2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #33588

# A Pilot Study Investigating STEM Learners' Ability to Decipher AI-generated Video

**Mr. Dule Shu, Carnegie Mellon University**

Dule Shu is a PhD student in Mechanical Engineering at Carnegie Mellon University. His research interest is machine learning, especially using deep neural network models for data generation.

**Dr. Christopher Doss, RAND Corporation**

Christopher Doss is an Associate Policy Researcher at RAND who specializes in fielding descriptive and causal studies in education. His research includes evaluations of early childhood policies, educational technology interventions, interventions that leverage behavioral economics, and interventions to improve teacher instruction. Dr. Doss has published RAND reports and articles in top education and economics journals.

**Dr. Jared Mondschein, RAND Corporation**
**Denise Kopecky, Challenger Center**

Denise Kopecky leads Challenger Center's team of education and technology experts whose work is at the core of the organization's STEM education mission. She oversees the development and implementation of all education products and programs, including curriculum, professional development, and assessment. She also manages relationships with program collaborators including partners, external vendors and evaluators. Ms. Kopecky is a member of the Senior Leadership team, providing direction on all strategic and operational issues to ensure an effective and coordinated effort to meet Challenger Center's strategic goals. Prior to joining Challenger Center, Ms. Kopecky spent 13 years in the classroom, having taught both upper and lower elementary grade levels. She developed and delivered curriculum and led professional development courses. Ms. Kopecky holds a professional certification in Instructional Design from University of Wisconsin-Stout, as well as a Bachelor of Science in Psychology, and a Master of Teaching from Virginia Commonwealth University.

**Ms. Valerie A. Fitton-Kane, Challenger Center**
**Dr. Lance Bush, Challenger Center**

Lance Bush is President and CEO of Challenger Center. With a goal to inspire more students, Dr. Bush has led the growth and expansion of the organization, including the development of a simulation-based program that can be delivered in the classroom. Under Dr. Bush's leadership, Challenger Center was recognized with the National Science Board's Public Service Award for its work to promote a public understanding of science and engineering.

Dr. Bush started his career at NASA as one of the chief engineers designing the next generation space transportation. He managed the International Space Station Commercial Development program. He also co-founded and served as the Chairman of the International Space Station Multilateral Commercialization Group comprised of the five partner space agencies (Canada, Europe, Japan, Russia, and the United States) and 16 countries.

Prior to Challenger Center, Dr. Bush served as the Chief Strategic Officer at Paragon Space Development Corporation, a space vehicle design and build company. He helped grow and mature the firm to become nationally recognized, and as a result, Paragon was included on the Inc. 5000 Fastest Growing Companies list for five straight years.

Dr. Bush is a member of the Cosmos Club, President of Sea Space Symposium, and Founder and an Advisory Board Member of Space Generation Advisory Council. He has previously held positions including Board Member of International Space University, Chair of Education for World Space Congress, and Vice President of Education for American Astronautical Society.

Dr. Bush holds a bachelor's degree in aerospace engineering from the Pennsylvania State University, a master's degree in mechanical engineering from Old Dominion University, and a Ph.D. in technology

policy and management from the Pennsylvania State University, along with an SSP from International Space University.

**Dr. Conrad Tucker, Carnegie Mellon University**

Conrad Tucker is a professor of mechanical engineering at Carnegie Mellon University. He focuses on the design and optimization of systems through the acquisition, integration, and mining of large scale, disparate data.

# A Pilot Study Investigating STEM Learners' Ability to Decipher AI Generated Video

# A Pilot Study Investigating STEM Learners' Ability to Decipher AI Generated Video


## Abstract

Artificial intelligence (AI) techniques such as Generative Neural Networks (GNNs) have resulted in remarkable breakthroughs such as the generation of hyper-realistic images, 3D geometries, and textual data. This work investigates the vulnerability of science, technology, engineering, and mathematics (STEM) learners to AI-generated misinformation in order to safeguard the public-availability of high-quality online STEM learning content. The COVID-19 pandemic has increased STEM learners' reliance on online learning content. Consequently, safeguarding the veracity of STEM learning content is critical to ensuring the safety and trust that both STEM educators and learners have in publicly-available STEM learning content. In this study, state-of-the-art AI algorithms are trained on a specific STEM context (i.e., climate change) using publicly-available data. STEM learners are then randomly presented with authentic and AI-manipulated STEM learning content and asked to judge the authenticity of the content. The authors introduce an approach that STEM educators can employ to understand correlations between STEM learning topics such as climate change, and students' susceptibility to AI-driven misinformation. The proposed approach has the potential to guide STEM educators as to the STEM topics that may be more difficult to teach (e.g., climate change), given students' susceptibility to AI-driven misinformation that promotes controversial viewpoints. In addition, the proposed approach may inform students themselves as to their susceptibility to AI-driven STEM misinformation so that they are more aware of AI's capabilities and how they could be utilized to alter their viewpoints on a STEM topic.


