Dungeons & Deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content

Saniat Javid Sohrawardi
saniat.s@mail.rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Y. Kelly Wu
kellywu@mail.rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Andrea Hickerson
andreah@olemiss.edu
University of Mississippi
Oxford, Mississippi, USA

Matthew Wright
matthew.wright@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

ABSTRACT

The evolving landscape of manipulated media, including the threat of deepfakes, has made information verification a daunting challenge for journalists. Technologists have developed tools to detect deepfakes, but these tools can sometimes yield inaccurate results, raising concerns about inadvertently disseminating manipulated content as authentic news. This study examines the impact of unreliable deepfake detection tools on information verification. We conducted role-playing exercises with 24 US journalists, immersing them in complex breaking-news scenarios where determining authenticity was challenging. Through these exercises, we explored questions regarding journalists' investigative processes, use of a deepfake detection tool, and decisions on when and what to publish. Our findings reveal that journalists are diligent in verifying information, but sometimes rely too heavily on results from deepfake detection tools. We argue for more cautious release of such tools, accompanied by proper training for users to mitigate the risk of unintentionally propagating manipulated content as real news.

CCS CONCEPTS

• **Applied computing** → **Investigation techniques**; • **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

deepfake, journalism, verification, role-play

ACM Reference Format:
Saniat Javid Sohrawardi, Y. Kelly Wu, Andrea Hickerson, and Matthew Wright. 2024. Dungeons & Deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3641973

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3641973

1 INTRODUCTION

In March 2019, Voice of Myanmar News published a video of the Chief Minister of Yangon, Phyo Min Thein, confessing to bribing the deposed State Counselor Aung San Suu Kyi. The video made rounds on Myanmar social media accounts as likely being doctored [48, 86]. Indeed, publicly available deepfake detection tools indicated that the video was fake, with one showing 98% confidence [71]. Soon after, expert analysis contradicted this analysis, pointing out that even though the confession itself may well have been forced, the video of it was unlikely to have been a deepfake [40]. Compression artifacts in the published video could have confused the detection tool, giving the users a false verdict. This incident illustrates a disconnect within the current ecosystem of manipulated videos and the use of Artificial Intelligence (AI) systems used to detect them.

Deepfakes are videos manipulated using deep learning (DL) technology so that a person could be shown saying things they never said. Although recently the term has evolved to additionally encompass audio and image manipulations, the focus in our work is on videos. The current publicly available deepfake generation techniques [47, 58, 67, 89, 89, 97] can yield convincing, high-quality results. A tech-savvy individual could learn how to use these techniques from online forums. Then with a high-powered consumer GPU, like the ones used for gaming, they could produce a very high-quality deepfake in just a few weeks. Since deepfakes can be used to make anyone say anything the creator wants, the potential to create misinformation is clear.

Many research efforts have been made in the past few years to detect deepfakes [1–4, 17–19, 22, 26, 42, 54, 57, 64, 68, 81, 82, 93], and several tools have been deployed [5, 24, 65, 101]. Most of the works report very high accuracy, with some reaching 100% on curated research datasets.

Unfortunately, current deep learning-based systems are unstable in open-world scenarios [20, 23, 55, 80], which also applies to deepfake detection systems [27]. This issue has led to very confident false-positive or false-negative deepfake detection results in deployed tools. Additionally, researchers have shown that it is possible to generate *adversarial examples* to trick deepfake detection models into providing inaccurate results [44, 46, 79, 95]. Errors like these could mislead users, creating false suspicion about real videos

– as with the video of Phyo Min Thein – or misplaced confidence that a fake video is authentic.

Journalists have been adopting various tools to help them verify the information and the authenticity of video, audio, and images [10, 11, 53]. Large media organizations typically have a variety of protocols in place for the thorough analysis of each type of medium. Given that deepfakes are still a new manipulation threat, however, there are very few reliable detection tools. Prior research [10, 102] indicates that journalists would use video verification tools as a part of their arsenal to aid them in their work. Unlike traditional verification tools, however, deepfake detection tools rely on potentially unstable deep-learning methods and output unreliable confidence estimates. Such tools thus create new challenges for journalists that other tools did not, and the impacts of these challenges have not previously been studied. Moreover, the current detection tools imitate the workflow of anti-virus software, providing positive versus negative detection results, which may not be ideal for the use case of media forensics. Thus, it is very important to study how journalists would interpret results from this new generation of tools for breaking news in a typical newsroom setting given the pressure for fast publication.

Thus, in order to better understand the potential contributions of deepfake detection tools to the journalistic workflows, our work aims to address three aspects of their usage:

(1) **Perception.** There is a lack of research into journalists' perceptions of deepfake detection tools, which have become increasingly available over the past three years. Given that these tools are new and different from traditional verification tools, it is important to understand journalists' readiness to adopt them. Thus we seek to understand the journalists' perception towards deepfake detection tools **(RQ1)**.
(2) **Workflow changes.** The potential effects and the importance of the deepfake detection tools may vary depending on when the journalists choose to use them. Literature on psychology and AI tools points to anchoring bias [90] as a cognitive bias where people may favor information they received early in the decision-making process. Hence, we want to observe when the journalists would opt to employ deepfake detection tools and why **(RQ2)**.
(3) **Overreliance.** The Myanmar incident showed how unreliable deepfake detection tools can cause serious harm by misleading users. Previous studies have also pointed out the risk of automation bias [62], where users tend to trust the output of decision-making tools too much. This could have disastrous effects if journalists publish fake videos as real news or dismiss real videos as fake. Therefore, we want to investigate how much journalists rely on the tools **(RQ3)**.

We wanted to design a study that would let journalists verify information in their own way, since different journalists have different verification methods [99]. To achieve that, typical Scenario-Based Design [15] would have been sufficient. Thus we took inspiration from the immersive and flexible environments provided in Dungeons & Dragons tabletop game, to design a semi-structured scenario-based role-playing exercise. Through this methodology, participating journalists are placed in a high-stakes newsroom scenario and asked to verify content containing a video with access to deep-learning-based deepfake detection tools and most importantly, free reign to create their own actions. While this greatly increases the complexity of the study, it allows the participants to act out their behaviors in a more natural environment, as opposed to structured questions where their answers may have less chances to align with their true behavior. To the best of our knowledge, this is the first research study of its kind to assess the verification behaviors of journalists interacting with AI tools. This study design allowed us to carry out the study in both online and in-person settings and remain in control of the interactions.

Through our interviews with 24 US journalists, our key findings were:

- Journalists expressed positive views of deepfake detection tools and looked forward to using them, but wanted more explanations of the results. Their trust in the tools may depend on the perceived reputation of the developers.
- Journalists tended to rely more on traditional verification methods first to establish context, and turned to deepfake detection tools when contextual verification was difficult or time was limited. Tools were used more when videos made very bold/unusual claims.
- Events with high social or political impact encouraged more verification steps from journalists, while urgent events sometimes led them to skip steps in favor of a quicker publication.
- A few journalists showed possible overreliance on the tools, especially when results confirmed their initial impressions, due to various biases. The trust in the tool seems to depend on their previous experiences with it. Thus while a positive experience may increase overreliance, bad experiences may instill diligence.
- The scenario-based role-play methodology proved engaging for participants and shows promise for training journalists on issues around manipulated video and improving verification skills.

The findings reveal the importance of improving the explainability of deepfake detection tools and training journalists on their nuanced interpretation as these AI assistance tools see wider adoption.

2 LITERATURE REVIEW

2.1 Deepfakes

Deepfakes, a portmanteau of *deep learning* and *fake* [105], first came into the spotlight in 2017. At the time of writing, the majority of fake videos on the web are either non-consensual pornography [77] or entertainment parodies. These examples draw a significant amount of negative attention to this technology. There are also positive uses, however, in arts, education, and therapy, including the Dali Museum interactive exhibit [61] and the David Beckham multilingual malaria campaign [83].

The focus of our work is on the malicious use of this technology to create and spread misinformation. The potential to destabilize our societal norms with manipulated video and audio content has been a point of discussion in various fields [16, 39, 51, 77, 108]. Silbey and Hartzog suggest that deepfakes provide an opportunity to reform our education system, journalism, and democratic systems to become more resilient to fake media [98].

The Russia-Ukraine conflict in 2022 saw the first use of malicious deepfakes created to affect a war. Two fake videos were circulated online, one each of the two opposing presidents announcing that their country was to surrender [35, 36]. More recently, deepfakes have been used in the 2024 U. S. Presidential campaign, both by unknown entities and by the campaigns themselves [73, 106]. Even the mere existence of deepfakes can cause erosion of trust and lead to the spread of misinformation. In Gabon in 2019, armed soldiers who believed that the New Year's presidential address video was a deepfake attempted a coup, even though the video was probably authentic [14]. The Myanmar video we mention in the introduction is also a recent example of this effect.

2.2 Verification Systems in Journalism

Journalists and independent fact-checkers have long used various tools to verify information. Brandztaeg et al. [10, 11] studied the practices and perceptions of a young generation of European journalists towards verification tools. They mentioned tools and services available for photo and video verification such as Google Image Search, TinEye, Exif, Topsy, Tungstene, Google Maps, Streetview, YouTube videos, and Storyful. Many of these tools were also mentioned in a study on the requirements of deepfake detection tools by Sohrawardi et al. [102]. All three studies involved a semi-structured interview format, which appeals to the conversational behavior of journalists. One of the studies by Brandtzaeg et al. [10] included social media users alongside journalists, and both groups had mixed feelings about online fact-checking tools. A very recent survey by Khan et al. [53] listed some of the more prevalent tools in the journalism world and suggested their limitations. In contrast, our study aims to first observe the behavior of journalists when presented with a task rather than solely relying on questions and answering.

For video verification, the InVID Project [104] is a well-known tool that allows users to reverse search frames, alter saturation and brightness, and view the impact of the video on social media. It is often used to verify the context or find original videos, but it does not possess deepfake detection capabilities. For deepfake detection, publicly accessible options are Deepfake-o-meter [65], Deepware [24], Sensity [5], and Reality Defender [25]. A more restricted and user-centered option is the DeFake tool [102], which was a result of academic research with journalists and is only available for that target population alongside researchers. Since their interface was designed with journalists in mind, it made sense to use it for the studies. We recreated the interface prototypes from the DeFake paper and altered them to match the context of the scenarios in our study.

2.3 Overreliance and Confirmation Bias in AI Tool Utilization

A pressing concern in the realm of deepfake detection tools is their potential to inadvertently increase confirmation bias [84]. This bias is characterized by the inclination to favor information that aligns with one's existing beliefs, often at the expense of contradictory evidence. Rastogi et al. [90] posited that such biases, coupled with overreliance, can arise when users stop looking for alternative evidence or viewpoints upon receiving machine-generated output. This observation is particularly relevant to our study.

Lee et al. [62] introduced the concept of automation bias [76] within the AI milieu, highlighting the pitfalls of suboptimal human-algorithm collaboration. Drawing parallels with Watson's rule discovery experiment [109], it's evident that forensic investigators [75] are not immune to confirmation bias. This susceptibility is glaringly evident in the realm of deepfake videos, as observed in public reactions to a manipulated video of President Trump [107].

