagree with these statements about this new idea of Intent Labels.". Then seven questions of scale 1 to 5 where 1 = Strongly disagree and 5 = Strongly agree are given: "I noticed the Intent Label easily.", "I understood the meaning of the Intent Label easily.", "I thought the Intent Label helped to complete the task faster.", "I thought the Intent Label made it easier to guess the purpose of the post.", "Knowing the poster's intent is important.", "Adding

Intent Label to social media platforms could help in assessing whether the post is authentic or deepfake." and "Adding Intent Label to social media platforms could help assessing whether the post is posted with genuine or malicious intent." to gain more understanding of how useful students see the Intent Label. This can give important information to develop the concept further.

4.4 Implementation of the survey

The experiment was implemented as part of a course "DTEK2029 Human Element in Information Security" which is a mandatory course for students in "MDP in Information and Communication Technology, Cyber Security" and "Tietotekniikka, Kyberturvallisuusteknologia (DI)" in the University of Turku. Over seventy students were taking the course in 2026, making it suitable sample for the experiment. The survey was given as individual mandatory assignment for the course, being part of topic discussing deepfake threats. Students were given two weeks to finish the task and answer the survey. Two weeks were given so that the students would have enough time to answer it as a part of other coursework. Students were asked to give their name in the survey, but all identifiable information was deleted before the course instructors provided the answers for further analysis for this thesis work.

After the results were gathered, a recording of overview of the results and the correct answers was given to students. Moodle discussion forum is also opened for students to give feedback about the survey and the experiment. This feedback is not part of the results of this thesis.

5 Results and Discussion

In this chapter, results from the survey-based experimental study are presented. First, overview of the data is given and quantitative findings are explained. Then, thematic analysis is implemented for qualitative data gotten from open-ended questions of the survey. Finally, limitations are discussed.

Survey got 73 responses. There was no missing data or incomplete responses. Average answering time was 1 h 3 min 11 s, excluding one student taking 2 d 13 h 20 min 56 s (probably due to the student starting the survey and continuing it later, so the actual answering time would be different), which suggests that the students were doing the survey thoroughly without long pauses in between. Answering times ranged from 5 h 4 min 35 s to 9 min 46 s. Students that filled the survey very fast might not have given it their full effort as it was noticed that some of these students had answered exactly same answers to every question about their confidence and malicious intent. However, these answers were not excluded from the data to prevent researcher bias as it cannot be proven that the reason for giving the same answers was especially due to lack of motivation as these students were still doing the detection task by giving unique answers on the actual detection part (selecting authentic image or deepfake) of the survey. Five students also mentioned use of AI tools and search engines specifically on one image that required external knowledge to better understand the context even though the instructions were to not use them in the detection task. It might be possible that also other students have used them

but not reported the usage, so this should be considered when analyzing the results. These answers were not deleted from the data as they can still give insights of how the students were doing the detection task, to prevent researcher bias and because the students that reported using the tools were not reporting using them for every set of images. One student also mentioned that using reverse image search would be profitable, but did not use it as the instructions were forbidding it. Background information of the students was not collected.

5.1 Quantitative analysis

Tools used for the quantitative analysis were Microsoft Excel[1] and jamovi[2]. First, mean accuracy of the overall survey was calculated including images in all conditions: images without context, with context and with context and Intent Label. Overall detection accuracy was approximately 62.24% which seems to support other human deepfake image detection results from existing research. However, to compare this result better with other human deepfake detection research, overall accuracy of images in condition one, only image, is better comparable to existing research because most of the research used only images in their experiments. Overall accuracy in condition one was approximately 65.14%, which is still similar to existing research. It was above chance level of 50% accuracy. When comparing accuracies of detecting authentic and deepfake images, they were very similar, authentic images having slightly better accuracy (62.47% and 62.01%). However, it has to be noted again that as there were images with different image and context pairings, comparing only the accuracies of authentic and deepfakes when there was only image would be better comparable with existing research. Students detecting authentic images in condition one (only image) achieved accuracy of approximately 70.68% and deepfake images

[1]https://www.microsoft.com/en-us/microsoft-365/excel

[2]https://www.jamovi.org/

in condition one achieved accuracy of approximately 59.59%. These findings show that students were better at detecting authentic images than deepfakes. These findings support findings from [13]. It can also be seen that some images were easier to detect than others, accuracies ranging from approximately 95.43% to 30.59%. This kind of variance is supported by findings from [2] where the accuracy widely varied between images generated with different models. Images achieving the highest and the lowest accuracies were both sets of deepfakes, the highest one with suspicious context and the lowest one with trustworthy context. From the qualitative data, it was noticed that deepfake images with highest detection ratings had often mentions of typical AI-generated indicators for example texture, smile and smoothness, meaning they corresponded to what kind of deepfakes the students had experienced the most or what they thought deepfakes would be. As was stated in one of the answers: *"Overall, this is a textbook example of AI-generated imagery."*. This means that there are different difficulty levels of deepfakes for humans to detect, and high resolution quality is not always corresponding to high deepfake quality. Paper [12] states that prior knowledge of synthetic media does not affect the performance, but findings of this thesis suggest that it can help detecting deepfakes that have similar features to the ones that people have seen before.

Detection accuracy varied between individuals. Some students were better at the detection task than others. Accuracies varied from approximately 88.33% to 25.00%. To further analyze whether the accuracy varied between different conditions and to research the impact of the context, this individual level variance should be considered. In the following sections, repeated measurement ANOVA is applied to test for differences between conditions, between groups, between conditions and groups of different image and context pairings and for confidence ratings in different conditions.

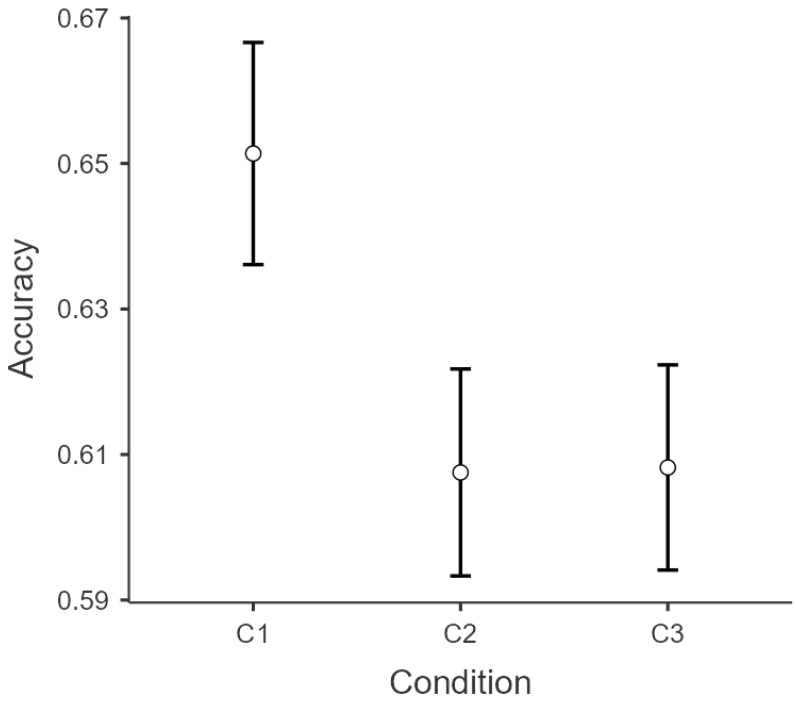

Figure 5.1: Differences of accuracies in different conditions

As seen in Figure 5.1, where C1 means images without any context, C2 means images with social media context and C3 means images with social media context and Intent Label, mean accuracy drops from approximately 65.14% to approximately 60.8% (60.75% in C2 and 60.82% in C3). Using jamovi, it was noticed that the data violates the assumption of sphericity, so the ANOVA test was corrected with Greenhouse-Geisser correction. Results show p value $p<.001$ when $\alpha = 0.05$ meaning there is significant difference between some conditions. Statistical details of this ANOVA test are seen in Figure 5.2. With Post Hoc Comparisons, it can be seen that the significant differences are between C1 and C2, and C1 and C3. There is no significant difference between C2 and C3 as p value gotten with Tukey's Honest Significant Difference test for that comparison was $p=0.996$ and $0.996>0.05$.

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Accuracy in different conditions	**None**	0.0921	2	0.04604	13.0	<.001
	Greenhouse-Geisser	0.0921	1.68	0.05490	13.0	<.001
Residual	**None**	0.5113	144	0.00355		
	Greenhouse-Geisser	0.5113	120.76	0.00423		

Note. Type 3 Sums of Squares

Figure 5.2: Statistical details of repeated measures ANOVA test for differences of accuracies in different conditions

Differences between different context groups are seen in Figure 5.3. Group A includes authentic images with trustworthy context, group B has deepfakes with trustworthy context, group C has deepfakes with suspicious context and group D has authentic images with suspicious context. Repeated measures ANOVA with Greenhouse-Geisser correction shows significance differences between groups as $p<.001<0.05$. Statistical details are presented in Figure 5.4. Figure 5.3 shows that groups A and C have higher detection accuracy and these are the groups where context supports the detection making. For groups B and D, context is hindering the detection. Post Hoc Comparison with Tukey's Honest Significant Difference test shows that there is significant differences between groups A and B with $p<.001<0.05$, A and D with $p<.001<0.05$ and B and C with $p<.001<0.05$ and C and D with $p=0.002<0.05$. There is no significant differences between groups A and C with $p=0.996>0.05$ and B and D with $p=0.999>0.05$. Differences between groups that include same type of images (authentic or deepfakes) but with different contexts are all significant.

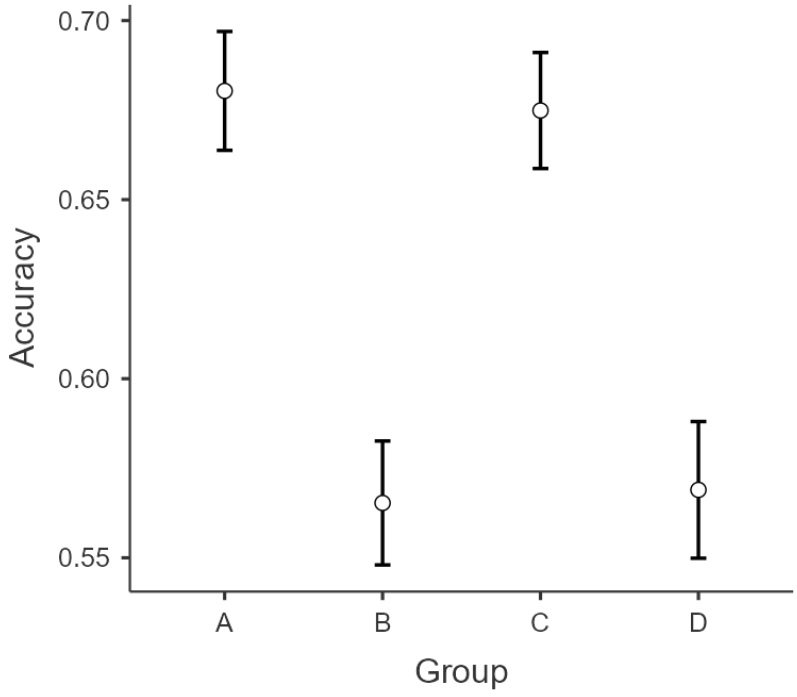

Figure 5.3: Differences of accuracies in different context groups

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Group	None	2.68	3	0.8930	13.4	< .001
	Greenhouse-Geisser	2.68	2.52	1.0633	13.4	< .001
Residual	None	43.66	654	0.0668		
	Greenhouse-Geisser	43.66	549.23	0.0795		