## 1. Introduction

The rapid expansion and adoption of communication technologies has led to the dissemination of information at ever increasing scales and speeds [1]. From a science, technology, engineering, and mathematics (STEM) education perspective, this unprecedented level of access to information has the potential to transform the manner in which students learn and engage with one another. This is particularly evident during the COVID-19 pandemic, as digital communication continues to serve as the primary medium for STEM educational knowledge exchange among both educators and learners [2]. In a brick-and-mortar classroom learning environment, an instructor may be able to guide students' access to information by limiting certain technological resources (e.g., a no cell phone during classroom instruction policy). In online instruction however, guiding students towards certain information sources may be more challenging due to the asynchronous nature of online instruction, coupled with the challenges of enforcing rules and policies in a remote setting [3]. These challenges are compounded by the fact that students regularly engage with technology outside of the classroom and can consume large amounts of information from various sources [4]. Furthermore, K-12 students are digital natives who use YouTube indiscriminately to assist themselves outside of the classroom in completing assignments, with varying degrees of judgement for the reliability of the sources [5]. The

emergence of ubiquitous computing has created, among many things, the ability for every-day individuals to disseminate digital content at scale. For example, there are over one billion videos existing on YouTube [6] with content ranging from Einstein's theory of relativity [7] to how to train a new dog [8]. However, an abundance of data does not automatically translate into an abundance of knowledge. Formally, *knowledge* is defined as: *"facts, information, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject"* [9]. Therefore, data becomes knowledge when it is acquired by a learner and demonstrated via practice.

Advancements in Artificial intelligence (AI) methods such as Generative Neural Networks (GNNs), have resulted in the ability to generate hyper-realistic data including images, videos and text, data types that are commonly used to teach in both brick-and-mortar and virtual learning environments [10]. GNNs have resulted in remarkable breakthroughs such as the creation of artwork [11], generation of 3D engineering designs [12], and the generation of educational game levels [13]. These breakthroughs present both an opportunity and a challenge to online STEM learning. On one hand, the ability of GNNs to generate hyper realistic data has the potential to personalize the delivery of STEM educational content by diversifying the contexts that are presented to learners beyond what a single instructor may be able to achieve due to scalability constraints. On the other hand, GNNs used for nefarious purposes, may inject misinformation within publicly-available STEM educational platforms and other common methods of online communication. As a result, there is a potential increased risk that students are exposed to, and believe inaccurate STEM information, reducing user trust in both the platforms and the STEM educational content.

This paper seeks to investigate the susceptibility of learners to AI-generated STEM educational learning content. The authors introduce an approach that STEM educators can employ to understand correlations between STEM learning topics such as climate change, and students' susceptibility to AI-driven misinformation. In addition, the proposed approach may inform students themselves as to their susceptibility to AI-driven STEM misinformation so that they are more aware of AI capabilities and how they could be utilized to alter their viewpoints on a STEM topic. The contributions of this paper are summarized below:

- A survey tool is designed that utilizes a GNN to synthesize fake videos of celebrities making statements opposite to their real beliefs on a specific STEM topic (e.g., climate change). To the best of our knowledge, this is the first time such a survey design is introduced in the context of STEM education and varying student populations (K-12 and undergraduate students).
- A data feedback loop enabling STEM educators and AI-researchers to quantify the differences between how humans determine what aspects make information real/fake (e.g., AI-generated movement of eyes in a STEM video), and how AI algorithms determine what aspects make information real/fake (e.g., changing certain pixels on a video to minimize a mathematical loss function).

- Preliminary survey results that provide insights as to the differences that may exist in how different STEM learner populations respond to AI-generated STEM content pertaining to STEM topics such as climate change.