Numerous studies in the field have explored AI-assisted disinformation detection [9]. Although these approaches introduce automation bias, a potential remedy lies in incorporating explanations, as demonstrated by Epstein et al. [34]. Their research revealed an enhanced ability to distinguish between genuine and fabricated news by American internet users. However, it is imperative to exercise caution in the handling of these explanations, as they have the potential to inadvertently promote biased behavior [90].

While Pennycook et al. [88] demonstrated that individuals with heightened analytical thinking are less prone to such biases, the Myanmar incident [86] underscored the vulnerability even among seasoned journalists. One plausible explanation for this behavior is the limited exposure or comprehension of the limitations inherent in AI-driven tools. A review of tools employed by journalists [53] reveals a scanty adoption of AI-based tools, with the few in operation being recent introductions.

Historical studies on fingerprint analysis [100, 103] have shown that pre-emptively providing participants with background information, or *priming* them, can inadvertently introduce bias. This suggests that even the most adept forensic experts are not entirely impervious to bias [59]. Byrd's work [13] offers insights into the myriad biases in forensics, which can be extrapolated to our context. In our research, our aim is to provide journalists with pertinent background information to discern if the context might inadvertently induce an overreliance on deepfake detection tools.

2.4 Dungeons & Dragons Tabletop Game

Dungeons & Dragons (D&D) is a popular tabletop role-playing game where one person serves as a Dungeon Master (DM) and others play as characters in medieval fantasy world. The DM creates and narrates the adventure, while the players openly decide what their characters would do and how they interact with the world. The game uses dice and rules to determine the outcomes of actions and events. The gameplay mechanics of D&D are based on the core rulebooks: the Player's Handbook, the Dungeon Master's Guide, and the Monster Manual [92]. These books contain, guides on how to create characters, as well as basic rules for combat, magic, and exploration.

The role of the DM is to be the game's lead storyteller and referee. The DM is responsible for preparing the adventure, describing the scenes and locations, playing the roles of non-player characters (NPCs) and monsters, adjudicating the rules, and keeping track of the game state. The DM also has the authority to improvise and modify the adventure as needed, depending on the players' choices and actions. The DM's goal is to make the game fun and engaging for everyone, while also challenging the players and creating a sense of immersion and wonder.

The immersive nature of the game, flexible interaction choices, and the role of the DM are the three core features we adopted from D&D to develop our user studies.

2.5 Software Analysis through Participation in Scenarios

Software analysis and design can be approached in myriad ways. Among these, role-play and scenario management stand out as interactive and often gamified methods. Parson [87] delineated four models where simulation gaming could guide decision-making in intricate policy issues. He highlighted the potential of simulations to either promote creativity and insights or to integrate knowledge. Our study resonates with the latter, as we aim to gauge the early implications of integrating AI-driven deepfake detection tools into journalistic processes.

Constructing semi-realistic scenario-based studies can be labor-intensive, but their value is undeniable, as evidenced by Eden et al. [29]. In qualitative role-play studies, typically conducted in groups, participants assume designated roles or personas and enact them throughout the study [28, 50]. Conversely, scenario management situates individual users within specific scenarios. Jarke et al. [49] offered a comprehensive review of scenario management across disciplines, including human-computer interaction. They championed a methodology for scenario development, underscoring the advantages of creativity, contextual awareness, and flexibility. Carroll [15] posited that scenarios should furnish a rich narrative, enabling participants to discern contextual implications. Munroe et al. [78] stressed the significance of high fidelity and minimal instructions to encourage genuine scenario interactions. They also highlighted the utility of trigger events to monitor behavioral shifts, an element we integrated into our study to amplify urgency.

Recent AI research has also spotlighted scenario-based design and role-play [8, 66, 111]. While Wolf et al. [111] adopted a theoretical stance, generating scenarios from prior research, our approach, akin to Eiband et al. [33], is more participatory. However, our focus is solely on the evaluation phase, unlike Eiband et al. who engaged users throughout the design process. A recurrent critique of these methodologies is their limited replicability. To try and address this, Geerts et al. [38] repurposed the Serious Game Design (SGDA) framework [72] to evaluate and guide the design of research games - games used as playful methods in HCI research to engage participants and collect user insights. The framework breaks down the game into Purpose, Content, Mechanics, Narrative, Aesthetics, and Framing with the emphasis on cohesiveness and coherence between the pieces. The authors found that the framework helps support a more systematic development of the game. Given that our study can be categorized as a serious research game, we used the SGDA framework post hoc for comprehensive evaluation. Paired with our description in this paper, we hope that it improves reproducibility, keeping in mind the context of the study.

Our approach is novel in its use of scenarios to understand the design of media verification tools for journalists. Consequently, we have encountered unique challenges in crafting realistic and demanding scenarios centered on video content verification.

2.6 Education and Training

Prior work has delved into the efficacy of gamified techniques for education and training. The What.Hack training program [110] showcased heightened awareness of phishing threats among corporate personnel. Similarly, initiatives like FakeYou![21] and MAthE[52] employed analogous strategies to bolster public defenses against disinformation. Landers et al. [60] found that gamified methods resonate particularly with those familiar with video games. Armstrong et al.[6] posited that melding gamification with instructional design tenets can enhance learning outcomes. Building on this foundation and leveraging the SGDA [38], we believe that our methodology could be refined and repurposed to train journalists in dealing with the fast-evolving and complex types of disinformation. This would empower journalists to hone their media verification skills in simulated settings featuring possible deepfakes.

3 METHOD

3.1 Sample and Consent

The study involved 24 participants. As shown in Table 1, nine participants engaged in individual sessions, while the remaining 15 participated in group sessions. Individual interviews spanned September 2020 to November 2021, with group sessions conducted in June 2022. All participants were active journalists with diverse experience in news publishing working on various beats (topics) that include tech, politics, health, and entertainment. While all participants were familiar with deepfakes, their understanding of the term varied, hence we used pre-briefing (see Section 3.2) to give everybody a similar starting point.

Recruiting journalists for hour-long research sessions posed a challenge, which contributed to our modest sample size. However, through strategic outreach, leveraging personal contacts, assistance from our funding agency, and employing snowball sampling techniques, we were able to assemble a diverse group of journalists representing both local and national news agencies. Each participant brought several years of journalism experience to the table, and all had dealt with digital multimedia in their professional roles, albeit to varying extents. This approach allowed us to build a representative sample, capturing a spectrum of verification habits from journalists across the United States.

The study and its methodology received approval from the Institutional Review Board (IRB) at the Human Subjects Research Office in Rochester Institute of Technology. Participants were informed about the study's focus on news verification concerning deepfakes and their detection. Prior to the study, participants received an Informed Consent document containing two sections: one seeking consent for participation and an optional section requesting permission to disclose their names. The latter was not solely for academic publication but also to discuss their participation in other mediums.

Upon conclusion of the session, participants were debriefed about the unpredictable nature of deep learning tools and the imperative of thorough verification when utilizing them.

While potential risks, albeit minimal, especially in terms of employment implications, may arise due to the potential loss of anonymity, it is noteworthy that majority of the participants willingly provided consent for the use of their names. The primary

ID	Gender	Scope	Beats	Ver
1L	M	Local	BUS, ENV, JST, POL, SOC	
2L	M	Local	EDU, HLT, JST, POL, SPO	
3L	M	Local	EDU, ENV, JST, POL, SOC	
1N	F	National	CUL, ENV, HLT, POL, SOC	✓
2N	M	National	CYB, JST, POL, SOC, TEC	
3N	F	National	CUL, EDU, JOR, SOC	
4N	F	National	DIS, HLT, POL, SOC, TEC	✓
5N	F	National	CUL, DIS, POL, SOC, TEC	✓
6N	M	National	DIS, POL, SOC, TEC	✓
1GL	M,M,F,F	Local	broad variety	
2GL	M,M,M,M	Local	broad variety	
3GL	F,F,F	Local	broad variety	
4GL	M,M,M,F	Local	broad variety	

Table 1: Anonymized List of Study Participants. *Ver* indicates whether the participant regularly engaged in verification of digital multimedia. Beats Legend: BUS - Business, CUL - Culture, CYB - Cybercrime, DIS - Disinformation, EDU - Education, ENV - Environment, HLT - Health, JST - Justice, POL - Politics, SOC - Society, SPO - Sports, TEC - technology.

objective of this research is to gain a deeper understanding of these tools and to offer guidance for the development of a reliable deepfake detection tool that empowers journalists to validate media sources, thereby enhancing the quality of their reporting. Such a tool holds particular value for journalists operating without the support of a dedicated fact-checking team. Given journalists' inherent appreciation for the importance of anonymity, they are well-positioned to assess the risks and benefits of participating in such studies.

3.2 Study Structure

Our studies were conducted using a semi-structured, scenario-based role-play approach, spanning one-hour sessions. Drawing inspiration from the D&D, we integrated its fictional scenarios and the flexible control of the *DM*. Following the left side of Figure 1, the participants were placed in a series of carefully designed newsroom scenarios, mirroring real-world situations, each accompanied by a piece of newsworthy information and a video description.

Pre-Briefing. During the pre-briefing sessions, we introduced the concept of deepfakes and familiarized participants with various verification tools, including those designed for deepfake detection. We also discussed prominent instances of deepfakes in the media. These tools were presented as cutting-edge AI solutions capable of accurately identifying manipulated videos. It is important to acknowledge that this introduction might have influenced participants in favor of utilizing deepfake detection tools, potentially amplifying concerns about the prevalence of deepfakes. Despite this, it's worth noting that, at the time of our research, promotional materials for these tools often portrayed them optimistically, and many news articles on deepfakes tended to adopt a sensationalist tone, predicting significant societal impacts in the way media is perceived.

Introduction to Scenarios. Participants were briefed on their role, typically that of an investigative journalist in a prominent news organization as shown on the top right of Figure 1, though roles could vary based on geographic context. The scenario's backdrop was then presented, offering information for verification and potential publication, supplemented by a video description. Subsequently, participants were asked to assess the trustworthiness of the source of the information and its potential newsworthiness using a three-point Likert scale. We emphasized the low stakes of these ratings, aiming to capture their impressions of the scenario's characters and quality. The introduction culminated in an initial verdict or *step 0*, where participants rated the scenario on a 5-point Likert scale ranging from "Real" to "Fake". This initial verdict aimed to capture participants' preliminary perceptions and monitor for potential *primacy effect* biases [13].

Verification Activity. Central to the study was the *step*-wise verification activity. At each step, participants were presented with a grid of potential actions to help guide their content verification process. When selecting an action, they received the corresponding response. Depending on the number of steps taken, an *event* might be introduced, elevating the scenario's urgency. Regardless of whether or not an event was triggered, participants then revisited the Likert verdict scale, indicating any changes in their judgment. This cycle was repeated until the participants decided to publish or dismiss the article based on their findings, after which the participants were notified whether or not the video was deepfake. Although the participants were presented with an initial set of actions, they were made aware that these actions were provided to help them make their decisions and they were free to invent their own actions at any point.

Debriefing. After completing three randomly selected scenarios, the participants underwent a debriefing session. This allowed clarifications, discussions on the efficacy of deepfake detection tools, best-practice recommendations, and potential study extensions.