Note. Type 3 Sums of Squares

Figure 5.4: Statistical details of repeated measures ANOVA test for differences of accuracies in different context groups

Figure 5.5 shows how accuracies differ between the groups in different conditions. Repeated measures ANOVA with Greenhouse-Geisser correction shows significant differences in this interaction with $p<.001<0.05$. Statistical details are presented in

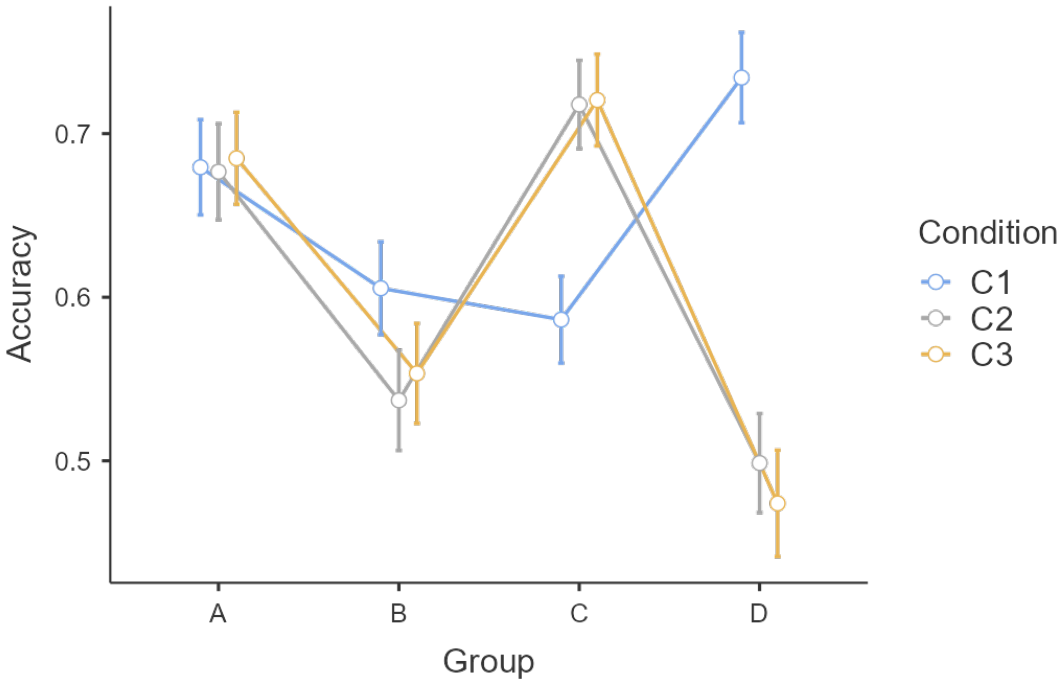

Figure 5.5: Differences of accuracies in different conditions and context groups

Figure 5.6. Post Hoc Comparison with Tukey's Honest Significant Difference test verifies significant difference between conditions C1 and C2 as well as C1 and C3 in group C and between conditions C1 and C2 as well as C1 and C3 in group D. There are no similar significant differences in groups A and B. These findings show that especially having suspicious context impacts the detection accuracy. Having authentic image with suspicious context decreased the accuracy and having deepfake with suspicious context increased the accuracy. Conditions inside groups having trustworthy context did not have significant differences, even though it can be seen from Figure 5.5 that trustworthy context still decreased the accuracy of detecting deepfakes in group B. There were no significant differences in any group between conditions C2 and C3, meaning that the addition of Intent Label did not impact the accuracy significantly. However, these findings suggest that context has impact and it can support but also mislead human deepfake detection.

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Condition	None	0.368	2	0.1842	12.97	< .001
	Greenhouse-Geisser	0.368	1.68	0.2196	12.97	< .001
Residual	None	2.045	144	0.0142		
	Greenhouse-Geisser	2.045	120.76	0.0169		
Group	None	2.679	3	0.8930	5.99	< .001
	Greenhouse-Geisser	2.679	2.32	1.1529	5.99	0.002
Residual	None	32.208	216	0.1491		
	Greenhouse-Geisser	32.208	167.31	0.1925		
Condition ✻ Group	None	3.695	6	0.6158	34.28	< .001
	Greenhouse-Geisser	3.695	3.87	0.9536	34.28	< .001
Residual	None	7.759	432	0.0180		
	Greenhouse-Geisser	7.759	278.96	0.0278		

Note. Type 3 Sums of Squares

Figure 5.6: Statistical details of repeated measures ANOVA test for differences of accuracies in different conditions and context groups

Repeated measures ANOVA with Greenhouse-Geisser correction was also done to analyze if confidence rating changed in different conditions. Figure 5.7 shows that confidence ratings increased with given context. Significant difference was found as $p<.001<0.05$. Statistical details are presented in Figure 5.8. Post Hoc Comparison with Tukey's Honest Significant Difference test shows significant difference between every group with $p<.001<0.05$ for C1 and C2 as well as C1 and C3 and for comparison C2 and C3 $p=0.008<0.05$, meaning that the Intent Label also impacted the confidence. However, it can't be fully said that specifically the Intent Label was the reason for the difference. As students were able to see the image a third time, this longer time of analyzing the same image could have impacted the results. In [14] and from qualitative data of the survey, it is noticed that display time of an image can have an effect, so the repetition of the same image might have affected the confidence in the last condition. It can be stated that having context however

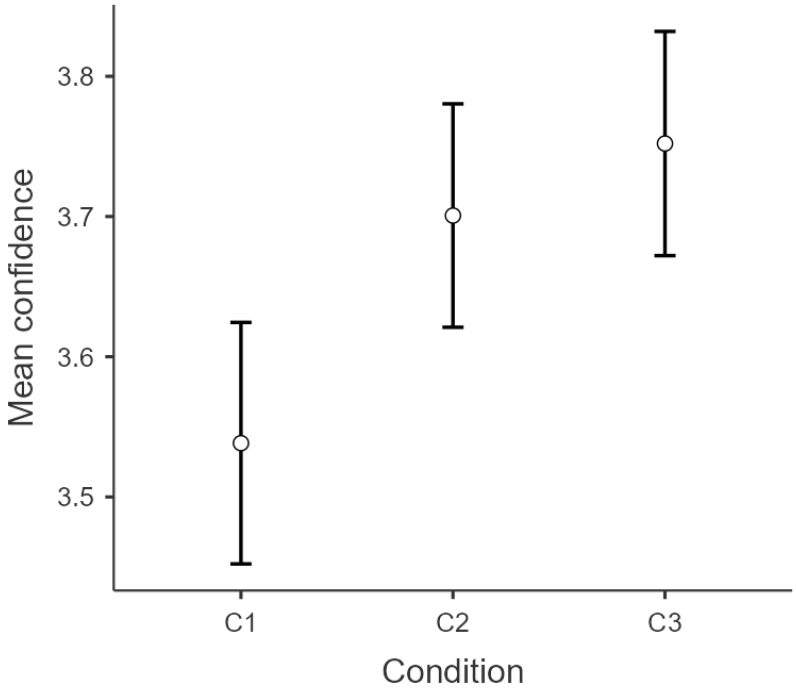

Figure 5.7: Differences of confidence ratings in different conditions

increases the confidence with or without Intent Label.

Confidence was also compared with accuracy. With jamovi, individual accuracies were compared with individual mean confidence ratings to test if there is correlation between these variables. Pearson correlation coefficient showed very weak negative correlation between the variables where r was approximately -0.052. This correlation is seen in Figure 5.9. Spearman's rank correlation coefficient also showed similar result of ρ=-0.062. This correlation is not significant as the p-value of Pearson correlation coefficient test as well as Spearman's rank correlation coefficient test were >0.05 (p=0.665 and p=0.602). Even though these findings do not support the claim in [50] that people would generally perform better when they are more confident, it can be seen from Figure 5.9 that some students are overconfident which supports the idea of overconfidence within groups with more experience on deepfakes such as cybersecurity students. However, these results are more similar to findings in [2] where it is stated that confidence is not related to accuracy. It can also be

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Condition	None	1.82	2	0.9083	27.4	< .001
	Greenhouse-Geisser	1.82	1.33	1.3677	27.4	< .001
Residual	None	4.78	144	0.0332		
	Greenhouse-Geisser	4.78	95.63	0.0499		

Note. Type 3 Sums of Squares

Figure 5.8: Statistical details of repeated measures ANOVA test for differences of confidence ratings in different conditions

observed that some students do not feel very confident but are performing better than chance level which might indicate that the detection task itself is difficult. Further analysis of the reasons for varying confidence is presented in Section 5.2.

Malicious intent was generally higher for images with suspicious context and reasons for changes in the score are further analyzed in Section 5.2. Percentages of students that reported using specific contextual cues at the end of the survey are seen in Figure 5.10. Most students used quality of the image, visual artifacts in images and post's description as cues which are all from the category of perceived quality. Source cues and intent were the next most used contextual cue categories. 45.2% of the students reported using only their intuition when making the choice, but it was noticed from the data that some students that chose this answer also chose other cues in addition, so the word "only" is not accurate in these results and this category can include students that used their intuition in some images but not in all of them.

It can be seen from Figure 5.11 that knowing the intent of the poster was seen as something important, but the Intent Label itself did not seem very useful for the students. It can be seen that some people did not notice the label easily and did not understand what it could be used for. This experiment setup was to test how