## 2. Literature Review

### *2.1. Generative Neural Networks*

Generative neural networks, a neural network implementation of generative models [14] are a class of deep neural networks used for generating user-specified data. In general, deep neural networks are a type of nonlinear computational model which are developed to accomplish the tasks of data classification or data generation. As shown in Figure 1, a deep neural network model consists of many elementary computation units named nodes or neurons. Each node or neuron is a mathematical function that often contains parameters referred to as weights. In a neural network model, the nodes are grouped into multiple layers, and the layers of nodes are connected in a sequential manner to implement the computational procedure of the model. The word "deep" in deep neural network or deep learning refers to the model having a large number of sequentially connected layers, which is a typical design in the state-of-the-art neural networks. Given an $m$-dimensional input $x$, a neural network model, denoted as $f_W$ where $W$ represents the set of weights in the model, defines a mathematical function that maps $x$ to an $n$-dimensional output $\hat{y}$, e.g., $\hat{y} = f_W(x)$. For example, in the case of video transformation using neural networks, $x$ represents the values of all pixels used to construct the image frames in an input video, and $\hat{y}$ represents the values of all pixels used to construct the image frames in the output video. The developer of a neural network model specifies a target model output, denoted as $y$, and uses numerical algorithms to search for the value of the weights $W$ which makes $\hat{y}$ approximate to $y$ as closely as possible. The process of searching for $W$ is referred to as the *training* of the neural network model, and the difference between $y$ and $\hat{y}$ is quantified by a mathematical function called the *loss function*. More information regarding the architecture and the application of deep neural networks can be found in [15].

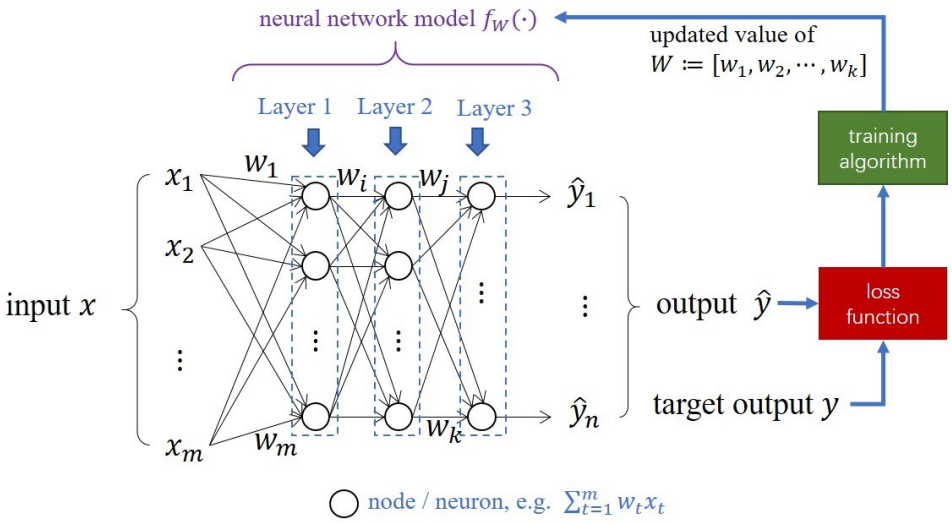



***Figure 1: An Example Architecture of a Neural Network Model***

To exploit the GNN's ability to generate user-specified data, researchers have proposed various GNN models for generating image and video content that seem authentic from a human observer's perspective [16]. Such content is often referred to as "deepfakes", since they are synthesized by deep neural networks rather than by recording devices that capture optical information from the real world. One common form of deepfake content is manipulated human imagery. Methods to generate manipulated imagery can be divided into three groups [17]: i) *Editing and Synthesis*, where a GNN is trained to encode different features of a human (e.g., race, gender and hairstyle) and a user can control the combination of these encoded features to create a new image (e.g., a new human face with user-customized race, gender and hairstyle) [18]; ii) *Replacement*, where a GNN is trained to identify different components of an image and replace a selected component in one image with the counterpart component in another image (e.g., replacing a person's dress in one image with another person's dress from a different image) [19]; and iii), *Reenactment*, where a GNN is trained to match the related components in a video (named the driving video) and an image (named the source image), such that a new video is generated as the animation of the source image by the driving video [20]. The Reenactment method is often used to swap people's faces in an input video. In this case, the related components to be matched are various features of a face including the eyes, the nose, the mouth, the hair, etc. The animation is done by making the face features in the source image track the motion of the face features in the driving video. Among the Reenactment GNN models, the First Order Motion Model proposed in [20], has been chosen in this paper to produce AI-generated videos to investigate learners' susceptibility to AI-generated contents. Compared with other GNN models, the First Order Motion Model has a stable performance in video manipulation and has multiple selections of pre-trained checkpoints on different training datasets. The overall process of generating the fake videos in the survey has been illustrated by Figure 2.