Role of the Dungeon Master. Analogous to the original D&D game, the DM role played a pivotal part in our study. A robust understanding of journalistic information verification was crucial for the DM, given the predefined responses for many actions, with occasional live adjustments based on the step sequence. The DM's expertise became especially vital when participants opted to invent their own actions, requiring spontaneous response generation. Additionally, the DM injected a sense of urgency, emulating the high-pressure newsroom environment prone to "Honest Errors"—unintended errors made through decisions that were not malicious [13]. In our study, the researchers themselves assumed the role of DMs, ensuring a deep grasp of journalistic methods, scenario structures, and study objectives. If a volunteer were to take on the DM role, simplified guides—though less intricate than those for the original game—would be necessary.

Interactions with the Video. While the scenarios involve potentially manipulated videos, as discussed earlier, it's important to note that we only provide participants with textual descriptions of the videos. In a manner akin to the D&D game, participants' interactions with the video rely on detailed descriptions provided by the DM. This approach is driven by the resource-intensive nature

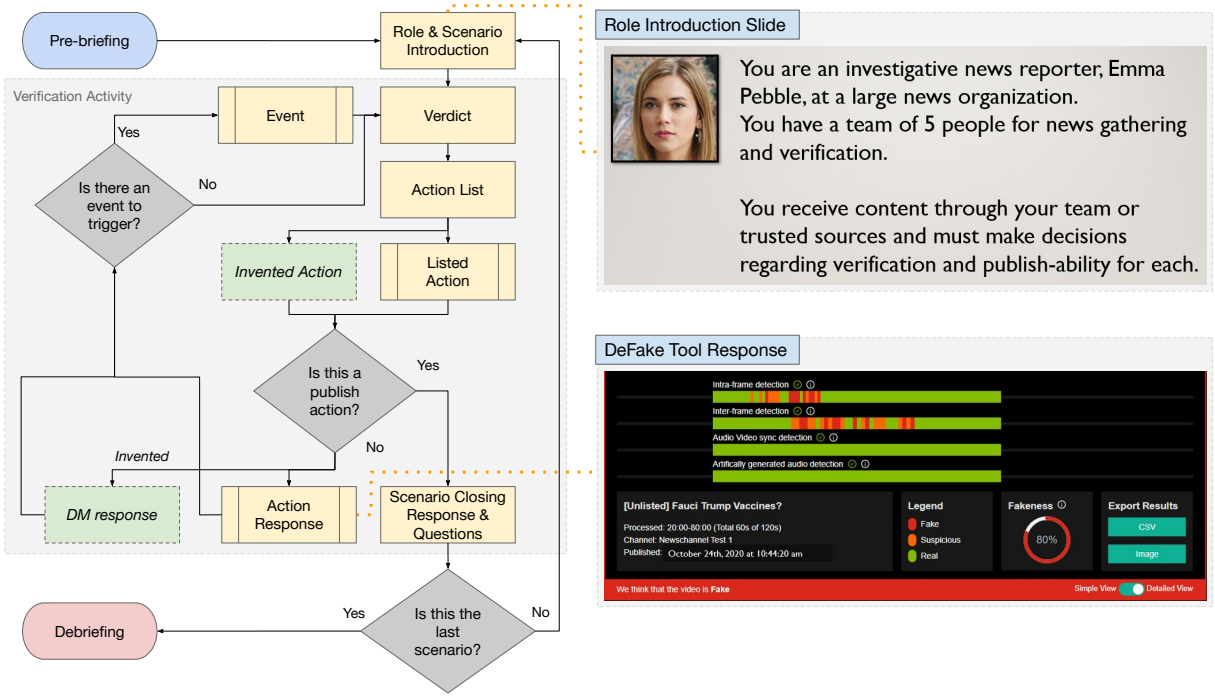

Figure 1: Left: Dungeons & Deepfakes study flow showing the various steps taken over the course of each session. Top Right: Introduction slide with an avatar generated using a face generator. Bottom Right: A cropped presentation slide corresponding to the participant's selection of the action to use a deepfake detection tool. The response reveals identified manipulation in the visual frames using two distinct methods, collectively indicating a 98% certainty, alongside predominantly pristine audio content. In this particular instance, the tool does not detect any inconsistencies in the synchronization between the audio and video signals.

of crafting a near-flawless deepfake. Our decision to refrain from sharing the actual manipulated video stems from a desire to avoid potential unintended artifacts that the participants could point to in order to raise their suspicions while ignoring other pieces of information [41].

It is crucial to acknowledge that a highly motivated adversary could feasibly possess the requisite resources to produce a high-quality deepfake for use in a disinformation campaign.

Group Setting Adaptations. For 15 of our participants, the study was conducted in a group setting with 3-4 participants at a time. This was necessitated by time constraints during in-person newsroom visits, but also offered insight into collaborative decision-making in newsrooms. A significant change in this setting was the designation of a *leader*. Thus, although we promoted discussions amongst the groups during the activity, the leader assigned actions to the group members and made the decisions. After each member received the response of their action, a group discussion ensued, potentially revealing the "Conformity Effect" [13]. The leader rendered the final verdict, and the steps proceeded until a collective publication decision was reached. To ensure efficiency in the face of potentially extended discussions, we introduced a total time limit for the scenarios and updated participants regularly on how much time remained.

3.3 Development

Solo study participants were presented with scenarios via Zoom meetings using Microsoft PowerPoint. Slides were built for introductions, rating scales, action card grids, responses, and events. For group studies, we adapted these slides for in-person presentations in newsrooms.

3.3.1 Scenarios. Our finalized study design included seven fictional scenarios, as outlined in Table 3. These scenarios blended elements of contemporary events: five were set in North America (NA) and two were international (INT). Real-world names were used to encourage participant familiarity and set expectations for character behavior. Participants were assigned roles with fictional names and synthetic faces.[1] The represented newsrooms, though fictional, were described in terms of scale, often drawing comparisons to real news organizations.

Primary subjects. in the scenarios were well-known figures with distinct behavioral patterns. For instance, Elon Musk's reputation for informal Twitter interactions and significant online announcements allowed us to craft a scenario where he unexpectedly announces his retirement. We leveraged Former President Trump's polarizing stances in scenarios NA3 and NA5. Dr. Fauci, known for his composure, was portrayed in a controversial light in NA5.

[1]Generated using https://www.thispersondoesnotexist.com

Character Introduction	
You are an investigative reporter, Ezekiel Potter, at a large news organization. You have a team of 5 people for news gathering and verification. You receive content through your team or trusted sources and must make decisions regarding verification and publishability for each.	
Referrer Information	**Video Description**
`October 28, 2020` You receive an email from a news verification agency run by your former colleague (Mark): *"Hi Ezekiel, We've been running through some videos that are getting a lot traction online and stumbled on this one. The context and metadata seem to checkout according to our team, and Mark thought it's more up your alley, so he suggested I forward it to you. And yeah, I know, the content did seem quite suspicious at first, especially with all the allegations. Regards, Dave."*	Biden's former campaign manager, Greg Schultz, claims that Biden has been severely ill for the past few weeks and has been going to rallies on painkillers. The video is slightly compressed, and Greg is visible from shoulders above sitting on what appears to be a couch. **Tool Verdict:** Fake with suspicious audio. **Real Verdict:** Fake
Actions	**Events**
[**regular**] Contact the sender; Visit Greg Schultz's Twitter; Search Google for context; Look for other news reports; Analyze original poster's account; Contact the original author; Message Greg Schultz; Analyze retweeters' accounts; Search Twitter for context; Invent action. [**technical**] Check metadata; Run through DeFake tool; Manually analyze video; Use InVID Project; Request biometric analysis (3 step cost). [**goals**] Publish the story as fake; Publish story as real; Do not publish anything.	Event 1: Video goes viral with 30,000+ retweets, Trump retweets. Event 2: Greg Schultz responds denying involvement.

Table 2: Scenario NA1 based in North America, involving the Presidential Candidate Joe Biden. Note that not all elements are shown to the participants.

These characterizations aimed to establish a foundation for the scenarios and potentially induce the "Primacy Effect" [13], potentially influencing participants towards an early verdict if they did not maintain thorough and objective verification.

The stories. constructed around these primary subjects were designed to have substantial societal and economic implications. The significance of the information, combined with the rapid spread of rumors on social media, and the pressure for a large news organization to be the first one to publish, could contribute to confirmation bias [30]. Table 2 illustrates one such scenario.

Our fictional adversaries in the scenarios employed more than just deepfake video manipulation. They combined these with cybersecurity attacks and propaganda tactics, including spear phishing, hacked Twitter accounts, spoofed email addresses, and compromised publisher websites. A notable example of such multifaceted misinformation campaigns emerged during the Russia-Ukraine conflict, which combined hacking of TV and radio stations with deepfake audio and video [74].

Incremental development. of our scenarios proved beneficial, allowing us to refine our narratives over time. Our goal with the scenarios was to ensure they contained highly newsworthy information, making it worthwhile for the participants to go to great lengths to verify the content for publication. Subsequently, using feedback gathered through the 3-point Likert scale, as outlined in the Introduction and during debriefing sessions, we discerned the scenarios that resonated most strongly with participants. This feedback highlighted the fact that some participants from major news organizations may have the resources to directly contact key figures like Elon Musk and expect a fairly prompt reply, foregoing a lot of the verification steps. Consequently, although we retained the Elon Musk scenario due to his activities being newsworthy, we strategically prioritized scenarios, featuring key figures whose involvement heightened the urgency for thorough verification. However, to avoid easy debunking, the DM had to come up with creative and logical responses. Moreover, insights from debriefing participant 4N prompted us to diversify our scenarios. Responding to the suggestion for scenarios with more restrictive and less trustworthy conditions, we introduced INT1 and INT2, situating them in distinct geographical regions.

3.3.2 Actions and Responses. Our initial set of verification actions and their corresponding responses were based on previous research [10, 11, 102]. As detailed in Table 2, actions were categorized into *regular*, *technical*, and *goals*. Through *regular* actions, participants interacted with scenario characters and performed contextual analysis. *Technical* actions delved into forensic video analysis using tools such as metadata readers, a deepfake detection tool [102], and the InVID video analysis tool [104]. The final category, *goals*, pertained to publication decisions. The actions were additionally given action identifiers for each category to improve the quality of life during the analysis phase.