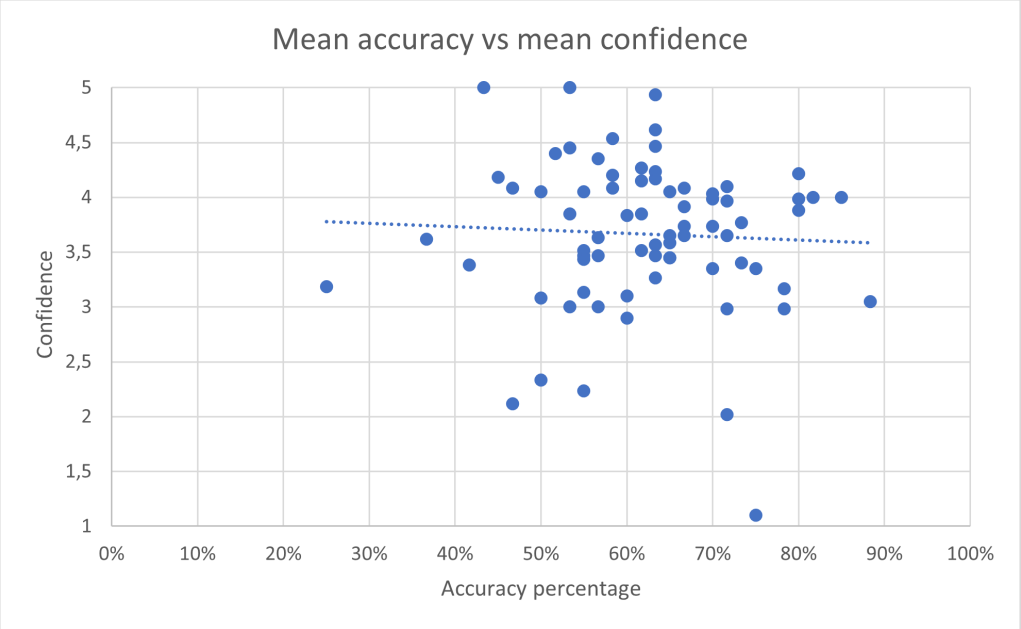

Figure 5.9: Correlation between accuracy and confidence

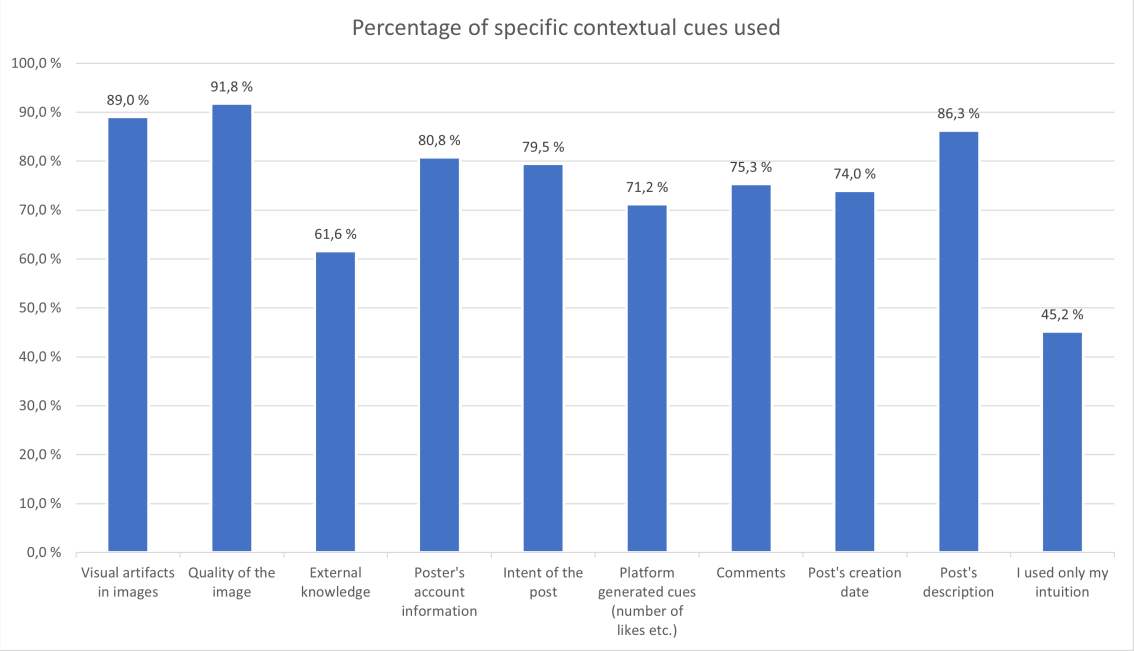

Figure 5.10: Percentage of used contextual cues

the students would interpret the label without explaining it to them before doing the detection task, because the survey wanted to experiment what cues students are using without giving them any helper information. However, the concept of Intent Labeling could have been seen as more useful if the experiment was designed differently and the label would have been explained more before doing the task. Results from this experiment did not show significance when comparing accuracies of images with context and with context and Intent Label. Confusion and the experiment setup might have affected the way students saw the usefulness of Intent Labels. However, because the intent was still seen as important, the idea of Intent Labels could be further experimented and developed in the future.

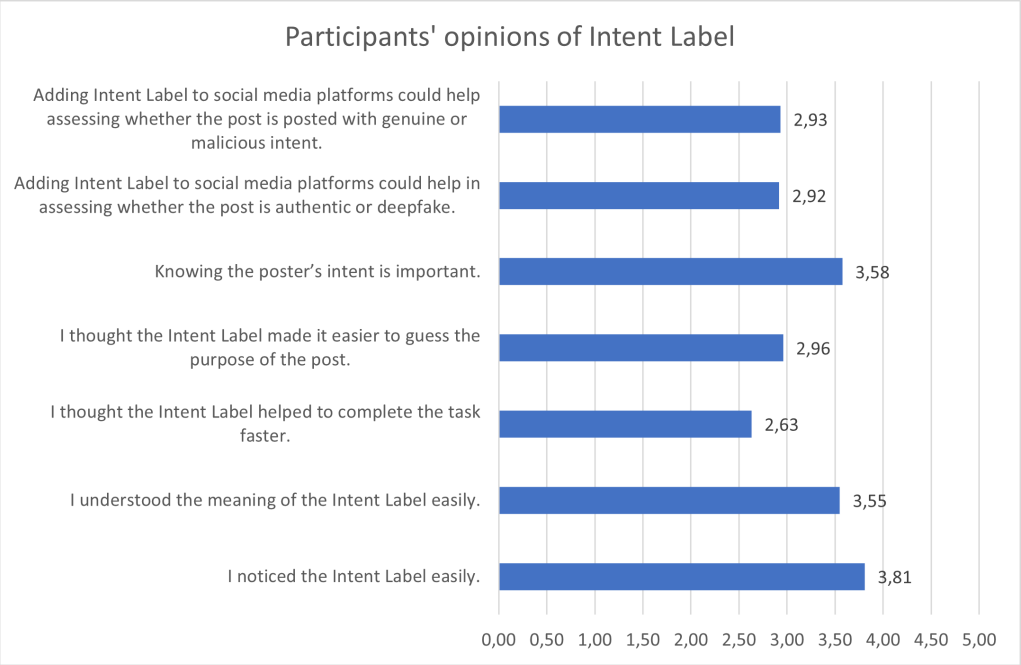

Figure 5.11: Students' opinions of Intent Labeling

5.2 Thematic analysis

There were 932 responses to open-ended questions asking what cues, features or strategy the student used when doing the detection task and 694 responses given to

question about changing the answers when more context was given. This thematic analysis was manually implemented following Braun and Clarke six-phase framework: reading the data, coding it, searching themes, reviewing themes, defining themes and writing the final analysis. Nine themes were defined, which are the following: Focusing on perceived quality, Focusing on consistency, Using external knowledge, Focusing on source, Interpreting intent, Focusing on interactions on the post, Using intuition, Impact of context and Human element and experiment design. Some of the themes matched the categorization of contextual cues presented in Chapter 4.1.1 in Table 4.1 and in addition, few other themes emerged from the data. Following sections will present these themes and give examples of responses from students to support each theme.

Focusing on perceived quality

Perceived quality can be divided into four sub-themes: "Focusing on visual details", "Focusing on visual quality", "Focusing on caption quality" and "I didn't find any artifacts so I believe this is authentic". Students used many different visual details when doing the detection task, including background, objects, people and text in the image. Background details used included analyzing nature, water, plants, individual strands and branches of trees, rocks, clouds, snow, weather, street, mess, cars, ceiling, windows, glass, reflections, shadows, pillars, poles, lines, architecture, design (of interior) and background depth. Instrument, fishing rod and gun are examples of specific individual objects that were used to do the detection. If the image portrayed a person, students used very detailed visual cues in the task, including posture, pose, face, skin texture, skin tone variation, wrinkles, moles, freckles, lips, texture of the lip, hands, fingers, finger nails, eyes, eyelids, clothes, fabric, materials, colors, shape, hair, individual hair strands, hair roots, smile, teeth, emotions, blush, ears, jewelry, accessories, feet and veins. Study [14] notices that people are often

suspicious especially of hair, eyes and ears. In this experiment, students focused on hair and eyes but instead of ears, skin and hands were mentioned more often and in more detail even though some students also mentioned using ears as visual cues. Aspects of visual quality that were used included blurriness, sharpness, lighting, angles, framing, arrangement, resolution, granularity, errors in image rendering, focus, style, cropping, smoothness and saturation. Issues with texture, lighting and geometry were also cues mentioned in [50] as well documented visual artifacts that were used in deepfake image detection task.

These details were compared to expectations and observations of real world. Students analyzed persons in the image and compared the observed details to what they believed would be authentic, like in this answer: *"The natural skin texture gave realistic aging details like wrinkles, skin tone variation."* Students also analyzed physical accuracy of the visual cues in the image. If it corresponds to the real world, as in this answer *"The reflection seem to capture the reverse picture properly"*, the image was seen authentic and if it doesn't correspond, as in this answer: *"The ruffles on the dress are too gravity-defying."* it was seen as deepfake. Students also analyzed whether the action the image was trying to portray seemed logical and correct. These answers give examples of their analyses: *"the hand is holding a chord on the guitar, it is not "randomly placed""* and *"The handgun is also weird as I don't recognize it, it looks like a weird mix of few different guns, it's weirdly stocky and clunky looking, and the pistol's iron sights seem to be useless as it doesn't even really seem to have all the parts needed for a functioning crosshair. It's also weirdly upsetting that the person is "aiming" a weapon while looking at the camera, while the fingers seem to have weird proportions and his thumb is whole left hand is positioned very weirdly for holding a weapon, and his thumb seems to also be impossibly far back."*. These findings are similar to categories of noticeable errors such as "functional implausibilities" and "violations of physics" presented in [14]. In

addition to focusing on one action, students also analyzed the overall situation the image was trying to portray, as in these answers: *"Bro has no fishing gear other than the rod, which is funny."*, *"COVID time + at airport with no mask? Someone would have stopped them probably and forced to wear a mask."* and *"I'm fairly certain the image is real because of these sorts of sculptures/thingamajigs are common enough to be here and there, they are likely doing maintenance on it by fixing paint or some bolts based on the tools strewn about"*.

Some participants focused on visual details that contradicted with their existing knowledge or expectations of the real world without analyzing them further. Relying on this type of cues often led to incorrect detection, for example *"The guitar has only two visible strings"* and *"The sky is sunny while its flood? The cars parking is kind of a bit off? (going into grass)."* were used as cues to incorrectly classify the images as deepfakes. However, some students analyzed more beyond what was seen in the image. With this strategy, another student was able to make the correct detection using the same visual cue of two guitar strings: *"At first I thought the strings were missing, but then I realized they were simply vibrating."*. Some students also noticed that AI would not be able to make images that required more high level of logical analysis and thinking, as in this answer: *"I feel like the image is real due to how the cars and water look, how the reflections look on the water, and how some of the cars are on the higher median so that they aren't sitting in the water, which seems like a thing that AI wouldn't predict as something people would do if there was a flood and thus wouldn't generate cars there. The clouds and trees also look real enough."*. These examples from the data support findings from [20] stating that cognitive reflection and holistic thinking have an effect on deepfake detection accuracy. It is therefore an important finding that students were able to make logical detections and use cognitive reflection to increase their detection accuracy, as this kind of strategy could be difficult for technical detection solutions.

Visual quality was also compared to expectations of real world as in this answer: *"The picture is in right sharpness. Also the picture is focused as usual cameras do."* and *"Picture quality is poor for a promotion"*. Quality was also further analyzed based on the situation the image would have been taken as in this answer: *"It just looks like the kind thing Pharrell and Pharrell-adjacent people would wear, and lighting and shadows look like the result of paparazzi flash photography, though with a high flash."*. Images with good quality, that the students described to look too perfect, were also seen to be result of professional setup, editing or AI. Study [50] also notices that "too perfect to be real" is used as a reasoning for an image to be deepfake. These findings slightly contradict with claim in [72] which states that good quality content would be seen as more credible, because it is seen as an effort toward the content, and therefore bad quality would be seen as less credible. It seems to not always be true, as some students are comparing the expected quality from specific source to the image. If the account seemed small or unprofessional but the image had excellent quality, it was found suspicious. When analyzing images of low quality, it was proved difficult to use visual cues in detection as is explained in these answers: *"The resolution is very low, so I am not sure."* and *"No, however there is a slight chance that this could be fake I would need to see a higher resolution image."*.