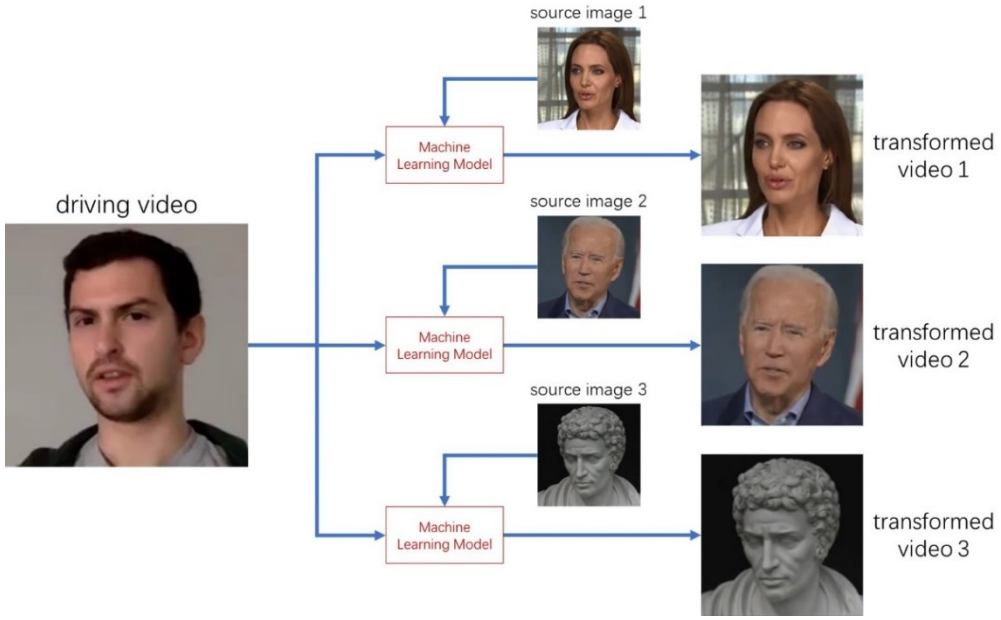



***Figure 2: Example Process of Generating Fake Videos in the Survey***

*2.2 Misinformation in the Digital Age.*

Exposure to misinformation, defined as information that is incorrect (by accident or deliberately), is of increasing concern in the United States [21]. Exposure to and belief in the authenticity of misinformation, have been shown to implant false memories of previous events [22], foster doubts about scientific consensus [23], and convince individuals of the veracity of conspiracy theories [24]. Unfortunately, both youths and adults tend to struggle to identify digital misinformation in a variety of contexts. For example, one small study found that high school students were readily convinced by a blog post containing misinformation about vaccine safety [25], often relying on flawed science-based reasoning. Further work examining readers' vulnerability to misinformation in news articles determined that their propensity to correctly identify authentic content was modulated by their individual view on the topic of discussion [26]. Several additional studies have found that images, including scenery and facial images, are particularly difficult for adults to distinguish as misinformation from authentic media [27]. Further, inoculating individuals against misinformation is a complex target. Recent studies have indicated that relatively simple techniques such as explicitly identifying the false information or providing warnings regarding the potential presence of false information tend to decrease individuals' ability to correctly identify misinformation. However, preliminary results indicate that training adults to identify manipulated images could improve their ability to correctly identify inauthentic facial images [28].

GNN-produced deepfakes threaten to lower barriers to achieving the mass synthesis of manipulated digital content, democratizing the production of misinformation [29]. To date, studies investigating vulnerabilities to deepfakes have largely been limited to adult populations and sociopolitical contexts [30], [31]. For example, one study tested Danish citizens' ability to correctly identify a deepfake video of a politician [32]. Here, the researchers found that viewers

often relied on the veracity of the politician's speech content to identify inauthentic videos; characteristics of the video itself did not influence the viewers. A second study found that ~16% of adults were tricked into believing that a deepfake video of President Barack Obama disparaging then-candidate Donald Trump was authentic and ~30% expressed uncertainty [31]. Ongoing research efforts have yet to study the vulnerabilities of K-12 or undergraduate students to deepfakes or of adults within educational contexts.

### *2.3 Cybersecurity in the Educational Context*

According to the National Center for Education Statistics, more than 56 million students are currently enrolled in K-12 public and private schools in the United States. In addition, there are over 16 million students enrolled in undergraduate programs at degree-granting postsecondary institutions [33], [34], as well as approximately three million students enrolled in postbaccalaureate degree programs. Most of these students have access to the internet in school and at home [34]. As of 2019, 99.2% of K-12 schools have the internet access needed to make digital learning available in their classrooms [35] and 73% of U.S. adults have high-speed broadband internet service at home [36]. Further, many of these students are using the internet to complete schoolwork. In 2015, a Pearson survey of 2,300 K-12 students showed that 53% of fourth and fifth grade students, 66% of middle school students, and 82% of high school students regularly used a smartphone, and 41% said they used a smartphone twice a week to complete schoolwork. Further, a 2017 survey found that over 71% of K-12 teachers allowed students to research subjects using the internet, and 58% used educational apps [37]. Technology use in education was projected to increase at that time and was known to have dramatically increased when schools closed during the COVID-19 pandemic [38].