Deepfake Detection Tool action responses were in the form of screenshots as shown on bottom right of Figure 1, DeFake tool [102]. We recreated both the collapsed and expanded interfaces mentioned in the original and adapted the results to match the scenario design, showing the collapsed interface first. Additionally we included the action for *Biometric Verification* from the paper, allowing the participant to request it if unavailable. In the scenarios, we marked all the verdicts as fake, and varied the amounts of fake content detected on each video as well as the fakeness of the audio signal. This decision was driven by the fact that the tool seemed to lean

ID	Referrer Information	Video Details
NA2	`August 16, 2020` | Dylan, a team member from your news gathering team, points out this Tweet from Elon Musk's verified account and suggests that it could be an interesting story to go on. [Tweet] *Stepping down as the CEO of Tesla after over a decade. Jérôme Guillen will be taking over the reins, and he has my full support as his resume speaks for itself. will still continue to be involved in the day-to-day decision making process as the founder, but the decrease in company responsibilities will give me more time to spend with the family, something I have not been able to afford in recent years.*	Announcement seems to have been made from an office space, with Elon looking straight at the camera. **Tool Verdict:** Suspicious with suspicious audio. **Real Verdict:** Fake
NA3	`October 18, 2020` | A team member from your news gathering team points out a viral Tweet from Kanye West. [Tweet] *I am pulling out on my election plans. Me and Kim refuse to work with a party that supports the vision of a terrible person like Trump. All the best wishes to Biden!*	Kanye West sitting on a couch at what looks like his home. Video shot in a conference call fashion. **Tool Verdict:** Fake with fake audio. **Real Verdict:** Real
NA4	`October 20, 2020` | Your producer wants to run a video with a headline that says: *"Biden is sick? His campaign manager appears to say so."* She tells the news director about it and he wants you to do a quick fact check run before going live. He notes that the news is time sensitive and they'd want to run with it soon. And have the anchors mention that the video has not been verified yet and say: *"Is it real? You decide."* In case verification isn't finished, he'd stick with the last part.	Biden's former campaign manager, Greg Schultz, claims that Biden has been severely ill for the past few weeks and has been going to rallies on painkillers. The video is slightly compressed, and Greg is visible from shoulders above sitting on what appears to be a couch. **Tool Verdict:** Fake with suspicious audio. **Real Verdict:** Fake
NA5	`October 24, 2020` | A competing highly reputable news organization, NewsPeople, publishes an article on a video of Dr. Fauci claiming that a vaccine already exists that is only reserved for POTUS and his family. The vaccine will be delayed to be released right prior to the elections to bolster the Trump administration's re-election chances. The administration blocked the publication of research related to the vaccine in fear that it would be developed into a cure abroad. Your editor wants you to write an article on it as soon as possible to get some of the early readership.	Dr. Fauci sitting in his home office talking to the webcam. **Tool Verdict:** Fake with real audio. **Real Verdict:** Real
INT1	`January 25, 2022` | A coworker at your news organization points out a video of secretary to Prime Minister Sheikh Hasina, Md. Tofazzel Hossain Miah admitting to bribery, that is going viral on TikTok. The news that could put a lot of pressure on both the secretary and the bribing party, East West Properties Development Ltd. The same party that was accused of fraud and embezzlement about 10 years ago.	Md. Tofazzel Hossain Miah is seen talking about receiving Tk. 75 lac from Ahmed Akbar Sobhan Shah Alam with the aim to facilitate a takeover of 20 acres of land in Ketun region on the outskirts of Dhaka city. **Tool Verdict:** Fake with real audio. **Real Verdict:** Fake
INT2	`January 25, 2021` | Myanmar government-controlled National news publishes an article with Maung Zarni, a Burmese activist for human rights, taking a payment from two other men in what appears to be a shipping dock. The two men are identified to be part of a recent round of arrests related to human trafficking.	Maung Zarni, are meeting 5 men at a shipping dock. Three men are armed. Maung Zarni opens a container. Armed men check the container. The interior of the container is not visible. The unarmed man gestures to one of the armed men, who hands off a gym bag to Maung Zarni. Maung Zarni walks out of the frame. **Tool Verdict:** Fake with fake audio. **Real Verdict:** Real

Table 3: List of scenario content introductions stripped of character introductions, actions, responses and events.

towards fake verdicts in practice, however the expanded view with a detection timeline helped evaluate the content.

Inventing actions. The participants had the flexibility to invent actions. This allowed them to freely define their own verification steps thus having relative freedom in their workflow. As we refined our scenarios, some of these invented actions were integrated into our base action list, like calling notable organizations or relevant characters.

Predetermined responses for all predefined actions took the pressure of the DM and ensured they provided sufficient information while limiting access to pivotal characters. Some responses were designed to bias participants towards a particular verdict. Depending on the scenario's progression, the DM may have adjusted variables within these responses, like the number of shares on social media, to maintain a sense of urgency and maintain logical order.

3.3.3 Events. To generate a sense of realism and urgency, the DM strategically introduced events as participants progressed through the study. Each verification action was assigned an associated *action cost*, accumulations of which dictated when an event would unfold. While the majority of actions carried a standard cost of one, those involving potential wait times exceeding a day were deemed to have a higher action cost. For instance, a biometric analysis request for a video subject incurred a cost of three, potentially necessitating an additional two days for completion.

These events primarily manifested as reactions on social media platforms, often marked by surges in retweets accompanied

by context and metrics. Some events provided deeper insights by incorporating character reactions, such as denials or delays. The DM dynamically adjusted variables within these events, such as the volume of social media discourse, based on participants' prior actions, creating a responsive and adaptive narrative experience.

3.4 Analysis

Our data sources included Zoom recordings from online sessions, in-person group meeting audio recordings, and handwritten notes from both. Zoom recordings were automatically transcribed, while in-person recordings were transcribed using Rev.ai.[2] Encouraging participants to think aloud enriched our insights into their decision-making processes.

Our primary objective was to understand participants' approaches to verifying video-based information and their interactions with deepfake detection tools. With this in mind, we conducted a thematic analysis of the recorded data. First, the two authors independently went through the transcripts from the online and in-person interviews and developed their own initial codebook using open coding. Then, the authors reconvened to share their codebooks and derive higher-level themes using axial coding [94]. During the collaboration phase, we merged conceptually similar themes and arrived at our final codebook. To enhance the accuracy of our analysis, we reviewed the original recordings while developing the codebook to avoid discrepancies that arose from automated transcription errors. The multiphase coding process allowed us to discern nuanced behavioral shifts and expressed opinions. We particularly observed how detection tool results influenced participants' content validity judgments and their reliance on these tools. To visually capture the variations in participants' decision-making processes—specifically, the number of steps taken—both before and after employing the detection tools and the resulting shifts in their verdicts, we constructed step graphs, as shown in Figure 2 and 3, using the *Deepfake Detection Tool* action as a reference point.

4 RESULTS

This work aims to gain insight into how journalists, as a specialized group of users, would approach news verification in the age of deepfake videos and detectors. We emphasize that the focus of this work is to assess near-term effects, as access to the deepfake detection tools is heavily democratized and developers rush to release their novel detection tools to the public. At the same time, we evaluate the risks of overconfident claims and the lack of disclaimers about detector shortcomings affect users without much prior experience with the tool.

4.1 RQ1: Perception of journalists towards deepfake detection tools

At the time of the studies, most of the participants were unaware of the existence of deepfake detection tools, and some were unfamiliar with the InVID tool. This is not surprising as Khan et al. [53] suggested that journalists often avoid tools due to a lack of technical knowledge. Participants 4N and 5N were, however, more accustomed to dealing with modern digital multimedia verification and were familiar with the various verification tools available to them. Participant 5N went as far as naming various other deepfake detection tools that were available at the time.

Participants had positive views on deepfake detection tools. All the participants indicated that having these detectors is vital and looked forward to using them. Echoing the findings from the previous work with journalists [102], journalists expressed gratitude during the debriefing sessions for the research that has been going into deepfake detection and mentioned the necessity of these tool sets in their verification arsenal. One caveat that would often come up in the debriefing regarding these tools is the lack of explanations to substantiate the results.

Trust in the app may depend on the developer reputation. Generally, the participants that used the tool felt comfortable receiving an answer from it. This may be attributed to the connection of the application to a neutral academic institution. When the discussions about trust emerged, the participants mentioned that they would trust a tool developed by a neutral entity. Group 2GL mentioned that they would be more comfortable if they knew that the tool was developed by a university source with transparent development and developer details that they could verify. They wanted to ensure that the tool was not a "part of an agenda" that could influence their opinions.

Positive experience with the tools may contribute to an increase in trust. If the tool provided users with correct responses in the past, there would be a higher chance of them trusting it. The participants from 1GL, in scenario NA5, chose to follow the output of the detection tool and made a publication decision based on that, because it gave them the correct answer in scenario INT1. On the other hand, more tech-savvy users may perform their own benchmark of the tool, something that participant 5N alluded to during the debriefing.

Time spent on the deepfake detection tool varies. The participants were presented with an interface for the deepfake detection tool [102], which displayed an aggregate score for the probability of the video being fake, as well as a more detailed view of individual scores for four detection methods per second of the video. Each participant spent varying amounts of time reviewing the results. Although every participant received a brief overview of the detection interface, experience with verification technology (4N, 5N) and group settings (1GL, 2GL) made participants pay more attention to individual detection results. Most of the participants did not raise any questions about the results unless prompted by the interviewer.

Opinions regarding the inclusion of the tool's results in publications varied. When asked whether participants would include the results of the deepfake detection tool in their publication, there was a wide spectrum of opinions. In many sessions, the participants mentioned the importance of the reputation of their own organization in deciding on the contents of their publication, as mistakes tend to tarnish the trust. Most participants were keen on adding the results to their publication for transparency of their methods. 3N and 3GL, however, specifically pointed out that they would not, remarking that the tool only served their own organization's needs for the verification. A participant from 2GL gave a different reason

[2]https://rev.ai

for not including the results: *"I feel like audiences will get bogged down. They don't really understand what it is."* As cautious as many journalists can be, another 2GL participant said: *'If we feel like we have a reliable tool that we're basing it off, I would reference that,"* which captured the essence of many participants' thoughts on this issue. The public access to the tool also came into question, with a few of the participants mentioning that it would be unproductive to include the results if the public could not test it for themselves.

4.2 RQ2: When and why deepfake tools are used

The study allowed us to take a closer look at the workflows of the journalists in a simulated newsroom scenario and their verification behavior when encountering video material. Although there are various guidelines for newsroom content verification [31, 32, 43], studies like the one by Lewis et al. [63] suggest that in practice, the verification process is less defined. Figures 2 and 3 show the disparity in the number of steps used to verify content and the order between various participants. The table was organized by using the *Deepfake tool* action as a reference, allowing us to observe the number of actions taken before and after it.

Journalists start with verification of context through traditional means. Unsurprisingly, participants chose to start with traditional journalistic actions of contacting characters in the scenarios most of the time. Participants from larger national newsrooms would extend this further, using more steps before using the deepfake detection tools. They attributed this to having to adhere to a more well-defined verification requirement when compared to local newsrooms. In group settings, each step allowed for three concurrent actions to be assigned. Even though the *Deepfake tool* action was usually included in the first step, the traditional actions were picked first. *"Typically we would reach out to all parties first,"* said the leader of 2GL. Traditional approaches were preferred due to a mix of comfort through experience and the need to put together the context for the events associated in the video. As noted by participant 3L, *"I'd want to minimize the value of the video,"*.

Participants who were more comfortable using technology for verification exhibited goal-oriented behavior when using the said tools. Their goals were to verify various pieces of context to form a more complete picture, thinking "I want to verify whether X is actually in this location/on this day." As an example of this, 2GL, 4N, and 5N used metadata analysis to identify date and location, while the latter two participants and 4GL used image reverse search to find related news and videos.