As the students were cybersecurity students they were familiar with AI, AI-generated images and deepfakes. That is one reason they were able to compare the visual details and quality also to their knowledge and expectations of AI. Students were focusing on facial features, fingers, hands, patterns, hair, creases on clothes and text because they thought that AI would make mistakes with these details. Similar advice, for example suggesting to focus on 'Strange Clothing Fabric', was given in [2] and it was found to lead to correct choices more often. Students also noticed when something that would have been difficult for AI to generate was missing, for

example when there was no pattern on the clothes or complex surfaces like skin with moles or authentic lip texture. AI-generated skin was instead described as rubbery, wax-like and having a plastic look. If skin or the overall image looked too smooth or too perfect it was seen as AI-generated. Some students focused on emotions like smiling and compared it to what kind of emotion AI usually creates as in this answer: *"The smile is very similar in a lurge chunk of AI slop posted online."*.

Students thought that AI would more likely portray popular brands like Burberry's scarf and make them more accurate. Some situations and arrangements were also seen more common for AI for example: *"The composition of a person doing a job and smiling widely into the camera is a very common scene in the world of deepfakes. The smile is bright, the skin is smooth, the person is sharp and the background is blurry. The fact that everything is well-organized and lined up in the background is another sign towards algorithmic generation. It looks as if a very generic clay working environment has been created by AI"*. Images having very generic background or setup were seen more likely to be deepfakes than images with complex setups, like images with multiple people, scattered objects or mess as is explained in this answer: *"The mess on the table is hardly reproduced by generative models"*. Study [14] also notices that complex scenes are easier to detect as they often have more details for AI to generate wrong. Typical AI art style was described and noticed by some students: *"The swirly texture, yellow-gray hue, and specific line thickness all look like typical AI illustrations"*. Specific type of lighting and angles were also linked to being a deepfake. Blurriness was seen as a cue for both, authentic images and deepfakes. Blurry background with sharp objects or person was seen as something that AI would create, but having consistent slight blur was seen more likely as authentic. If no artifacts or clear cues associated with AI were found in the picture, some students used that as a cue of the image being authentic as in this answer: *"My confidence has dropped from the previous ones. I didn't find anything*

that was particularly common for AI, and based my decision on that.". Paper [15] presents strategy of audio deepfake detection where people seemed to believe that the audio was authentic, if no specific fault was found. This finding could therefore be extended to image detection.

In addition to visual details and quality in the actual images, the caption quality was analyzed by some students. Language of the caption was described as clear, formal, corporate-like, dismissive, positive and shaming. The style of the language affected how malicious the image was seen. Having a link in the caption, suggesting to send a DM or using specific hashtags were also seen as signs of malicious intent and to belong to deepfake posts, as is explained in this answer: *"Without context the image seems real. With context the image is likely a deepfake. The link in the post is suspicious and indicating that the post is fake news."*. If the student agreed with the caption, it was more likely seen as unmalicious. If the description text of the image looked AI-generated or nonlogical, the image itself was also seen to be more likely a deepfake. One student was also giving an example of a character LLMs often use: *"the text are using - sign which is common on LLM"*. Analyzing the text this far requires detailed knowledge of LLMs.

Focusing on consistency

Consistency was mentioned by many students. Consistency was generally linked with authenticity and inconsistency with deepfakes. In the image itself, for example consistent lighting and visual details such as consistent width, shape and pattern were inspected. In addition to consistency within one object in the image, consistency of logos in different accessories was observed. In [2] 'Asymmetric Earrings' was also given as an example of advice that led to correct choices more often and it has a similar idea of checking consistency between similar objects or objects that should be matching in authentic images.

However, it can't be said that inconsistency would always be linked to deepfakes. Some students were able to further analyze the reason for inconsistency in specific situations, as in this answer: *"The books on the right aren't consistent in their appearance, but that could also just be a real symptom of yearly releases, for example."*. This analysis led to correct detection, while many other students found the inconsistency, and therefore chose deepfake even though the image was authentic. The overall consistency of the image was also observed as well as consistency of a post. One student realized inconsistency within the image quality and the situation the image seems to portray after thinking it longer and was able to make a correct detection: *"Without any context, the image could appear real; however, in retrospect, the low resolution and the studio environment are somewhat inconsistent, which may indicate that it is a deepfake."*.

Account and information from description were also observed in the viewpoint of consistency. Low amount of likes compared to high number of followers was seen as fake as well as low number of followers compared to high number of accounts followed. If the information given in the description didn't match the image, the post was seen inconsistent as in this answer: *"No Adidas logos in the ad, the link is buried in the text to trick fast clickers, there's a typo in the link, and the account has too few followers and posts for Adidas, and it is just created."*. Image was also compared to the timing of the post as in this answer: *"At first it looked like an old photo, but its posted in 2026. and the creation date is in 2026."*. Inconsistencies were also found when comparing the interpreted intent and the Intent Label in categories of News and Informing. In addition, consistency between caption and the comments, talking about the same topic, made the image seem more likely authentic.

Using external knowledge

External knowledge was used as a cue in the detection task. Some students reported being familiar with the person that was portrayed in the image, making the detection easier. Being familiar with the person made it possible to compare the looks but also the information from the context for example timing and description and see if they match with the real events. Familiarity is noticed for example in this answer: *"Seeing the date this was posted and the fact that this is supposed be Jason Bourne -related makes me much more confident that this is a deepfake. Maybe there's a TV show I missed, but that's not what Matt Damon or Jeremy Renner look like. The others look like AI rip-offs of real actors too, like Bill Skarsgård or Luke Evans."*. When the student was not familiar with the person, the detection was described to be difficult. Similarly, familiarity was said to improve detection accuracy in [17] and [55]. Importance of topic knowledge was also noticed by some students, as in this answer: *"The image doesn't really have anything you can form a opinion on, except if you know what the British museum looks like I guess?"*. This is similar to [57] stating that domain specific knowledge is important. Some students also wrote that they had seen the image before, so they could be more confident that the image was authentic. However, it seems like some students had only a false memory or feeling of seeing the image before, leading to incorrect choice. It is stated in [84] that familiarity with the content leads to increased credibility and trust but not increased accuracy. This could explain why some students thought some images looked familiar and therefore classified them incorrectly as authentic.

In the previous themes, students were also using their expectations and knowledge of the real world and AI when looking at the image. More detailed knowledge of the objects portrayed in the image can help making a correct detection as in this answer: *"That gun looks like a Beretta m9 but the sights on a beretta do not look like that"*. Knowledge of cameras can also help making deeper analysis of the image

as in this answer: *"lighting consistent with strong camera flashes consistent with paparazzi pictures. Comprehensible background text, background dimmed enough to be consistent with DSLR cameras with strong flashes."*. If the student knows more about cybersecurity and social media, possible malicious intents can be further analyzed as in this answer: *"Account is trying to sell clothes and directs viewers to http://... links that are not https://... and overall seem to be rally suspicious."*. However, if the student has false information or assumption of the topic, it can lead to incorrect choice. For example, in the authentic image that portrayed designs from Burberry, some students knew that the scarf was Burberry's design but then they thought the logo in earrings was not Burberry's logo and based their decision on that detail, even though in reality the logo in the earrings was one of Burberry's logos made for a special event.

Analyzing the timing of the post in the context of historical knowledge and the current timing can also be seen as external knowledge. Older posts are seen to be less likely deepfakes. Many participants relied more on the timing of the post than visual cues, because they thought that it was not possible to make deepfakes of good quality during the time when the image was posted. Similarly, when the image is recently posted, it is more likely to be a deepfake, because current models are seen more capable. Some of the students thought that many advertisements currently are made with AI, as well as AI art as is seen from these answers: *"When comparing to the world nowadays, many adds that I encounter seem to use a lot of AI in their images or adds. So why would this not be using AI as well."* and *"AI-generated "art" on t-shirts is getting very common these days and this looks like that kind of thing."*.

Focusing on source

Students used the source cues including account creation date, account post amount, number of followers and number of followed accounts in combination of other contextual cues to profile the accounts and then compare them to their expectations of other similar accounts. Influencers, celebrities, content creators and bigger brand accounts were seen as natural and less likely to post deepfakes as their reputations could suffer from it. Fan accounts and smaller accounts were seen more likely to post deepfakes. In addition, it was not seen impossible for a bigger brand to also use deepfakes for marketing if it has clear monetary benefit. Private profiles were considered to not post deepfakes and instead share real memories on their profile. Also specific accounts like social media account of a retirement home should not use AI images of the residents, so it would be more likely that they only post authentic images. Students were able to identify fake accounts, accounts trying to represent someone popular, bot accounts, spam accounts, hacked accounts and stolen accounts based on the source cues and their knowledge and expectations. These accounts would more likely post deepfakes. Signs of a fake account included having high amount of posts, being recently created and having inconsistent follower-followed or follower-likes ratios.

With further analysis some students noticed that newly created accounts could be authentic new accounts and accounts with low engagement can be unused accounts that are still authentic. Overall, old accounts and popular accounts were still seen more likely to post authentic images than new or small accounts as is explained in these answers: *"The likelihood that the post is deepfake, in my opinion, decreases because the account is made about 3 years prior to the post."* and *"The post changed my stance to authentic because it was from a popular account. This would need to be verified with further investigations."*. Impact of popularity and use of bandwagon heuristic was noticed also in [72] and importance of good reputation of the uploader

to how trustworthy the content is seen was highlighted in [85]. Some students expressed the importance of validating the source as in this answer: *"Environmental aspects, such as legitimacy of poster, real timing, contextual validity (sharing the image one day after the artist's death) are very important when identifying fabricated material. The visual cues were there from the beginning, but as AI advances, we can rely on these less, and in my opinion validating the source is more important"*. Some students also thought that social media account is never a trustworthy source without further research of the topic: *"I wouldn't trust this type of source in the first place until researching the topic further"*.

Interpreting intent

Students were also using intent of the post when they were doing the detection task. They were not only thinking the question of "Why is this image posted?" but also "Why would this need to be a deepfake?". Some students thought that if there is no clear reason for the image to be deepfake, then it is more likely authentic as in these answers: *"The pictures look easy to take so I don't think they need deepfake"* and *"There's no point for an account with 2 followers to post deepfake imagery or misleading information. The post is not promoted, it is not an advertisement, likely just a real moment."*. Similarly, if there is no clear reason to take authentic photo, it's more likely a deepfake as is explained in this answer: *"In the legs of the woman you can see some weird lines and the photo seems just weird. Why would someone have taken this."*.