Similar to corporations, schools can control the applications and websites their users access on school devices and networks. However, this approach becomes more challenging when learners are off-campus and not utilizing school networks/devices. Per the K-12 Cybersecurity 2019 Year in Review, there were 348 publicly-disclosed incidents involving 336 educational agencies across 44 states in 2019 – three times as many as there were in 2018. Key issues include the lack of qualified cybersecurity professionals working in schools, the absence of a common standard of practice or risk management framework to which most school districts adhere, and few cybersecurity tools designed to specifically meet the needs of the K-12 context [39].

Further, many students use personal devices on non-school networks (e.g., mobile phone networks, home networks) to conduct schoolwork both inside and outside of schools and use the internet and various digital media platforms for personal use [40]. Schools have little control over the content accessed from those devices or on non-school networks.

## 3. Methods

### *3.1 Study Population*

Participants were recruited from Carnegie Mellon University and K-12 schools to complete the survey. In addition to the Carnegie Mellon University and K-12 student populations, the RAND Corporation team solicited feedback pertaining to the survey design and instrumentation from

internal experts. The study team at Carnegie Mellon University recruited undergraduate students from the College of Engineering, the Cybersecurity Institute, and the Institute for Politics and Strategy. A total of 13 students from Carnegie Mellon University responded to the survey.

Challenger Center for Space Science Education (Challenger Center) recruited two groups of middle school students from K-12 schools, one in Missouri and one in Illinois, to participate in the pilot study. The students were attending school virtually, so they completed the survey at home using school and personal computing devices. A total of 37 students completed the survey.

*3.2 Experimental Setup and Survey*

The survey was approved by the Institutional Review Boards (IRBs) for Carnegie Mellon University, RAND Corporation, and Challenger Center, and consent was obtained for each survey participant. Adults and students enrolled in higher education were presented with a consent form that they were asked to carefully review and sign prior to their participation in the survey. Parental consent was obtained for all survey participants from K-12 schools.

Table 1 lists the questions that were included in the survey instrument. Additional questions collected background demographic information (e.g., age, gender, etc.). The surveys fielded to Carnegie Mellon University and middle school students were identical except that middle school students were not asked about their political orientation.

**Table 1**. Questions that were included in this survey instrument and the source they were adapted from.

| Survey Question | Source the Question was Adapted From |
|---|---|
| How confident are you that the video was fake or real? | [32] |
| Which aspects of the video below helped you decide if the video was real or fake? | [41] |
| What is your view of scientists' understanding of global warming? | [42] |
| What is your personal view of global warming? | [42] |
| What does the "greenhouse effect" refer to? | [42] |
| Which gases in the atmosphere are good at trapping heat from the Earth's surface? | [42] |
| What causes ocean acidification? | [42] |
| What best describes your political orientation? | [43] |
| On average, which sources do you use to learn about climate change and how often do you use them? | [44] |
| How much do you trust the news and information about climate change that you learn from those sources? | [44] |
| Outside of school or work, which device do you primarily use to access the internet? | [45] |
| Which social media platforms do you use and how often do you use them? | [46] |

| | |
|---|---|
| What device are you using to respond to this survey? | [Developed for this survey] |
| How often do you think fake science, technology, or math information is seen on the Internet? | [Developed for this survey] |
| How much do you think that fake science, technology, or math information on the Internet poses a risk to successfully completing your professional or schoolwork? | [Developed for this survey] |
| How well do you think you are able to detect fake science, technology, or math content? | [Developed for this survey] |
| How much does fake digital science, technology, or math content pose a risk to the overall functioning of society? | [Developed for this survey] |

Carnegie Mellon University used current AI and Machine Learning (ML) techniques to create a bank of eight videos. The bank contained two videos for each of the four people: Timothy Gallaudet, current Assistant Secretary of Commerce for Oceans and Atmosphere and former Acting Administrator of the National Oceanic and Atmospheric Administration; Richard Lindzen, Professor Emeritus of Meteorology at the Massachusetts Institute of Technology; Greta Thunberg, climate change activist; and Naomi Seibt, climate change skeptic. Timothy Gallaudet and Greta Thunberg believe the scientific consensus that climate change is occurring, while Richard Lindzen and Naomi Seibt do not. For each person, the video bank contained an authentic clip where each espoused their views on climate change, and a manipulated clip where each are made to espouse the opposing view. Each survey recipient received one video clip from each person but was randomly assigned either the authentic or the manipulated version. Randomization occurred within person, meaning the same respondent can receive an authentic clip of one person, but a manipulated clip of another.