Deepfake detection is used when context verification is difficult. The deepfake tool was often picked if contextual verification could not yield confident conclusions. However, in contrast to the goal-oriented behavior, the participants showed a discovery-oriented behavior when opting to use deepfake detection tools. A majority of the participants demonstrated a 'let us see what this tool says about the video' mindset while deciding to use the deepfake tool. This inclination may be caused by their limited familiarity with the tool's functionality and performance. For instance, in scenario INT1, participants from group 2GL hesitated initially about the utility of using the deepfake detection tool, but ultimately decided to try it as their final action in *step 2*.

Figures 2 and 3 show that in both individual and group studies, the detection tool was used earlier in scenario NA5. The primary subject in scenario NA5, Dr. Anthony Fauci, makes a very bold and unusual statement in a video interview. *"Honestly for like a story of this magnitude, wouldn't you not just like do anything possible before publishing it?"* said a participant from 2GL while discussing the possible actions, pointing to the magnitude of this news. Participant 2N was less interested in the results of deepfake detection tools in the first scenario they faced, saying that they would want to dissociate the video from the news mentioned in it. However, for the Fauci video, they used it early, stating that the odd behavior and the fact that the subject was alone in the video made manipulation more likely. Deepfakes most commonly take the form of single talking head videos. Therefore, it is sensible that they may be concerned if the behavior of the subjects is out of the ordinary.

It is worth noting that participant 4N completely avoided using deepfake detection tools in all of the scenarios. In the debriefing however, they eluded to the fact that they would definitely have considered using the tools if the various contextual steps proved to be fruitless. The participant talked about needing to develop various layers of evidence in order to come to a conclusion, a part of which would have been results from not one, but various detection tools to get diverse opinions on the content.

Time was also a defining factor when choosing to use the verification tool. *"No tool is perfect and I am trying to do everything I can to add more context allowing the reader to see the bigger picture,"* said 5N, admitting that technology-assisted solutions often lead to faster albeit incomplete results. In contrast, traditional methods were more tried and tested, but would often take longer to yield results like a response from characters they tried to contact directly. Time is scarce in modern breaking news scenarios. As a member of 1GL stated during one of the scenarios, *"We only got a little bit of time, so DeFake, we're gonna trust that."*

4.3 RQ3: Potential for overreliance on tools.

Overreliance due to cognitive biases on automation in human-AI collaboration has been a topic of many recent studies [37, 90] and will continue to be as our interactions with AI-driven tools evolve. Similarly to these works, we touch on the potential effects of cognitive biases on the journalistic verification process given a hectic newsroom scenario. On average, it was reassuring to find that three individual participants and three groups displayed a healthy amount of skepticism toward the performance of the deepfake detection tools. A nice example was a discussion amongst the participants in 3GL during scenario INT2 regarding the possible lack of effectiveness of the detection tool on a video from a surveillance camera feed where the faces were less clear.

Overreliance. The novelty and optimistic description of the deepfake detection tools in the pre-briefing may have introduced an automation bias [62] wherein the participants display overreliance on the tool due to blind trust. While it is hard to pinpoint overreliance, we marked the instances (△) where participants used very few contextual verification actions and decided on a final verdict within two actions of using the tool on Figures 2 and 3. In the session with 4GL, one participant showed extreme interest in the detection tool and said: *"I didn't know that technology existed before*

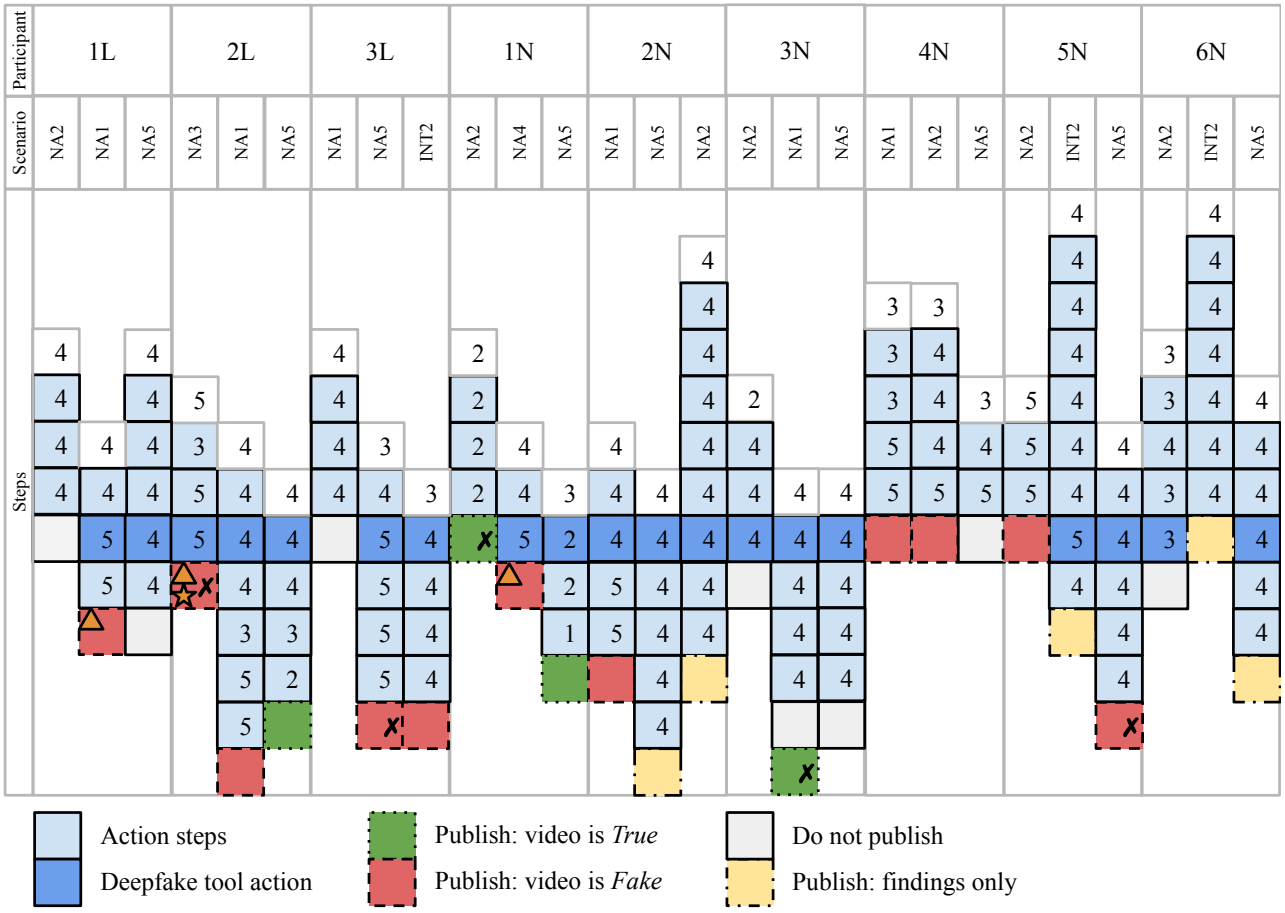

Figure 2: Individual Study. Analysis of the verification actions taken by the participants per scenario. Actions start at the top (white box) and move down, keeping the location of the deepfake verification tool action as a reference point. The action boxes contain the verdict Likert response (1: "Real" to 5: "Fake") after using the action. Publication boxes are marked with ✗ for a wrong verdict, ⋆ for reaffirmation, and △ for overreliance. The top box in each scenario denotes the initial verdict before verification. Please refer to Table 1 for the participant codes, and Tables 2 & 3 for the scenario codes.

just now. If I had access to it, I'd be running so much stuff through it". Although it was nice to see a high acceptance rate of the technology among the participants, developers must be aware of the high dependency of users on the technology. An example we saw was from the session with group 1GL. Having seen that the detector provided one accurate result for the previous scenario, they then trusted the tool too heavily in scenario NA5, in which the tool actually provided the wrong result. Participants 1L, 2L, 1N, and 1GL showed signs of possible overreliance on the tool through their activity patterns and their reasoning as they talked about their actions. Right after using the detection tool in scenario NA1, participant 1L said, *"That's looking pretty fake and with my hackles already up on this thing,"* and after using one step that happened to support their suspicion, they chose to publish the article. During the debriefing, participants suggested that visual cues indicating causes that may affect the robustness of the results would have been helpful.

Reaffirmation. While overreliance may have been based on naivety and lack of experience, reaffirmation targets the aspects of confirmation bias and "Primacy Effect" [13], which is a tendency of the people to give more weight to the information gathered early in the verification process. Given a prior of the participants' knowledge of the characters in the scenarios, and the initial verdicts given in *step 0*, we can observe what happens when the initial verdict and the results of the tool are in agreement. If the detection result leads participants to revert to their initial *step 0* verdict, considering the possibility that their judgment may have changed during the process due to other verification actions, we interpret this as reaffirmation. We also assume reaffirmation if the participants decide to publish as soon as the result of the tool agrees with their verdict in *step 0* without further analysis. The example of the event in Myanmar can be an example of this behavior, since journalists went in thinking that the video was manipulated, only to have their suspicions exacerbated by the results of the tool. Given our definition,

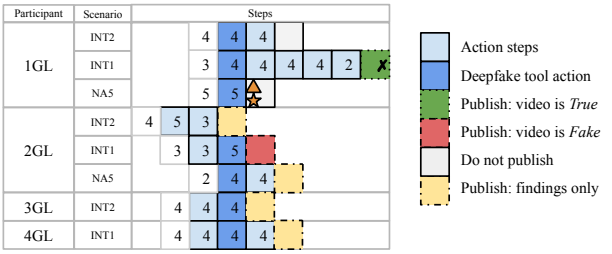

Figure 3: Group Study. Analysis of the verification actions taken by the groups per scenario. The sequence begins on the left (white box) and moves right, keeping the location of the deepfake verification tool action fixed as a reference point. Each action box represents a set of three actions and contains the verdict Likert response (1: "Real" to 5: "Fake") after using the selected actions. Publication boxes are marked with ✗ for a wrong verdict, ⋆ for reaffirmation, and △ for overreliance. The leftmost box in each scenario denotes the initial verdict before verification. Please refer to Table 1 for the participant codes, and Tables 2 & 3 for the scenario codes.

the effects of overreliance and reaffirmation are not mutually exclusive and can co-occur. However, we thought it would be important to pay attention to reaffirmation, as this may cause the participants to pay less attention to detail in the detection results. We identified reaffirmation in sessions 2L-NA3 and 1GL-NA5, denoted by ⋆ in Figures 2 and 3.

There seems to be added uncertainty due to the existence of deepfakes. In many scenarios, participants opted to invent an action in which they published an article with informational content about the events that transpired and their verification progress instead of a definitive verdict. The participants chose this action after extensive verification strategies did not yield the full picture they were hoping for, with the result of the tool sometimes contributing to more confusion. This action lets them publish on an important video to inform the public while mentioning that its validity is still uncertain. For example, participant 2N in scenario NA5 maintained their suspicion about the video of Dr. Fauci, thinking that it was manipulated, even though several other verification steps pointed towards it being real. They chose to publish an article with their findings and stated that they were not yet able to reach Dr. Fauci to make a statement, so they could not confirm the video's validity.