Intentions to post deepfakes that students mentioned included monetary gain, saving money, politics, getting attention, endorsement farming, phishing and other scams. Advertising was seen as very common intention to use deepfakes. Students had assumptions about specific intent categories such as Advertisement, Networking, Humor and Entertainment, that can be seen for example in these answers: *"When*

I saw that it was an advertisement I changed my answer to a deepfake, because advertisers are lazy and cheap", *"Yes, suspicious of the post context. Unless someone created a separate "professional" profile, a new profile immediately creating engagement posts is suspicious."* and *"Image seemed photoshopped but not AI. Post date made me feel it was maliciously being spread but 'entertainment' context made me change my mind."*. Malicious intent was sometimes connected to being a deepfake, as was explained in source theme where for example fake profiles were seen more likely to post deepfakes. However, in some cases malicious intent did not guarantee the image being a deepfake as in this answer: *"Post context, suspicious link (I doubt the genuine Adidas would use an account created in 2025 and not have the domain adidas.com already registered). Scam, so definite malicious, doubtful on whether it's deepfake or reused image"*. It was noticed that unmalicious intent can also lead to classifying image as a deepfake as in this answer: *"The left arm seems to just disappear into the dress on the right side of the picture, the dress doesn't seem to have any type of coherent way the ruffles are forming which shouldn't really be a thing as they too need to be attached to the dress, the upper part of the dress seems weirdly rigid, the ruffle bow feels like it's defying gravity, and there's something weird going on just under it on the bottom left from the bow. The post doesn't seem, malicious however, and more like a promotion for products by a company that's using AI to instill products on ads so that they can save on modeling costs."*. The student still explained the intent to use deepfake, even when it was not seen malicious. Some students also saw harmless and relatable posting intent supporting their choice of the image being authentic as in this answer: *"All of the details seem genuine: the nature, fishing. Social media context makes it seem like a man posted himself fishing fish, not credentials."*.

Focusing on interactions on the post

Platform cues together with networked cues can be defined as interactions on the post. Higher number of interactions was seen as more likely authentic unless the amount was unbalanced for example having over 2000 comments and few likes as that was seen more like spam and not genuine interactions. Images having small amount of interactions such as likes and comments were seen to be more likely deepfakes. Shares were also mentioned as cues, and fast sharing was seen as authentic as explained in this answer: *"It seemed like the meme was already being shared quite quickly. I believe the image is authentic."*.

Comments were further analyzed by some students. The content, style and timing of the comments affected the interpreted authenticity. Images that had comments that seemed more bot-like were seen as more likely deepfakes and comments that seemed human-written were seen more authentic for example in this answer: *"The chats snippet shows a form of real human communication. The posture of the woman also looks real."*. If comments include links they look malicious and can make the image look more likely a deepfake, however, some students though that the comments might not be related to the post or the poster at all and be separate people trying to scam in the comment section of another post. Students also seemed to trust comments pointing out a scam or use of AI and use them to support their decision as in this answer: *"The image seems genuine but the link to the site is clearly a scam. Someone even says in the comments that the site is a scam"*. Further analyzing the comments included seeing some unmalicious looking comments as a spiel, trying to make the post purposely look more authentic and therefore increasing the maliciousness score of the image. Students also considered the timing of the comments. If the comment is left right away, the post is more likely to be a deepfake as is explained in this answer: *"It's unlikely that someone would immediately comment on a post if it's the first one."*. The original poster re-

plying or reacting to a comment made the image look more authentic. This finding contradicts findings in [85], stating that discussion between users were not affecting how credible the post is seen.

Using intuition

Intuition includes feeling that something is off, weird or uncanny while not being able to exactly explain what caused the feeling, using expression of "gut feeling" or feeling the image is deepfake or authentic without any specific reason. Using gut feeling was described for example in this answer: *"The image doesn't seem quite right. I can't be precise on what's wrong with it, I just get a weird feeling about it. This is basically only based on my gut. Now that I think more about this, if this was a legitimate ad made by adidas, the account wouldn't have been made in 2025 (although I don't know if it's common for them to make new accounts for advertisements) and I don't think that adidas would have only 2,5k followers and follow about 1k people. Now I think that this add is malicious (I should have also increased the malicious intent in the previous one). This might be a fake add to redirect the person clicking on the link to a malicious page. I'm not sure that my interpretation is correct, this is just the feeling that I get."*. Some students first used visual details as cues, but decided to trust their gut feeling when making the final choice as in this answer: *"I tried to look at the reflections and faces, my gut feeling told me that the image is deepfake."*. This type of strategy often led to making a wrong choice, meaning that intuition is not a trustworthy cue to use when detecting high quality deepfakes.

Some students were able to point out a specific detail in the photo that felt or looked weird or uncanny as in these answers: *"The people on the photo seem off, I can not quite tell why."* and *"glasses look off, clothes look off, face looks uncanny"*. The situation that the image portrayed could also cause the feeling that something is off. Some students also described the image feeling AI-generated, emotionless,

lifeless or real. Some students seemed to be very confident in their intuition as in this answer: *"It just looks AI-genarated. I dont know how to explain it, but I am pretty confident."*.

Impact of context

Impact of the context varied between students. Some students reported that they didn't change their answers, by answering "No" to the question asking whether they changed their answer after they saw the context or when they saw the context with Intent Label or not. Some of these answers were due to students' initial thought being similar to what the contextual cues supported, so these answers of "No" are not very useful. Instead, answers that explained more the reasoning behind not changing the answers gave more information about why the context did not make an impact. Some students thought that there was no reason to change the answer, as the image stayed the same in all conditions as is explained in this answer: *"I saw the image as a deep fake because of the eyes and the instrument. The image didn't change in my opinion."*. Some started to question their answers, but did not change them as can be noticed from these answers: *"No it did not change but I started to questioning myself and my answer."* and *"No. The additional information did make me take double take"*. However, for many students, context affected the choice of deepfake or authentic, confidence and the maliciousness rating. In the following sections these impacts are further analyzed.

For changes in the actual detection, it was noticed that getting more information from context made students change their choices as in this answer: *"The first photo looked like an advertisement, and I wasn't sure whether it was real or fake. It could have gone either way. I changed my mind when I saw which brand the ad was supposed to be advertising."*. Source cues were also seen as reasons to change the choice as in this answer: *"Yes, the picture itself looked real enough, but I do not*

trust it from this source.". In some cases students noticed malicious intent and changed their choice from authentic to deepfake as in this answer: *"The resolution and aspect ratio immediately raised suspicion, but this was not enough for making a decision. The context indicated that this is likely a deepfake made for malicious purposes."*. Some students noticed that they might have analyzed visual cues wrong and changed their choice because of the context: *"Context switched me from deepfake to authentic - the suspicious amount of upper teeth may be just low image quality that I misinterpreted."*. Suspicious context was able to mislead some students into changing their choice to a deepfake even when they first saw the image correctly as authentic: *"I initally thought it might be authentic, but seeing the account age and post date, along with the spam comment and accusatory tone, made me re-evaluate the image. i think it's a deepfake after all, and not even a repurposed image."*. In some cases when the choice didn't change, the students noticed that it verified their choice as in these answers: *"My initial answer was questioning, but the second image confirmed my suspicion."* and *"I originally just guessed that it was a deepfake as it looked a bit weird but the posts text confirmed it."*. In these cases, it can be claimed that giving context still had an impact on their choices even when students didn't change their answers.

In addition, context made the students change their confidence ratings. Some students reported that their confidence increased when context was given without telling further what was the exact reason from the contextual cues to change it. More detailed answers revealed that common reasons that increased confidence included realistic account numbers, old posting date, old account creation date, clear intent and consistency between many contextual cues. Reasons for decreased confidence included uncertainty about facts and intentions, contradictions with existing assumptions and interpreted malicious intent with authentic image.

Context also affected malicious intent ratings. Getting more information from

the context combined with external knowledge of cybersecurity that the students had, often made the students to increase the maliciousness score as in this answer: *"Weird links in the comments, people looks bit unreal, "today only" tries to make people act fast, no likes, no brand seen"* and *"Malicious intent raised due to account being new and bot like comments"*. Description style could also make the image look more malicious as well as it including a link. If the profile was not seen as authentic or it posted about specific topics such as COVID or children, the post was seen more malicious. For the rating to decrease, entertainment and humor categories made the image look less malicious, even if the image was seen as a deepfake.

In general, students expressed their surprise of how much impact the context actually had as in this answer: *"Yes, context of social media post made me think it was more likely to be authentic. It surprised me how big of a difference it made."*. They noticed the difficulty of detecting deepfakes without any context given as in these answers: *"Yes, without context, it's seems hard to detect."* and *"I was not sure about which to select at first but additional info provided context which helped in deciding"*.

Human element and experiment design

This theme includes sub-themes of students not being able to tell whether the image was authentic or deepfake, confusion of the task and survey design and other notices specific to human element that might have impacted results. It was noticed from the results that when students reported not being able to tell whether the image was authentic or deepfake, in some cases they used internet to check for additional information, image search tools or AI, relied on their intuition or made a guess. One student explained the choice to use internet in the answer: *"I had to check images of BB King to actually compare the man as I had no idea what he looked like."*. Students using other tools explained their detection strategy for example like this:

"Image search tools like Google lens, Gemini AI. My final understanding in short - Found the source of it (Jan Persson from news agency AP). Maybe photoshop like tools used to softening the image that give an oil paint-like texture on it, but it's not AI. Also, I used the AI and the output says the Visual Elements are Confirmed, No AI.". Even though the use of tools was forbidden in the instructions, relying on the tools can tell about the difficulty of the detection task as well as knowledge or earlier experience of using these tools. As the students are cybersecurity students, they would probably have knowledge of tools that can be used in deepfake detection task. Some answers support the observation of difficulty of the task, for example this answer: *"it is possible that the image is real I have no way of telling."*. One student also knew about the available tools and suggested using them as is clear from this answer: *"Watched for shadows, teeth, eyes, and hair. the shadows seem constant, light source probably coming from the front-right of the supposed child models. Flat background, velvet type clothing with no challenging patterns, so hard to guess. After post context definitely a scam so chances of it being deepfakes are higher, but then again it may very well be just a reused image. Reverse google image search would work perfect here, but I've been artificially limited in the tooling that I can use."*. This also highlights the issue of the survey not matching to real world situation, where the students would have been able to use any tools they have access to. However, as the purpose of the survey was to research human deepfake detection accuracy and how context impacts it, the use of tools and internet was forbidden to get better results of what cues the students would use if they have to make the detection fully by themselves.

Another observation made from the results that might have affected the results and make the experiment differ from real world situation, was repetition of the same image. Some students reported that when they saw the same image longer and though about it more, it affected their confidence as is explained in this answer:

"Also the change for the confidence of the picture being real was from just thinking about it more". Some students also expressed that the amount of images in the survey started to make them see everything as deepfake as is expressed in these answers: *"These could be real people but I started to doubt every picture."* and *"At this point in the survey I am becoming suspicious of the images without context. With context the image seems normal and authentic."*. These answers were given for the image set seven and thirteen, where the students had seen multiple images already.

It was also noticed that some students got confused of the survey design itself. At least one student explained the reason for changing the answer to be: *"Yes, I didn't understand how the survey works"*. The experiment setup was also questioned and some students did not understand the reason for repetitions, which still did not lead to changing the answer but might have affected confidence and signal about challenges in the survey design. Answers expressing confusion of the survey design include: *"No, It can be a way to manipulate your decision making with same picture and questioning your reasons/answers"*, *"yes little bit I got confused because of the repetition of question with the same image but then I thought its not the case to confuse myself so yes I feel that this image is real."* and *"No. Apparently the social media thing around the image is supposed to look fake? This survey could use a back button or some instructions..."*. Some students were not sure what was meant by malicious intent, which caused confusion as is explained in this answer: *"Though I don't know what the framing of "malicious intent" which the survey embodies really."*.