Randomization means that, in expectation, background characteristics of survey respondents should be uncorrelated with the receipt of authentic or manipulated videos, thus providing an unbiased estimate of the ability to detect deepfake videos. In this paper we explore three outcomes of interest, i) the probability that respondents correctly identified the authenticity of a video, ii) the probability that they incorrectly identified the authenticity of the video, and iii) the probability of responding "I cannot tell." We use the last outcome as a measure of "uncertainty." That is, we interpret an increase in the proportion of respondents who cannot tell the authenticity of the videos as a sign that the fake videos cause uncertainty in the population of respondents. To understand the effect of receiving a manipulated video on the above outcomes, we stack the responses from each video such that observations are at the respondent-video level. We then employ the following Ordinary Least Squares models:

$$Y_{ip} = \beta_0 + \beta_1 Manipulated_{ip} + \alpha_i + \varepsilon_{ip} \quad (1)$$

Where $Y_{ip}$ represents the outcome of interest for respondent, $i$, on video of person, $p$; $Manipulate_{ip}$ is an indicator for getting the manipulated version of a person's video; $\alpha_i$'s are respondent fixed effects that control for all stable person level characteristics; and $\varepsilon_{ip}$ is a person-

video level idiosyncratic error term. We cluster our standard errors at the respondent level to account for the fact that there is more than one observation per individual, which causes responses within individuals to be correlated. The coefficient of interest is $\beta_1$ which is the effect of receiving the manipulated video on the probability of responding as indicated by the outcome. Due to the small sample sizes, we cannot make firm inferences, as all results are statistically insignificant. We therefore present stacked bar charts of responses by authentic and manipulated video so that differences in responses can be compared. However, we see these analyses as exploratory and suggestive. Firmer conclusions will be made after the full survey effort is completed. A contribution of this paper is the design of the experimental survey that enables STEM educators to repeat/reproduce the results as well as to explore research questions and AI-generated STEM content beyond what is presented in this work.

We are also interested in how individuals' views, as measured by responses to the contextual questions, moderate their ability to detect deepfake content. In this paper we explore this relationship among two views: (1) whether or not the person believes that deepfakes are "common" or "everywhere" as opposed to "doesn't exist," "very rare," or "fairly common"; and (2) whether a respondent believes they can identify deepfake videos "most" or "all" of the time as opposed to "not at all" or "sometimes." We choose these views because of past research which has shown that they can moderate the ability to detect misinformation. We see these analyses as illustrative due to our lack of power to detect statistically significant effects and hence, refrain from exploring a larger set of moderating variables.

We explore how these beliefs moderate effects in two ways. First, we estimate the effect of receiving manipulated videos on each subgroup separately. In these instances, we use models in the form of Equation 1 above on the relevant subgroup of interest. We formally test the difference in effects between subgroups with the following model:

$$Y_{ip} = \beta_0 + \beta_1 Manipulated_{ip} + \beta_2 Belief_i + \beta_3 Manipulated_{ip} * Belief_i + \alpha_i + \varepsilon_{ip} \qquad (2)$$

Equation 2 is identical to Equation 1, except we add a main effect for the person's response to the contextual question of interest and an interaction term between the indicator for receiving a manipulated video and the relevant belief. The coefficient interest is now $\beta_3$ which is an estimate of the difference in effects of the manipulated video on the subgroup that holds a specific belief. For example, the coefficient will indicate the differential effect of receiving a manipulated video on the subgroup of respondents who believe deepfakes are more prevalent (as opposed to less prevalent).

*3.3 Research Questions*

In this paper, we explore two general research question:

1. What is the effect of receiving a manipulated video on a person's ability to correctly identify a video as authentic?
2. How does that effect vary by a person's beliefs regarding the prevalence of deepfake videos and their own ability to detect deepfakes?