We observed that often the results of the detection tools added to the participant's uncertainty if its results contradicted their other verification steps. Sometimes a response from the detector could affect their perception of subsequent and even previous actions, and they could become suspicious of other characters in the scenario. For example, participants 5N and 6N hesitated to publish after a several verification steps where the deepfake tool contradicted other actions and preferred to send the video to expert groups. While this slows down publishing speed, we believe it is a safe decision to preserve the integrity of the news.

4.4 Impact of the group verification setting.

The group studies gave the participants the advantage of multiple opinions and may have created a more natural newsroom environment. However, these interactions are a double-edged sword, as they may lead to the "Conformity Effect" [85] where the detection tool's results are inflated by peers.

All groups used the detector in their verification process within the first two steps. Team discussions before and after the steps allowed the journalists to make mutual decisions about the actions. Also, since participants needed to select three actions per step, there would often be room left for the detection tools. Thus, we could assume that in team environments these tools would see greater use in parallel with traditional journalistic methods.

We mention that when carrying out the group studies, we allowed the group to elect their own leaders who would be in charge of assigning actions and making final decisions. This works well if there is a comfortable dynamic between the participants. However, in some studies, the preexisting hierarchy within the newsroom may play a part. For example, in 3GL, the participant elected to be the leader was a junior employee, which had an effect on their confidence while making final decisions.

4.5 Study Evaluation and Training

This was a novel qualitative research study methodology, in which we tried to simulate the natural workspace of professional journalists by placing them in fictional scenarios that mirrored our world and then asking them to think out loud while verifying the content. For an hour-long study, we hoped to make the procedures engaging enough to elicit interesting and unique discussions, and memorable enough for useful lessons to be learned.

Evaluation. While we draw inspiration from D&D, our study falls under the category of *research games*. According to Geerts et al. [38], we could use the SGDA [72] framework to evaluate the study. Following this framework, we compiled a diagram of *coherence and cohesiveness* of our study, as shown in Figure 4. The diagram shows to what extent different aspects of the game were strongly or weakly consistent and aligned with the other aspects, based on both the responses from the participants and our own analysis.

When analyzing the coherence and cohesiveness between all the SGDA elements, it was clear that the mechanics of the game alongside the rest of the pieces are strongly coherent with our aims. Throughout the course of the study, we received positive feedback from the participants. They were consistently engaged in the study, and even more so in group scenarios, where they could discuss the process among themselves. The only critical feedback came from one participant, who noted that the study could be tiring due to its intensity. The mechanics of having to pick actions, face events, and deal with consequences may sometimes be overwhelming given the realistic setting. However, this was a design decision we made to elicit more realistic responses.

The SGDA components of fiction/narrative and framing, together with content/information, worked well for the most part whenever the participants were in a more relatable scenario within the US. However, when the scenarios place the participants outside of the

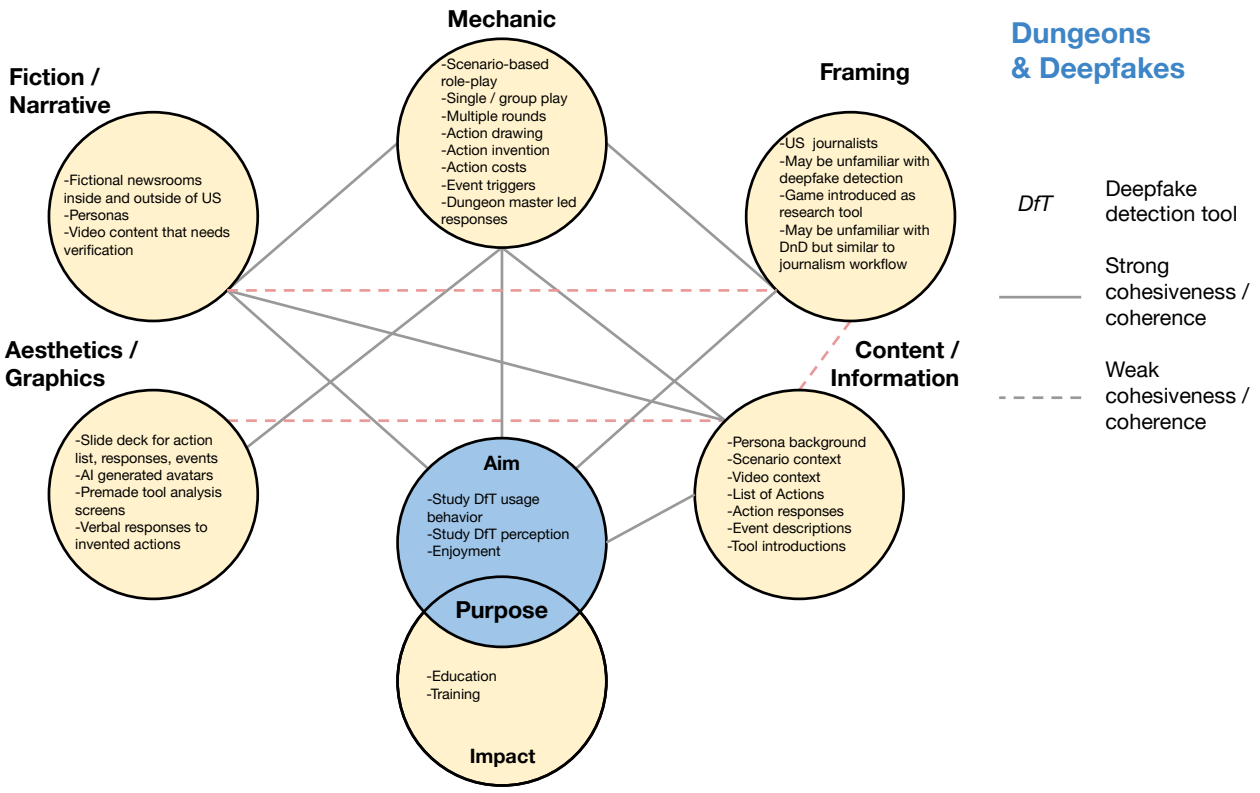

Figure 4: Coherence and cohesion between SGDA elements for the study.

US region, the DM would have to thoroughly describe the limitations, journalistic freedom, and government tendencies of the unfamiliar context. Additional context and clarifications would often need to be provided and repeated throughout the study.

Another pair of components with weaker cohesion between them are aesthetics/graphics and content/information, as we couldn't provide the participants with live videos to look at, but rather text-based descriptions of the content. Although that did not seem to limit engagement, the scenarios would benefit from either low-fidelity visual outlines of the content or high-quality generated videos in the future.

Training. The groups were all local newsrooms, and all felt that deepfakes would have less impact on them, as they often record their own videos and seldom source them from social media. However, one of the groups faced a recent incident where miscontextualized information from a social media source made it into their news broadcast. Hence, the danger from a targeted deepfake video may not be far-fetched. Thus, there is a need for deepfake awareness training for local newsrooms a statement that is echoed in the study of journalists by McClure Haughey et al. [69].

What.Hack [110] showed that gamified role-playing training systems could improve the behavior of office Internet safety by improving the resilience of participants to phishing emails. Although our study requires an active DM to provide live responses, making it harder to use as a drop-in training module for an organization, we were curious what our participants thought about its potential as a training simulation for journalists. During our post-study debriefing sessions, all participants agreed that newsrooms would benefit from a gamified version of the news verification practice used in our work. Participant 2N mentioned: *"This is very fun and I could see this being used in my newsroom."* These positive responses suggest a direction for future work in developing training methods for investigative journalists.

5 DISCUSSION

The Dungeons & Deepfakes study primarily follows a qualitative approach, with minimal incorporation of quantitative elements. Through this study, we have engaged participants in an exploration of their thought processes while performing a familiar verification task, offering valuable insights into our research questions.

Journalists want deepfake detection solutions. The participating journalists displayed a high level of receptivity towards the tool and expressed a keen interest in incorporating it into their verification toolkit. It became apparent that participants with greater experience in digital multimedia verification, particularly when using other technologies, exhibited a greater level of comfort and discernment when using the deepfake detection tool. Some participants even discussed the possibility of conducting their own assessments of the tool to gain a more comprehensive understanding of its capabilities.

Experience and explanations reduce overreliance. While the results were not catastrophic by any means, some participants displayed a considerable degree of trust in the outcomes produced by deepfake detection tools. Those with a higher level of technological proficiency would have a smaller chance of being overreliant on the output of the tools. When verifying digital media with technology, it is vital to spend a sufficient amount of time analyzing the results to overcome biases [90]. Hectic breaking news scenarios may limit the time available, so it is imperative that developers provide explainable findings that journalists can use to verify the results.

Reliance on tools is inversely proportional to the ability to independently verify a video. Section 4.2 findings highlight that seasoned journalists initiate verification through contextual measures. They reach out to characters in the story and try to construct a comprehensive narrative. Often, the compilation of information from diverse sources renders the video's structure inconsequential, as the news it conveys can be independently verified. However, when the trustworthiness of responses from characters within the scenarios diminishes, there is a corresponding increase in reliance on verification tools. In difficult scenarios such as organized, state-funded disinformation campaigns and personal blackmail videos, where assembling a coherent narrative proves exceptionally challenging, deepfake detection tools will be increasingly required.

Training will help empower newsrooms while developers catch up. While researchers and developers tackle the lack of robustness, usability, and explainability of forensic tools, it is necessary to alleviate the dangerous impact of deepfakes and unreliable detection tools on contemporary news. To that end, it is essential to train newsrooms to understand the utility, strengths, and weaknesses of verification tools. They must be able to use these tools with an awareness of their potential pitfalls and without having to learn on the fly when under deadline stress. Based on the participants' lack of exposure to deepfake detection, is not surprising that a substantial number of our participants chose either not to publish anything or to release articles reporting only on verified facts while continuing their verification processes.

Scenario-based role-play is a viable, though complex, user study methodology. Strategic implementation of a game-like role-play experience injected enthusiasm into the study while efficiently extracting valuable insights. It is not simple to design and execute, however. Role-play studies, known for their complexity, demand meticulous curation of roles and scenarios. Our design, inspired by D&D, further necessitates the creation of thoughtful stories, actions, responses, and events, along with a DM possessing sufficient domain knowledge to effectively respond to participants' inventive actions. Nevertheless, our results affirm the effectiveness of our methodology, and the positive feedback received from participants hints at the potential extension of this approach into a training module to address the issues illuminated in our findings.

Broader Impacts on the Public. Assuming that reliable and efficient deepfake detection, though potentially achievable, is still several years away, platforms hosting potentially manipulated information may not yet have the necessary solutions in place to filter or tag such content. Furthermore, considering the fragility of detection tools, releasing them for public use could pose dangers, as any errors, coupled with confirmation bias, may lead to significant controversies, as evidenced by incidents in Myanmar [48] and Gabon [14]. The responsibility for informing the public about the truthfulness of content in a logical manner largely falls on journalists and fact-checkers. To that end, our work points to the importance of making sure that they are well informed of the strengths and limitations of the evolving deepfake detection tools, and are equipped with best practices to enhance their verification efforts. While human error is inevitable, it is essential to mitigate potential exacerbation through mistakes made by automated technologies designed to assist them.