Another confusion was caused by the layout. Multiple students expressed their confusion of the platform layout even when the instructions told to assess only the authenticity of the image and not the layout itself. Thinking about the layout as part of the image caused some students to classify images with context as deepfakes as in these answers: *"The image tries to imitate the Instagram app, but this is not*

how the Instagram mobile application displays." and *"The first answer was that the image was not authentic. The 2nd and 3rd I picked deepfake because I'm not sure what to analyze: the image or the Temu looking version on Twitter or Instagram."*. Layout was also analyzed further than was intended by one student: *"If the hearts on the comments mean that the poster liked them, then the post is malicious as it quite obviously seems to want to direct people into questionable and unsecure links. However, if the poster didn't like them, then bots could have hijacked the comments to scam people, but overall, I think it depend on what the poster has done. But overall, probably malicious, to be honest."*. In this answer, the student had seen the filled hearts as liked comments even though the hearts were part of the layout. The layout should have been more clear and include only the needed aspects to show the contextual information. Another option would have been to use real layout from Instagram for example. Intent Label also caused confusion for some students, because it was not familiar feature from any popular existing social media platform and some students didn't know what it was supposed to be used for without any explanation.

Finally, it was noticed that some students did not remember what choices they made in the previous steps of the survey and as the going backward functionality was not in use, they couldn't check their answers as is explained in this answer: *"I don't remember. I think I changed and it was because I saw the number of likes."*. At least one student mentioned being tired with the survey and that caused forgetting whether the answer changed or not: *"I don't remember. I am mentally checked out at this point."*. Possibility of selecting answers by accident and doing a mistake in the survey was also mentioned by one student: *"not sure if i changed, if i did that was by accident"*. Survey was designed to not cause cognitive overload, but having many repetitive tasks seemed to be frustrating and tiresome for some students as in this answer: *"Blurry image, I'm starting to get frustrated with this task"*. Other

research papers [51] and [53] have also noticed that cognitive strain might have affected their results. Losing focus of the detection task could have caused some error in the results especially for the questions that were presented last.

5.3 Limitations

This section discusses the limitations of the experiment and of this thesis. As explained in Chapter 3, results of this thesis are not generalizable as the sample chosen consisted of only university cybersecurity students of one course. To have generalizable results, participants should be chosen from more diverse group of people, as deepfakes can affect almost everyone using internet and social media. Cybersecurity students might be more familiar with deepfakes and AI than people with no technical background, and that might have affected the results even though the overall accuracy was similar to existing human deepfake detection research and not significantly better. This thesis did not consider challenges such as cognitive decline in memory or low vision, that are more common with older people but can also affect people in all age groups.

It should be noted that the course the students took was mandatory for their degree, and the survey was given as mandatory assignment. There is a possibility that some participants were not doing their best in the detection task. Even though the survey was designed to not cause cognitive overload, fatigue and confusion was observed from the qualitative data. Experiment setup was not supervised or controlled, and some students used tools even though it was forbidden in the instructions. Because students were doing the survey with their own devices, it could have been meaningful to collect data of what devices they were using. As many students were using visual cues in the detection task, the device they used could also have affected the results. Because usage of tools was forbidden, same image was repeated multiple times and the students were told about possibility of deepfakes,

the experiment setup was not fully corresponding to real-world situation.

Dataset created for the experiment included a diverse set of images and contextual cues. However, this means that the dataset was not any benchmark or commonly used dataset and could have errors or bias caused by using OpenFake as its base dataset, despite being cautious and considering diversity when creating it. Having 20 unique images in the deepfake detection task, 5 in each group of different image and context pairings, is a low amount of images and therefore each image has more impact on the results than they would have if the total amount of unique images was higher. This means that if there are images that have errors or are much easier to detect than other images, they can distort the overall results. It should be noted that the dataset included only images, but other types of deepfakes such as videos are often encountered on social media and should also be considered.

6 Conclusion

This thesis explored the impact of contextual information on human deepfake detection accuracy with a systematic literature review and survey-based experimental study gathering quantitative and qualitative data. It contributes to the field of human deepfake detection by proposing and evaluating a new human deepfake detection improvement strategy based on contextual information in social media posts. Intent Labeling was proposed as a new human deepfake detection improvement strategy focusing on decreasing cognitive overload of users on social media platforms and helping them to focus on posts that are meaningful for further cognitive reflection.

Research questions one and two were answered by summarizing findings from the systematic literature review. Different types of contextual information in social media posts were gathered and categorized into seven categories: perceived quality, external knowledge, source and cross-references, intent, networked context, platform context and metadata. Contextual information related to trustworthiness included that everything in the post corresponded to viewer's expectations, the source was popular with high amount of followers and human-like interactions, the account creation date or the date the post was published was before there were AI models capable for generating high-quality deepfakes and that the image, description or interactions did not have any clear artifacts. Contextual information related to malicious intent included that something in the post was not corresponding to expectations of the viewer, different cues were inconsistent, source had low number of

followers and interactions or was newly created, post was of low quality and that the source was posting about topics such as giveaways, cryptocurrencies, pornography, the war in Ukraine or debates on COVID-19 and vaccinations.

Existing human deepfake detection improvement strategies included developing assistive tools, training, raising awareness and content labels and warnings. Existing improvement strategies have challenges of being dependent on the accuracy of technical deepfake detection solutions and needing to be constantly updated to correspond to fast improvement of deepfake generation models. A new human deepfake detection improvement strategy, Intent Labeling, was designed to avoid these challenges but it can be debated if this approach is only trying to dodge the problem of low human deepfake detection accuracy and not improving it.

Research question three was "Does contextual information increase, decrease or have no impact on the deepfake detection accuracy of cybersecurity students?". Survey-based experimental study was conducted with sample size N=73 as within-subjects study testing human deepfake detection accuracy in three conditions: images without any context, images with social media context and images with social media context and Intent Label. Context given for each unique image was created to be either trustworthy or suspicious. Based on the results, especially suspicious context has an impact on the accuracy. It significantly increases the accuracy when presented with deepfake and decreases it when presented with authentic image. These findings suggest that focusing on contextual cues can improve human deepfake detection accuracy in real-world situations, where the context is authentic. However, it should be noticed that in this study some of the participants were using tools and search engines, have not understand the survey design and have been affected by cognitive overload due to the survey design. These limitations might have affected the results.

In addition to quantitative data, the survey-based experimental study gathered

insights of strategies and cues that university level cybersecurity students were using when doing a deepfake detection task. Thematic analysis presents detailed findings of the cues students were using. Results show that cybersecurity students were using cues from every contextual cue category. Visual details, quality, consistency, source and interactions were such cues that students used in the detection task. Students were using multiple contextual cues, comparing many different cues and using their intuition in the task. It was noticed that same cues were sometimes used to classify the image as authentic but also as a deepfake. This suggests that people are focusing on different aspects even when they are seeing the same cues and they have different intuitions. As previously explained in Section 2.3.3, paper [52] presents three strategies: media-based, knowledge-based and search-based strategies that are used when detecting deepfake videos. These categories correspond well to results of this thesis, even though [52] did not consider contextual cues in addition to the video. These strategies could be expanded to include analyzing the source and interactions in addition to the actual content.

When testing how students would use given Intent Labels, it was observed that there was no significant difference between the accuracy of conditions where Intent Label was given and where it wasn't. From qualitative data, it was noticed that some students were focusing on intent when they were doing the task, and considered knowing the intent important. Despite that, Intent Labels were not seen very useful. In future work, Intent Labels could be further developed and tested with more specific experiment design, as the survey created in this thesis was designed to get more general information about the use of contextual cues and not only on Intent Labeling, which caused confusion and fatigue for some students. It would have been better to explain the meaning of the Intent Label before the experiment, as they are not a common feature in popular social media platforms. Results of thematic analysis show that cognitive reflection is important to improve human deepfake

detection accuracy, which means the idea of Intent Labeling could still be relevant if it is further developed in the future.

In conclusion, contextual information should be considered when researching human deepfake detection corresponding to real-world situations, as there will be some kind of context included and it has an impact on the accuracy. In future work, contextual cues should be researched with generalizable sample to get more insights of how age and background can affect the results. When creating new improvement strategies or improving existing strategies, limitations of cognitive decline in memory and low vision should be addressed, because these aspects are not enough considered in current research of human deepfake detection. Results from this thesis can be a starting point for further research of how contextual information could be used in improving human deepfake detection. Emphasizing cognitive reflection could be the key to increase human deepfake detection accuracy. Using context is critical for that as context can provide more cues that can be used to make a correct decision. It has been noted that cognitive overload is a problem related to tasks requiring cognitive reflection, and if the idea of Intent Labels is further developed, it could still help reduce that cognitive overload, and therefore help improving human deepfake detection.

References

[1] J. Twomey, D. Ching, M. Peter Aylett, M. Quayle, C. Linehan, and G. Murphy, "What is so deep about deepfakes? a multi-disciplinary thematic analysis of academic narratives about deepfake technology", *IEEE Transactions on Technology and Society*, vol. 6, no. 1, pp. 64–79, 2025. DOI: 10.1109/TTS.2024.3493465.

[2] S. D. Bray, S. D. Johnson, and B. Kleinberg, "Testing human ability to detect 'deepfake' images of human faces", *Journal of Cybersecurity*, vol. 9, no. 1, tyad011, Jun. 2023, ISSN: 2057-2085. DOI: 10.1093/cybsec/tyad011.

[3] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation", *Heliyon*, vol. 11, no. 3, e42273, 2025, ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2025.e42273.

[4] M. A. Taha et al., "Emerging threat of deep fake: How to identify and prevent it", in *Proceedings of the 6th International Conference on Future Networks & Distributed Systems*, ser. ICFNDS '22, Tashkent, TAS, Uzbekistan: Association for Computing Machinery, 2023, pp. 645–651, ISBN: 9781450399050. DOI: 10.1145/3584202.3584300.

[5] S. R, A. K. T A, A. Jayaraj, A. Snil, C. M. Varghese, and H. Lakshman, "A comprehensive analysis on web-based deep fake detection techniques", in *2024 Second International Conference on Intelligent Cyber Physical Systems and*

Internet of Things (ICoICI), 2024, pp. 888–891. DOI: 10.1109/ICoICI62503.2024.10696708.

[6] N. Soares, S. Seiden, I. Baggili, and A. Webb, "On the application of synthetic media to penetration testing", in *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, ser. WDC '23, Melbourne, VIC, Australia: Association for Computing Machinery, 2023, pp. 1–10, ISBN: 9798400702037. DOI: 10.1145/3595353.3595886.

[7] F. Sharevski, A. Zeidieh, J. V. Loop, and P. Jachim, "Blind and low-vision individuals' detection of audio deepfakes", in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24, Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 4867–4881, ISBN: 9798400706363. DOI: 10.1145/3658644.3690305.

[8] D. Ching, J. Twomey, C. Linehan, and G. Murphy, "Encountering deepfakes: A thematic analysis of comments on a music video featuring deepfake content", *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. 7, Oct. 2025. DOI: 10.1145/3757406.

[9] R. A. Frick, "Towards explainable and robust deepfake detection and attribution: Enhancing multimedia forensics for the next generation of synthetic media", in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '25, Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4863–4865, ISBN: 9798400715259. DOI: 10.1145/3719027.3765570.