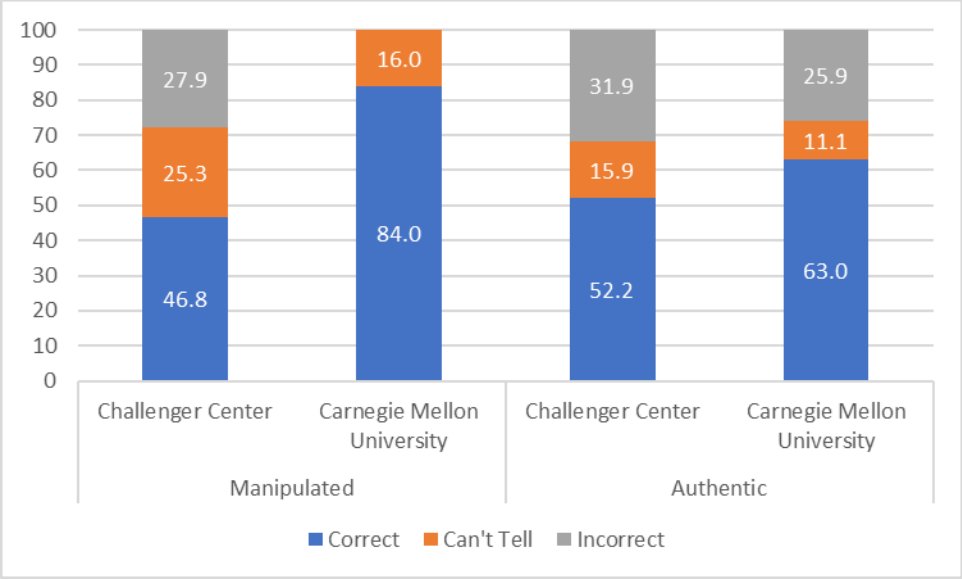
## 4. Results and Discussion

### *4.1 Main Results*

We begin with Research Question 1, which explores the effect of receiving a manipulated video on a respondent's ability to correctly identify the authenticity of a video. Figure 3 below presents the percentage of each type of response among those who received the manipulated videos and among those who received the authentic videos separately for the Challenger Center and Carnegie Mellon University samples. Challenger Center respondents were able to correctly identify the authentic video 52.2 % of the time, while respondents were able to correctly identify the manipulated videos 46.8 % of the time. Previous studies have also found that survey respondents correctly identify authentic and deepfake media with approximately 50% success rates [31]. Meanwhile respondents reported not being able to tell the authenticity of the video 15.9 % of the time when receiving the authentic version and 25.3 % of the time when receiving the manipulated version. Thus, respondents incorrectly identified the authentic videos 31.9 % of the time (calculated as 100% - 52.2% - 15.9% = 31.9%, which corresponds to the grey area of the third column from left to right in Figure 3) and the manipulated videos 27.9 % of the time (calculated as 100% - 46.8% - 25.3% = 29.9%, which corresponds to the grey area of the first column from left to right in Figure 3). In aggregate, the results suggest that the fake videos may have caused more uncertainty which reduced the instances in which respondents correctly identified the authenticity of the video but also decreased the instances in which the respondents incorrectly identified the video.

Results on the Carnegie Mellon University sample differ. Compared with the Challenger Center's result, Carnegie Mellon University's result shows a higher success rate in identifying both the authentic videos (63.0 % vs 52.2 %) and the manipulated videos (84.0 % vs 46.8 %). When receiving the authentic videos, 11.1 % of the time the Carnegie Mellon University participants reported not being able to tell the authenticity of the video, while this rate of uncertainty is 16.0 % when the participants were shown a manipulated video. This result suggests that, among the Carnegie Mellon University participants, the manipulated videos lead to increased instances of correctly identifying either type of videos (authentic or manipulated), despite causing more uncertainty in identification. Note that this is different from what the Challenger Center's results indicate, and such difference is likely caused by the difference of sampled populations (college students with engineering background vs middle school students). It is important to note that in Carnegie Mellon University's survey results, there are no instances of a respondent incorrectly identifying a manipulated video. The authors postulate that this result may be due to student respondents from Carnegie Mellon University being engineering-majored students who already have background knowledge in programming and are aware that the state-of-the-art AI technology is able to generate realistic-looking imagery data.

These findings align with previous studies that have found large amounts of uncertainty among individuals asked to discern between deepfake and authentic digital media, although these literature results are mixed regarding which content – deepfake or authentic – results in greater uncertainty [31], [47]. This uncertainty has cascading effects on consumers' news consumption, often increasing distrust in news content viewed on social media [31].




Notes: Results are derived from 37 middle school students recruited by the Challenger Center and 13 undergraduate students at Carnegie Mellon University judging the authenticity of 4 videos each. No differences in the probability of correctly or incorrectly identifying a video's authenticity or in responding "I can't tell" are statistically significant.