Limitations. The study only recruited participants based in the US, meaning that the findings might not reflect the practices and opinions of other journalists, especially those from different journalistic cultures. However, the participants come from several different media organizations of different scopes and sizes, providing us with varying thought processes and practices. For our study samples, we only focused on the scope of the organization (national versus local) to separate behaviors. Other attributes like gender, age, experience, and geographic location could yield other types of results in future work.

The study procedures are intentionally designed to bias participants and thwart some of their typical practices. This was necessary to construct meaningful and challenging scenarios in which the deepfake detection tool could be seen as helpful, even when it was mostly new to the participants. It suggests, however, that our scenarios are more challenging than ones journalists might typically face, even when deepfakes are involved. In that sense, the design of a deepfake detection tool should not rely too heavily on the results of this study and should also consider related work [101].

Section 4.3 relies on observations of the verification process to evaluate the effects of the deepfake detection tool on the thought process of the participants, where we classified less desirable behavior as potential *overreliance* or *reaffirmation* behaviors with respect to the deepfake detection tool. While we believe that our conclusions are reasonable, it would be better in future studies for the DM to identify these behaviors in real time and ask participants to elaborate on their decisions during debriefing.

Finally, as eluded to in Section 3.2, we did not use actual videos in our studies. This prevented participants from applying their expertise in visual verification to examine the videos. Despite the challenge of visually identifying manipulations in high-quality deepfakes, where most people are not reliable [12, 41, 56, 96], our studies specifically involved journalists, a more seasoned group of users. Journalists and fact-checkers often leverage their experience to interpret body language and detect anomalies. On the other hand, awareness of deepfakes – particularly in the context of our study – might lead to participants suspecting that a genuine video is fake due to compression artifacts or issues with lighting. While our study does not directly address this phenomenon, exploring experts' ability to discern deepfakes and generated media, akin to the work conducted by Ask et al. [7], could serve to extend our methodology and mitigate this limitation.

Safeguards. While the focus of this study was not the development of safeguards, but rather a creative assessment of the dangers, our findings in Sections 4.2, 4.3 and 4.4 showed that it is possible

for journalists to misinterpret the results of the deepfake detection tools and overrely on them. Thus, we argue that AI-based media forensics tools must support journalists to understand how the tools may be inaccurate. We propose the following to safeguard journalists and other users of their AI-based forensics tools:

(1) Add clear warning messages alongside detection results, alerting users to possible inaccuracies and the reasons for them.
(2) Add *tooltip-driven* onboarding for first-time visitors to attune them to the important pieces of the application, including warnings. Mets Kiritsis [70] found tooltip-driven onboarding to be more effective than other popular onboarding methods.
(3) Provide training materials or sessions to organizations and verification groups that could use the tools and include both contextualized warnings and examples of both correct and incorrect results.

While further development to improve the robustness and domain-specific explainability continues, the above improvements should reduce the unintended effects of the assistive technology.

Future Work. This study has the potential for various extensions. It would be interesting to see the differences between the behaviors in this study of journalists from different regions of the world. As Humprecht [45] showed, levels of journalistic professionalism vary between countries. It would be fascinating to observe the influence of hierarchy and experience on the verification and publication decisions in a group setting. Another direction would be to examine different ways to offer domain-specific explainability and transparency in a deepfake detection tool for journalists and seek to understand how they might use these when constructing news pieces. However, it would be important for the explainability to be handled and studied carefully so that it does not exacerbate failures [34, 90]. The study design itself may benefit from potential automation with the help of recent advancements in Large Language Models (LLMs) to design various aspects of the study and even take the role of the DM.

6 CONCLUSION

Journalists are the gatekeepers of truth in the published media. This work provides novel insights into their behavior and perception towards deepfake detection tools through an engaging scenario-based roleplay methodology. Our key findings reveal that while journalists have positive views on these new AI assistance tools and look forward to incorporating them into their workflow, most tend to rely first on traditional verification methods to establish context. The tools see more use when contextual verification is difficult or time is limited. We observe that urgent breaking news scenarios sometimes lead journalists to skip verification steps, while high-impact stories result in more diligent checking.

Additionally, we noticed risky overreliance by a few participants due to potential cognitive biases, particularly when tool results confirm initial impressions. This signals the need for cautions around deepfake detector deployment and highlights the importance of improving explainability. Given the participants' interest, our scenario methodology shows promise for training to improve verification skills.

Given the current surge in interest surrounding Large Language Model (LLM) Chatbots [91], our findings may be extrapolated to indicate that users might overrely on their outputs due to a lack of experience and the prevailing social media hype regarding their efficacy. Despite recent applications warning users about result instability, employing a scenario-based role-play methodology can help assess changes in workflows across various industries, gauge user perceptions of these tools, and determine the degree to which users heed provided warnings.

In conclusion, as deepfake detection tools see wider adoption, it is vital we understand their impacts on professional workflows and provide adequate training. This will empower journalists to harness AI assistance for quality reporting while mitigating unintended harms from shortcomings. Further interdisciplinary efforts between journalists, developers and researchers in this space are crucial.

ACKNOWLEDGMENTS

We would like to thank all the journalists who took time off their busy schedules to participate in these studies and provide valuable feedback. This material is based upon work supported by the National Science Foundation (NSF) Grant no. 2040209 and by the John S. and James L. Knight Foundation.

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, Hong Kong, China, 1–7. https://doi.org/10.1109/WIFS.2018.8630761
[2] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. Detecting Deep-Fake Videos from Appearance and Behavior. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, New York, NY, USA, 1–6. https://doi.org/10.1109/WIFS49906.2020.9360904
[3] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 2814–2822. https://doi.org/10.1109/CVPRW50498.2020.00338
[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Long Beach, CA, USA.
[5] Sensity AI. [n.d.]. Sensity | Deepfakes detection. https://sensity.ai/deepfakes-detection/. Accessed: 2021-10-10.
[6] Michael B Armstrong and Richard N Landers. 2018. Gamification of employee training and development. *International Journal of Training and Development* 22, 2 (2018), 162–169.
[7] Torvald F Ask, Ricardo Lugo, Karl Veng, Jonathan Eck, Muhammed-Talha Özmen, Basil Bärreiter, Benjamin J Knox, Stefan Sütterlin, et al. 2023. Cognitive Flexibility but not Cognitive Styles Influence Deepfake Detection Skills and Metacognitive Accuracy. (2023).
[8] S. Avin, R. Gruetzemacher, and J. G. Fox. 2020. Exploring AI Futures Through Role Play. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020). https://doi.org/10.1145/3375627.3375817
[9] Alessandro Bondielli and Francesco Marcelloni. 2019. A Survey on Fake News and Rumour Detection Techniques. *Information Sciences* 497 (2019), 38–55.
[10] Petter Bae Brandtzaeg, Asbjørn Følstad, and Maria Ángeles Chaparro Domínguez. 2018. How Journalists and Social Media Users Perceive Online Fact-checking and Verification Services. *Journalism Practice* (2018).
[11] Petter Bae Brandtzaeg, Marika Lüders, Jochen Spangenberg, Linda Rath-Wiggins, and Asbjørn Følstad. 2016. Emerging Journalistic Verification Practices Concerning Social Media. *Journalism Practice* (2016).
[12] Sergi D Bray, Shane D Johnson, and Bennett Kleinberg. 2023. Testing human ability to detect 'deepfake'images of human faces. *Journal of Cybersecurity* 9, 1 (2023), tyad011.
[13] Jon S Byrd. 2006. Confirmation Bias, Ethics, and Mistakes in Forensics. *Journal of Forensic Identification* 56, 4 (2006), 511.
[14] Sarah Cahlan. 2020. How misinformation helped spark an attempted coup in Gabon. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.
[15] John M Carroll. 1997. Scenario-based design. In *Handbook of human-computer interaction*. Elsevier, 383–406.

[16] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107 (2019), 1753.

[17] Akash Chintha, Aishwarya Rao, Saniat Sohrawardi, Kartavya Bhatt, Matthew Wright, and Raymond Ptucha. 2020. Leveraging Edges and Optical Flow on Faces for Deepfake Detection. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE.

[18] A. Chintha, B. Thai, S. J. Sohrawardi, K. M. Bhatt, A. Hickerson, M. Wright, and R. Ptucha. 2020. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. IEEE Journal of Selected Topics in Signal Processing (Early Access) https://ieeexplore.ieee.org/document/9105097.

[19] Umur Aybars Ciftci and Ilke Demir. 2019. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *arXiv preprint arXiv:1901.02212* (2019).

[20] James Clayton. 2023. Intel's deepfake detector tested on real and fake videos. *BBC* (Jul 2023). https://www.bbc.com/news/technology-66267961

[21] Lena Clever, Dennis Assenmacher, Kilian Müller, Moritz Vinzent Seiler, Dennis M Riehle, Mike Preuss, and Christian Grimme. 2020. FakeYou!-a Gamified Approach for Building and Evaluating Resilience Against Fake News. In *Disinformation in Open Online Media: Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26–27, 2020, Proceedings 2*. Springer, 218–232.

[22] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510* (2018).

[23] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395 [cs.LG]

[24] deepware.ai. [n.d.]. Deepware | Scan & Detect Deepfake videos. https://deepware.ai/. Accessed: 2021-10-10.

[25] Reality Defender. [n.d.]. Enterprise-Grade Deepfake Detection. https://realitydefender.com/. https://realitydefender.com Accessed: 2023-10-10.

[26] Ilke Demir and Umur Aybars Ciftci. 2021. Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking. http://dx.doi.org/10.1145/3448017.3457387. *ACM Symposium on Eye Tracking Research and Applications* (May 2021). https://doi.org/10.1145/3448017.3457387

[27] Brian Dohalsky. 2020. https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.

[28] Allison Druin. 2002. The Role of Children in the Design of New Technology. *Behaviour and information technology* 21, 1 (2002), 1–25.

[29] Hal Eden, Eric Scharff, and Eva Hornecker. 2002. Multilevel Design and Role Play: Experiences in assessing support for neighborhood participation in design. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*. 387–392.

[30] Stephanie Edgerly, Rachel R Mourão, Esther Thorson, and Samuel M Tham. 2020. When Do Audiences Verify? How Perceptions About Message and Source Influence Audience Verification of News Headlines. *Journalism & Mass Communication Quarterly* 97, 1 (2020).

[31] FactCheck Editors. 2020. Our Process. https://www.factcheck.org/our-process/. *FactCheck* (Aug 2020).

[32] USA Today Editors. 2020. USA TODAY's Fact Check Guidelines. https://www.factcheck.org/our-process/. *USA Today* (Feb 2020).

[33] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. 211–223.

[34] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 183–193.

[35] Dan Evon. 2022. Bad Deepfake of Zelenskyy Shared on Ukraine News Site in Reported Hack. https://www.snopes.com/news/2022/03/16/zelenskyy-deepfake-shared/.

[36] Dan Evon. 2022. Putin Deepfake Imagines Russian President Announcing Surrender. https://www.snopes.com/fact-check/putin-deepfake-russian-surrender/.

[37] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1362–1374.