[10] M. Pleskach, *Legal loopholes in the eu artificial intelligence act: Addressing disinformation and human rights protection – the need for further refinement*, 2025. DOI: 10.1109/ACIT65614.2025.11185881.

[11] J. Ricker, D. Assenmacher, T. Holz, A. Fischer, and E. Quiring, "Ai-generated faces in the real world: A large-scale case study of twitter profile images", in *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, ser. RAID '24, Padua, Italy: Association for Computing Machinery, 2024, pp. 513–530, ISBN: 9798400709593. DOI: 10.1145/3678890.3678922.

[12] D. Cooke, A. Edwards, S. Barkoff, and K. Kelly, "As good as a coin toss: Human detection of ai-generated content", *Commun. ACM*, vol. 68, no. 10, pp. 100–109, Sep. 2025, ISSN: 0001-0782. DOI: 10.1145/3729417.

[13] A. Diel, T. Lalgi, I. C. Schröter, K. F. MacDorman, M. Teufel, and A. Bäuerle, "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers", *Computers in Human Behavior Reports*, vol. 16, p. 100 538, 2024, ISSN: 2451-9588. DOI: https://doi.org/10.1016/j.chbr.2024.100538.

[14] N. Kamali, K. Nakamura, A. Kumar, A. Chatzimparmpas, J. Hullman, and M. Groh, "Characterizing photorealism and artifacts in diffusion model-generated images", in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, Yokohama, Japan: Association for Computing Machinery, 2025, ISBN: 9798400713941. DOI: 10.1145/3706598.3713962.

[15] K. Warren et al., ""better be computer or i'm dumb": A large-scale evaluation of humans as audio deepfake detectors", in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24, Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 2696–2710, ISBN: 9798400706363. DOI: 10.1145/3658644.3670325.

[16] S. Uesugi, "Fostering critical thinking on social media: Combating ai-generated fake posts upon natural disasters", in *Proceedings of the 2024 11th Multidis-*

ciplinary International Social Networks Conference, ser. MISNC '24, Bali, Indonesia: Association for Computing Machinery, 2024, pp. 73–80, ISBN: 9798400717550. DOI: 10.1145/3675669.3675681.

[17] K. Somoray, D. J. Miller, and M. Holmes, "Human performance in deepfake detection: A systematic review", *Human Behavior and Emerging Technologies*, vol. 2025, no. 1, p. 1 833 228, 2025. DOI: https://doi.org/10.1155/hbe2/1833228.

[18] H. H. Le, V. S. T. Nguyen, T. L. C. Dang, V. T. K. Nguyen, T. T. H. Nguyen, and H. Cao, "Multimedia verification through multi-agent deep research multimodal large language models", in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25, Dublin, Ireland: Association for Computing Machinery, 2025, pp. 14 034–14 040, ISBN: 9798400720352. DOI: 10.1145/3746027.3762033.

[19] I. N. Sherman, J. W. Stokes, and E. M. Redmiles, "Designing media provenance indicators to combat fake media", in *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses*, ser. RAID '21, San Sebastian, Spain: Association for Computing Machinery, 2021, pp. 324–339, ISBN: 9781450390583. DOI: 10.1145/3471621.3471860.

[20] J. Frank et al., "A representative study on human detection of artificially generated media across countries", in *2024 IEEE Symposium on Security and Privacy (SP)*, 2024, pp. 55–73. DOI: 10.1109/SP54263.2024.00159.

[21] K. Meding and C. Sorge, "What constitutes a deep fake? the blurry line between legitimate processing and manipulation under the eu ai act", in *Proceedings of the 2025 Symposium on Computer Science and Law*, ser. CSLAW '25, Munich, Germany: Association for Computing Machinery, 2025, pp. 152–159, ISBN: 9798400714214. DOI: 10.1145/3709025.3712218.

[22] A. Ayub Khan et al., "Digital forensics for the socio-cyber world (df-scw): A novel framework for deepfake multimedia investigation on social media platforms", *Egyptian Informatics Journal*, vol. 27, p. 100 502, 2024, ISSN: 1110-8665. DOI: https://doi.org/10.1016/j.eij.2024.100502.

[23] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection", *ACM Comput. Surv.*, vol. 57, no. 7, Feb. 2025, ISSN: 0360-0300. DOI: 10.1145/3714458.

[24] A. K. Saini, G. M. Upadhyay, and P. Vats, "Unmasking reality: A comprehensive review of deepfake detection techniques and their evolving landscape", in *2025 International Conference on Intelligent Control, Computing and Communications (IC3)*, 2025, pp. 311–318. DOI: 10.1109/IC363308.2025.10957379.

[25] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges", *Alexandria Engineering Journal*, vol. 125, pp. 265–277, 2025, ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2025.04.007.

[26] K. Sree Kammari et al., "A comprehensive review of deepfake detection techniques utilizing remote photoplethysmography", *IEEE Access*, vol. 13, pp. 183 557–183 578, 2025. DOI: 10.1109/ACCESS.2025.3624284.

[27] S. Jadhav, M. Bartere, and S. Patil, "Review of deep fake detection using deep learning convolutional, recurrent, and graph networks", in *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, 2024, pp. 1–7. DOI: 10.1109/ICAIQSA64000.2024.10882268.

[28] S. Sadiq, T. Aljrees, and S. Ullah, "Deepfake detection on social media: Leveraging deep learning and fasttext embeddings for identifying machine-generated tweets", *IEEE Access*, vol. 11, pp. 95 008–95 021, 2023. DOI: 10.1109/ACCESS.2023.3308515.

[29] W. Wimalasena, H. Herath, and I. Hewapathirana, "A systematic literature review of deepfake face image detection with transfer learning techniques", in *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 2025, pp. 1–7. DOI: 10.1109/SCSE65633.2025.11031009.

[30] E. Mihăilescu and D. F. Chiper, "Watermarking techniques for content integrity verification, tamper detection and forensics in synthetic media", in *2025 International Symposium on Signals, Circuits and Systems (ISSCS)*, 2025, pp. 1–4. DOI: 10.1109/ISSCS66034.2025.11105620.

[31] D. L. R and B. B. Sujitha, "Advancements in deepfake detection: A comprehensive review of ai-driven approaches", in *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*, 2025, pp. 1007–1011. DOI: 10.1109/ICMLAS64557.2025.10967658.

[32] B. Zegeye, H. Atinafu, S. Sherif, R. Dave, and M. Bhavsar, "Deepfake detection using machine learning: A comprehensive literature review", in *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, 2025, pp. 1–7. DOI: 10.1109/ACDSA65407.2025.11166436.

[33] S. Ren et al., "Do deepfake detectors work in reality?", in *Proceedings of the 4th Workshop on Security Implications of Deepfakes and Cheapfakes*, ser. WDC '25, Hanoi, Vietnam: Association for Computing Machinery, 2025, pp. 21–26, ISBN: 9798400714191. DOI: 10.1145/3709022.3736545.

[34] N. Pasupuleti and D. V. Lakshmi, "Deepfake methods and statistics: A comprehensive review", in *2025 5th International Conference on Expert Clouds and Applications (ICOECA)*, 2025, pp. 964–971. DOI: 10.1109/ICOECA66273.2025.00168.

[35] B. Biswas et al., "Navigating the obscured: A novel deepfake detection framework tackling trending occlusions in social media", in *2024 27th International Conference on Computer and Information Technology (ICCIT)*, 2024, pp. 2986–2991. DOI: 10.1109/ICCIT64611.2024.11021926.

[36] C. Kang, S. Jeong, J. Lee, D. Choi, S. S. Woo, and J. Han, "Hidf: A human-indistinguishable deepfake dataset", in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, ser. KDD '25, Toronto ON, Canada: Association for Computing Machinery, 2025, pp. 5527–5538, ISBN: 9798400714542. DOI: 10.1145/3711896.3737399.

[37] S. Yuan, J. Dong, and Y. Li, "Where the devil hides: Deepfake detectors can no longer be trusted", in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 8764–8774. DOI: 10.1109/CVPR52734.2025.00819.

[38] P. Liu, Q. Tao, and J. Zhou, "Robust deepfake detection by addressing generalization and trustworthiness challenges: A short survey", in *Proceedings of the 1st ACM Multimedia Workshop on Multi-Modal Misinformation Governance in the Era of Foundation Models*, ser. MIS '24, Melbourne VIC, Australia: Association for Computing Machinery, 2024, pp. 3–11, ISBN: 9798400712012. DOI: 10.1145/3689090.3689386.

[39] D. Singh, P. Singh, and R. Bhandari, "A comprehensive review of deepfake detection in advanced neural network architectures and deep learning strategies", in *2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*, 2024, pp. 1147–1152. DOI: 10.1109/GlobalAISummit62156.2024.10947978.

[40] A. Hou, L. Lin, J. Li, and S. Hu, "Rethinking individual fairness in deepfake detection", in *Proceedings of the 33rd ACM International Conference on*

Multimedia, ser. MM '25, Dublin, Ireland: Association for Computing Machinery, 2025, pp. 11 424–11 433, ISBN: 9798400720352. DOI: 10.1145/3746027.3755244.

[41] C. Chen, D. H.-L. Goh, H. R. Qiu, and C. Neo, "Generation z's fight against deepfake videos: A survey on identification strategies", *Proceedings of the Association for Information Science and Technology*, vol. 62, no. 1, pp. 892–896, 2025. DOI: https://doi.org/10.1002/pra2.1309.

[42] D. Prudký, A. Firc, and K. Malinka, "Assessing the human ability to recognize synthetic speech in ordinary conversation", in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2023, pp. 1–5. DOI: 10.1109/BIOSIG58226.2023.10346006.

[43] S. Ahmed, A. W. T. Bee, S. W. T. Ng, and M. Masood, "Social media news use amplifies the illusory truth effects of viral deepfakes: A cross-national study of eight countries", *Journal of Broadcasting & Electronic Media*, vol. 68, no. 5, pp. 778–805, 2024. DOI: 10.1080/08838151.2024.2410783.

[44] E. Bozkir, C. Riedmiller, A. N. Skodras, G. Kasneci, and E. Kasneci, "Can you tell real from fake face images? perception of computer-generated faces by humans", *ACM Trans. Appl. Percept.*, vol. 22, no. 2, Nov. 2024, ISSN: 1544-3558. DOI: 10.1145/3696667.

[45] R. Nichols, C. Rathgeb, P. Drozdowski, and C. Busch, "Psychophysical evaluation of human performance in detecting digital face image manipulations", *IEEE Access*, vol. 10, pp. 31 359–31 376, 2022. DOI: 10.1109/ACCESS.2022.3160596.

[46] N. B. Mohamed, G. Bogdanel, and H. G. Moreno, "Is training useful to detect deepfakes? : A preliminary study.", in *2023 18th Iberian Conference on In-*

formation Systems and Technologies (CISTI), 2023, pp. 1–5. DOI: 10.23919/CISTI58278.2023.10211622.

[47] H. Weigelt, E. Segev, G. Kurtz, O. Kahana, and N. R. Fogel, "Enhancing students' critical thinking literacy in a generative ai context: Eye movement patterns of deepfake detection", *Computers & Education*, vol. 244, p. 105 529, 2026, ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2025.105529.

[48] E. Preu, M. Jackson, and N. Choudhury, "Perception vs. reality: Understanding and evaluating the impact of synthetic image deepfakes over college students", in *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2022, pp. 0547–0553. DOI: 10.1109/UEMCON54665.2022.9965697.

[49] M. Ramon, M. Vowels, and M. Groh, "Deepfake detection in super-recognizers and police officers", *IEEE Security & Privacy*, vol. 22, no. 3, pp. 68–76, 2024. DOI: 10.1109/MSEC.2024.3371030.

[50] M. Högemann, J. Betke, and O. Thomas, "What you see is not what you get anymore: A mixed-methods approach on human perception of ai-generated images", *Frontiers in Artificial Intelligence*, vol. 8, 2025, ISSN: 2624-8212. DOI: 10.3389/frai.2025.1707336.

[51] C. Han, P. Mitra, and S. M. Billah, "Uncovering human traits in determining real and spoofed audio: Insights from blind and sighted individuals", in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, Honolulu, HI, USA: Association for Computing Machinery, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642817.

[52] D. H.-L. Goh, ""he looks very real": Media, knowledge, and search-based strategies for deepfake identification", *Journal of the Association for Infor-*

mation Science and Technology, vol. 75, no. 6, pp. 643–654, 2024. DOI: https://doi.org/10.1002/asi.24867.

[53] M. R. Khan et al., "Exploring neurophysiological responses to cross-cultural deepfake videos", in *Companion Publication of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23 Companion, Paris, France: Association for Computing Machinery, 2023, pp. 41–45, ISBN: 9798400703218. DOI: 10.1145/3610661.3617148.

[54] Z. R. Tidler and R. Catrambone, "Effects of neurodivergence on deepfake-video detection: Autism spectrum disorder", *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, vol. 13, no. 1, pp. 157–159, 2024. DOI: 10.1177/2327857924131020.

[55] E. Nas and R. de Kleijn, "Conspiracy thinking and social media use are associated with ability to detect deepfakes", *Telematics and Informatics*, vol. 87, p. 102 093, 2024, ISSN: 0736-5853. DOI: https://doi.org/10.1016/j.tele.2023.102093.

[56] M. Appel and F. Prietzel, "The detection of political deepfakes", *Journal of Computer-Mediated Communication*, vol. 27, no. 4, zmac008, Jul. 2022, ISSN: 1083-6101. DOI: 10.1093/jcmc/zmac008.

[57] A. Batra, J. Khemani, A. Gumber, A. Kumar, A. Jain, and S. Gupta, "Socialdf: Benchmark dataset and detection model for mitigating harmful deepfake content on social media platforms", in *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, ser. MAD' 25, Association for Computing Machinery, 2025, pp. 81–89, ISBN: 9798400718915. DOI: 10.1145/3733567.3735573.

[58] S. S. Prasad, O. Hadar, T. Vu, and I. Polian, "Human vs. automatic detection of deepfake videos over noisy channels", in *2022 IEEE International Conference*

on Multimedia and Expo (ICME), 2022, pp. 1–6. DOI: 10.1109/ICME52920.2022.9859954.

[59] A. Boutadjine, F. Harrag, and K. Shaalan, "Human vs. machine: A comparative study on the detection of ai-generated content", *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 24, no. 2, Feb. 2025, ISSN: 2375-4699. DOI: 10.1145/3708889.

[60] J. Mink et al., "It's trying too hard to look real: Deepfake moderation mistakes and identity-based bias", in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, Honolulu, HI, USA: Association for Computing Machinery, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3641999.

[61] Y. Zhai, X. Xue, Z. Guo, T. Jin, Y. Diao, and J. Jeung, "Hear us, then protect us: Navigating deepfake scams and safeguard interventions with older adults through participatory design", in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, Association for Computing Machinery, 2025, ISBN: 9798400713941. DOI: 10.1145/3706598.3714423.

[62] S. J. Sohrawardi, Y. K. Wu, A. Hickerson, and M. Wright, "Dungeons & deepfakes: Using scenario-based role-play to study journalists' behavior towards using ai-based verification tools for video content", in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, Honolulu, HI, USA: Association for Computing Machinery, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3641973.

[63] A. S. Walker, "Preparing students for the fight against false information with visual verification and open source reporting", *Journalism & Mass Communication Educator*, vol. 74, no. 2, pp. 227–239, 2019. DOI: 10.1177/1077695819831098.

[64] F. Shahid, S. Kamath, A. Sidotam, V. Jiang, A. Batino, and A. Vashistha, ""it matches my worldview": Examining perceptions and attitudes around fake videos", in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391573. DOI: 10.1145/3491102.3517646.

[65] M. Boler, H. Gharib, Y.-J. Kweon, A. Trigiani, and B. Perry, "Promoting mis-/disinformation literacy among adults: A scoping review of interventions and recommendations", *Communication Research*, vol. 0, no. 0, p. 00 936 502 251 318 630, 0. DOI: 10.1177/00936502251318630.

[66] J. Walker, G. Thuermer, J. Vicens, and E. Simperl, "Ai art and misinformation: Approaches and strategies for media literacy and fact checking", in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23, Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 26–37, ISBN: 9798400702310. DOI: 10.1145/3600211.3604715.

[67] D. Gamage, D. Sewwandi, M. Zhang, and A. K. Bandara, "Labeling synthetic content: User perceptions of label designs for ai-generated content on social media", in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, Association for Computing Machinery, 2025, ISBN: 9798400713941. DOI: 10.1145/3706598.3713171.

[68] C. Wittenberg, Z. Epstein, G. Péloquin-Skulski, A. J. Berinsky, and D. G. Rand, "Labeling ai-generated media online", *PNAS Nexus*, vol. 4, no. 6, pgaf170, May 2025, ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgaf170.

[69] G. Huang and B. Hu, ""a warning is not enough. teach me how to spot deepfakes.": Testing media literacy interventions for combating deepfakes", *Science*

Communication, vol. 0, no. 0, p. 10 755 470 251 382 889, 0. DOI: 10.1177/10755470251382889.

[70] "Ieee standard for transparent human and machine agency identification", *IEEE Std 3152-2024*, pp. 1–42, 2025. DOI: 10.1109/IEEESTD.2025.10998963.

[71] C. Guo, N. Zheng, and C. (Guo, "Seeing is not believing: A nuanced view of misinformation warning efficacy on video-sharing social media platforms", *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW2, Oct. 2023. DOI: 10.1145/3610085.

[72] X. Jin et al., "Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics", *New Media & Society*, vol. 27, no. 3, pp. 1651–1672, 2025. DOI: 10.1177/14614448231199664.

[73] M. Choudhary, S. S. Chouhan, and S. S. Rathore, "Beyond text: Multimodal credibility assessment approaches for online user-generated content", *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 5, Nov. 2024, ISSN: 2157-6904. DOI: 10.1145/3673236.

[74] C. Chartier, A. Watt, O. Lin, A. Chandawarkar, J. Lee, and E. Hall-Findlay, "Breastgan: Artificial intelligence-enabled breast augmentation simulation", *Aesthetic Surgery Journal Open Forum*, vol. 4, ojab052, Dec. 2021, ISSN: 2631-4797. DOI: 10.1093/asjof/ojab052.

[75] M. Khamis, R. Panskus, H. Farzand, M. Mumm, S. Macdonald, and K. Marky, "Perspectives on deepfakes for privacy: Comparing perceptions of photo owners and obfuscated individuals towards deepfake versus traditional privacy-enhancing obfuscation", in *Proceedings of the 23rd International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '24, Association for Computing Machinery, 2024, pp. 300–312, ISBN: 9798400712838. DOI: 10.1145/3701571.3701602.

[76] T. Sanchez, "Examining the text-to-image community of practice: Why and how do people prompt generative ais?", in *Proceedings of the 15th Conference on Creativity and Cognition*, ser. C&C '23, Virtual Event, USA: Association for Computing Machinery, 2023, pp. 43–61, ISBN: 9798400701801. DOI: 10.1145/3591196.3593051.

[77] P. Pataranutaporn, C. Archiwaranguprok, S. W. T. Chan, E. Loftus, and P. Maes, "Synthetic human memories: Ai-edited images and videos can implant false memories and distort recollection", in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, Association for Computing Machinery, 2025, ISBN: 9798400713941. DOI: 10.1145/3706598.3713697.

[78] A. Skoularikis, S.-I. Papadopoulos, S. Papadopoulos, and P. C. Petrantonakis, "'humor, art, or misinformation?': A multimodal dataset for intent-aware synthetic image detection", in *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, ser. DHOW '25, Ireland: Association for Computing Machinery, 2025, pp. 95–104, ISBN: 9798400720574. DOI: 10.1145/3746275.3762215.

[79] B. Kasthuriarachchy, M. Chetty, G. Karmakar, and D. Walls, "Pre-trained language models with limited data for intent classification", in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–9. DOI: 10.1109/IJCNN48605.2020.9207121.

[80] M. Hameleers, T. G. van der Meer, and T. Dobber, "Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes", *Computers in Human Behavior*, vol. 152, p. 108 096, 2024, ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2023.108096.

[81] S. S. Sundar, M. D. Molina, and E. Cho, "Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?", *Journal of Computer-Mediated Communication*, vol. 26, no. 6, pp. 301–319, Aug. 2021, ISSN: 1083-6101. DOI: 10.1093/jcmc/zmab010.

[82] M. Hameleers, "Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting", *International Journal of Public Opinion Research*, vol. 36, no. 1, edae004, Mar. 2024, ISSN: 1471-6909. DOI: 10.1093/ijpor/edae004.

[83] E. Jang, H. M. Lee, S. Lee, Y. Jung, and S. S. Sundar, "Too good to be false: How photorealism promotes susceptibility to misinformation", in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '25, Association for Computing Machinery, 2025, ISBN: 9798400713958. DOI: 10.1145/3706599.3719796.

[84] B. Hu and G. Huang, "What looks good and familiar seems real: How heuristic processing of ai-powered synthetic and real videos shapes user perceptions and detection accuracy", *Social Media + Society*, vol. 11, no. 4, p. 20 563 051 251 401 815, 2025. DOI: 10.1177/20563051251401815.

[85] I. Hilmansyah, D. Wiryawan, and H. Hartono, "Analysis of student trust in the credibility of deepfake video information on social media", in *2024 9th International Conference on Information Technology and Digital Applications (ICITDA)*, 2024, pp. 1–6. DOI: 10.1109/ICITDA64560.2024.10810015.

[86] X. Hu, Z. Guo, J. Chen, L. Wen, and P. S. Yu, "Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media", in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23, Taipei, Taiwan: Association

for Computing Machinery, 2023, pp. 2901–2912, ISBN: 9781450394086. DOI: 10.1145/3539618.3591896.

[87] R. A. Frick and T. Kwon, "Improved estimation of check-worthy social media content using the analysis of commentaries", in *Proceedings of the 1st ACM Workshop on Deepfake, Deception, and Disinformation Security*, ser. 3D-Sec '25, Association for Computing Machinery, 2025, pp. 16–21, ISBN: 9798400719028. DOI: 10.1145/3733813.3764367.

[88] M. Umair, A. Bouguettaya, A. Lakhdari, M. Ouzzani, and Y. Liu, "Exif2vec: A framework to ascertain untrustworthy crowdsourced images using metadata", *ACM Trans. Web*, vol. 18, no. 3, Apr. 2024, ISSN: 1559-1131. DOI: 10.1145/3645094.

[89] V. Livernoche et al., *Openfake: An open dataset and platform toward real-world deepfake detection*, 2025. arXiv: 2509.09495 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2509.09495.