[38] David Geerts, Marije Nouwen, Evert Van Beek, Karin Slegers, Fernanda Chocron Miranda, and Lizzy Bleumers. 2019. Using the SGDA framework to design and evaluate research games. *Simulation & Gaming* 50, 3 (2019), 272–301.

[39] Chandell Gosse and Jacquelyn Burkell. 2020. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication* (2020). https://doi.org/10.1080/15295036.2020.1832697 arXiv:https://doi.org/10.1080/15295036.2020.1832697

[40] Sam Gregory. 2021. The World Needs Deepfake Experts to Stem This Chaos. https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/. *WIRED* (Jun 2021).

[41] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119.

[42] David Güera and Edward J Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *AVSS*.

[43] Angie Drobnic Holan. 2022. PolitiFact's checklist for thorough fact-checking. https://www.politifact.com/article/2022/mar/31/politifacts-checklist-thorough-fact-checking/. *PolitiFact* (Mar 2022).

[44] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. 2023. Evading DeepFake Detectors via Adversarial Statistical Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12271–12280.

[45] Edda Humprecht. 2020. How do they debunk "fake news"? A cross-national comparison of transparency in fact checks. *Digital Journalism* (2020).

[46] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. 2021. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3348–3357.

[47] iperov. [n.d.]. DeepFaceLab. https://github.com/iperov/DeepFaceLab.

[48] The Irrawady. 2021. Myanmar Junta Accused of Using Deepfake Technology to Prove Graft Case Against Daw Aung San Suu Kyi. https://www.irrawaddy.com/news/burma/myanmar-junta-accused-using-deepfake-technology-prove-graft-case-daw-aung-san-suu-kyi.html. *The Irrawady* (March 2021).

[49] Matthias Jarke, X Tung Bui, and John M Carroll. 1998. Scenario management: An interdisciplinary approach. *Requirements Engineering* 3, 3 (1998), 155–173.

[50] Edel Jennings, Mark Roddy, Alexander J Leckey, and Guy Feigenblat. 2015. Use of scripted role-play in evaluation of multiple-user multiple-service mobile social and pervasive systems. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 7, 4 (2015), 35–52.

[51] Deborah G Johnson and Nicholas Diakopoulos. 2021. What to do about deepfakes. *Commun. ACM* 64, 3 (2021), 33–35.

[52] Anastasia Katsaounidou, Lazaros Vrysis, Rigas Kotsakis, Charalampos Dimoulas, and Andreas Veglis. 2019. MAthE the game: A serious game for education and training in news verification. *Education Sciences* 9, 2 (2019), 155.

[53] Sohail Ahmed Khan, Ghazaal Sheikhi, Andreas L Opdahl, Fazle Rabbi, Sergej Stoppel, Christoph Trattner, and Duc-Tien Dang-Nguyen. 2023. Visual User-Generated Content Verification in Journalism: An Overview. *IEEE Access* (2023).

[54] Aminollah Khormali and Jiann-Shiun Yuan. 2021. ADD: Attention-Based DeepFake Detection Approach. *Big Data and Cognitive Computing* 5, 4 (2021), 49.

[55] Will Knight. 2020. Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them. *Wired* (Jun 2020). https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/

[56] Nils C Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled Twice: People cannot detect deepfakes but think they can. *Iscience* 24, 11 (2021).

[57] Pavel Korshunov and Sébastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *arXiv preprint arXiv:1812.08685* (2018).

[58] Marek Kowalski. [n.d.]. faceswap. https://github.com/MarekKowalski/FaceSwap.

[59] Jeff Kukucka, Alexa Hiley, and Saul M Kassin. 2020. Forensic Confirmation Bias: Do Jurors Discount Examiners Who Were Exposed to Task-Irrelevant Information? *Journal of Forensic Sciences* 65, 6 (2020), 1978–1990.

[60] Richard N Landers and Michael B Armstrong. 2017. Enhancing instructional outcomes with gamification: An empirical test of the Technology-Enhanced Training Effectiveness Model. *Computers in human behavior* 71 (2017), 499–507.

[61] Dami Lee. 2019. Deepfake Salvador Dalí takes selfies with museum visitors. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum.

[62] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[63] Justin Lewis, Andrew Williams, and Bob Franklin. 2008. A compromised fourth estate? UK news journalism, public relations and news sources. *Journalism studies* 9, 1 (2008), 1–20.

[64] Yuezun Li, Ming-Ching Chang, Hany Farid, and Siwei Lyu. 2018. In ictu oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv preprint arXiv:1806.02877* (2018).

[65] Yuezun Li, Cong Zhang, Pu Sun, Lipeng Ke, Yan Ju, Honggang Qi, and Siwei Lyu. 2021. DeepFake-o-meter: An Open Platform for DeepFake Detection. In *International Conference on Systematic Approaches to Digital Forensic Engineering (SADFE)*.

[66] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[67] Shao-An Lu. [n.d.]. faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN.

[68] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACVW*. IEEE.

[69] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the work of investigative journalists reporting on problematic information online. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[70] Lea Mets Kiritsis. 2022. Improving the onboarding experience: A qualitative analysis of onboarding slides and tooltips.

[71] Milk Tea Alliance Myanmar [@Milktea_Myanmar]. 2021. Twitter, https://twitter.com/Milktea_Myanmar/status/1374372690128035851.

[72] Konstantin Mitgutsch and Narda Alvarado. 2012. Purposeful by design? A serious game design assessment framework. In *Proceedings of the International Conference on the foundations of digital games*. 121–128.

[73] Tatyana Monnay. 2023. Deepfake Political Ads Are 'Wild West' for Campaign Lawyers. https://news.bloomberglaw.com/business-and-practice/deepfake-political-ads-are-wild-west-for-campaign-lawyers/. Online; accessed 12-September-2023.

[74] Jenna Moon. 2023. Deep fake video of Putin declaring martial law is broadcast in parts of Russia | Semafor. *Semafor.com* (Jun 2023). https://www.semafor.com/article/06/05/2023/putin-deep-fake-broadcast-in-parts-of-russia-declares-martial-law

[75] Scott Moser. 2013. Confirmation bias: the pitfall of forensic science. *Themis: Research Journal of Justice Studies and Forensic Science* 1, 1 (2013), 7.

[76] Kathleen L Mosier, Linda J Skitka, Susan Heers, and Mark Burdick. 2017. Automation bias: Decision making and performance in high-tech cockpits. In *Decision Making in Aviation*. Routledge, 271–288.

[77] Molly Mullen. 2022. A New Reality: Deepfake Technology and the World around Us. *Mitchell Hamline Law Review* 48, 1 (2022), 5.

[78] Belinda Munroe, Thomas Buckley, Kate Curtis, and Richard Morris. 2016. Designing and Implementing Full Immersion Simulation as a Research Tool. *Australasian Emergency Nursing Journal* (2016).

[79] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. 2021. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 923–932.

[80] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.

[81] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. arXiv:1906.06876 [cs.CV]

[82] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2018. Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. *arXiv preprint arXiv:1810.11215* (2018).

[83] Omar Oakes. 2019. "Deepfake" voice tech used for good in David Beckham malaria campaign. https://www.prweek.com/article/1581457/deepfake-voice-tech-used-good-david-beckham-malaria-campaign.

[84] Margit E Oswald and Stefan Grosjean. 2004. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* 79 (2004).

[85] Divya Padalia. 2014. Conformity bias: A fact or an experimental artifact? *Psychological Studies* 59 (2014), 223–230.

[86] Nay Paing. 2021. Is this guy for real? In Myanmar, the fear of deepfakes may be just as dangerous. https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/. *Coconuts Yangon* (May 2021).

[87] Edward A Parson. 1996. What Can You Learn From a Game. *Wise Choices: Games, Decisions, and Negotiations. Harvard Business School Press, Boston* (1996).

[88] Gordon Pennycook and David G Rand. 2019. Lazy, Not Biased: Susceptibility to Partisan Fake News is Better Explained by Lack of Reasoning than by Motivated Reasoning. *Cognition* 188 (2019).

[89] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

[90] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22.

[91] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* (2023).

[92] TRPG Resources. [n.d.]. Basic Rules for Dungeons & Dragons. https://dnd.wizards.com//what-is-dnd/basic-rules

[93] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *arXiv preprint arXiv:1901.08971* (2019).

[94] Johnny Saldaña. 2021. *The Coding Manual for Qualitative Researchers*. SAGE. 1–440 pages.

[95] Shaikh Akib Shahriyar and Matthew Wright. 2022. Evaluating Robustness of Sequence-based Deepfake Detector Models by Adversarial Perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes*. 13–18.

[96] Bingyu Shen, Brandon RichardWebster, Alice O'Toole, Kevin Bowyer, and Walter J Scheirer. 2021. A study of the human perception of synthetic faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.

[97] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*. 7135–7145.

[98] Jessica Silbey and Woodrow Hartzog. 2018. The upside of deep fakes. *Md. L. Rev.* 78 (2018), 960.

[99] Craig Silverman. 2013. New research details how journalists verify information. https://www.poynter.org/reporting-editing/2013/new-research-details-how-journalists-verify-information/

[100] Laura Smalarz, Stephanie Madon, Yueran Yang, Max Guyll, and Sarah Buck. 2016. The perfect match: Do criminal stereotypes bias forensic evidence analysis? *Law and Human Behavior* 40, 4 (2016), 420.

[101] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards Robust Open-World Detection of Deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2613–2615.

[102] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2020. DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection. In *Computation + Journalism Symposium*.

[103] Sarah V Stevenage and Alice Bennett. 2017. A biased opinion: Demonstration of cognitive bias on a fingerprint matching task through knowledge of DNA test results. *Forensic Science International* 276 (2017), 93–106.

[104] Denis Teyssou, Jean-Michel Leung, Evlampios Apostolidis, Konstantinos Apostolidis, Symeon Papadopoulos, Markos Zampoglou, Olga Papadopoulou, and Vasileios Mezaris. 2017. The InVID Plug-in: Web Video Verification on the Browser. In *Proceedings of the first international workshop on multimedia verification*. 23–30.

[105] Pablo Tseng. 2018. What Can the Law do About Deepfake. https://mcmillan.ca/insights/what-can-the-law-do-about-deepfake/.

[106] Loreben Tuquero. 2023. How a deepfake video of Ron DeSantis dropping out of the presidential race went viral. https://www.poynter.org/fact-checking/2023/how-a-deepfake-video-of-ron-desantis-dropping-out-of-the-presidential-race-went-viral/. Online; accessed 12-September-2023.

[107] Hans Von Der Burchard. 2018. Belgian socialist party circulates 'deep fake' Donald Trump video. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/. [Online; accessed 27-January-2020].

[108] Karin Wahl-Jorgensen and Matt Carlson. 2021. Conjecturing Fearful Futures: Journalistic Discourses on Deepfakes. *Journalism Practice* 15, 6 (2021), 803–820. https://doi.org/10.1080/17512786.2021.1908838 arXiv:https://doi.org/10.1080/17512786.2021.1908838

[109] Peter C Wason. 1960. On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology* 12, 3 (1960), 129–140.

[110] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. 2019. What.Hack: Engaging Anti-phishing Training Through a Role-Playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[111] Christine T Wolf. 2019. Explainability Scenarios: Towards Scenario-based XAI Design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 252–257.