***Figure 3: Judgements of Video Authenticity***

*4.2 Moderation Results*

These overall results could be moderated by different beliefs held by the respondent. We only explore moderation analyses in the Challenger Center sample, as the Carnegie Mellon University sample is too small to look at differences in effects by subgroups. Figure 4 below shows how effects vary by perceived ability to detect deepfakes. We create two subgroups of respondents, those [26], [32], [48] where respondents indicated that they believe that they can detect deepfakes "most" or "all" of the time, and those who believe they can detect deepfakes "sometimes" or "none of the time." Figure 4 shows that those who believe they can identify deepfakes may be more likely to correctly identify the authenticity of the video when receiving a manipulated one. This increase in correct identification may be accompanied with a decrease in incorrect identification and no effect on not being able to tell. Meanwhile those with less confidence may be less likely to correctly identify manipulated videos. This decrease in correctly identifying manipulated video may be mostly explained by a corresponding increase in not being able to tell. Thus, it seems that respondents may be accurately able to judge their ability to detect deepfakes, a finding that agrees with prior survey results which found that individuals with greater self-perceived skills were more likely to correctly identify authentic and fake digital content [49]. However, this question was asked after presenting the videos so participant responses are likely informed by their experiences judging the authenticity of the videos.

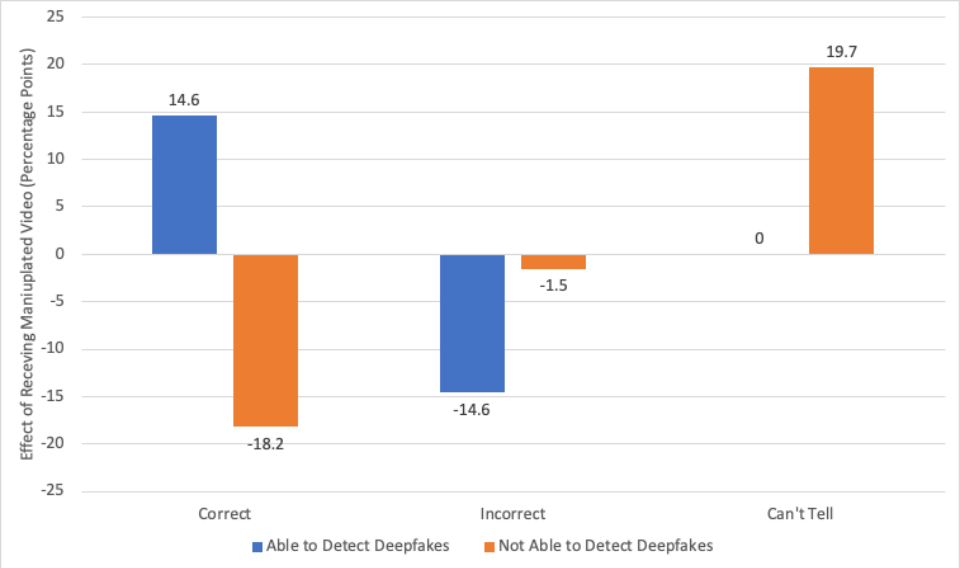



***Figure 4: Differences in Effect of Receiving Manipulated Videos by Perceived Ability to Detect Deepfake Videos***

Notes: Results are derived from 37 middle school students judging the authenticity of 4 videos each. Vertical bars indicate the effect of receiving a manipulated video on the probability of choosing the relevant option regarding the video's authenticity in the subgroup of interest. "Able to Detect Deepfakes" indicates the respondents believe they can detect a manipulated video most or all of the time. "Not Able to Detect Deepfakes" indicates the respondents believe that they can detect deepfake videos sometimes or not at all. No differences in effects are statistically significant.

Figure 5 shows that among those who believe deepfakes are common or everywhere, receiving the manipulated video has small effects on their responses. In contrast, those who believe they are less common may be less able to correctly and incorrectly identify manipulated videos and more likely to be unable to tell. Thus, the uncertainty caused by deepfakes may be concentrated on those who believe deepfakes are less common. This result differs from previous works, which in one study found that prior knowledge of deepfakes did not impact participants' ability to correctly identify authentic and deepfake videos [32], contrasting with a latter study which found that greater awareness of deepfakes tended to increase the likelihood that individuals believed the videos they were watching were fake, regardless of the videos' authenticity [47]. Indeed, this variance across studies indicates that additional work is needed to better understand how prior knowledge and experience affects vulnerabilities to deepfake digital content.

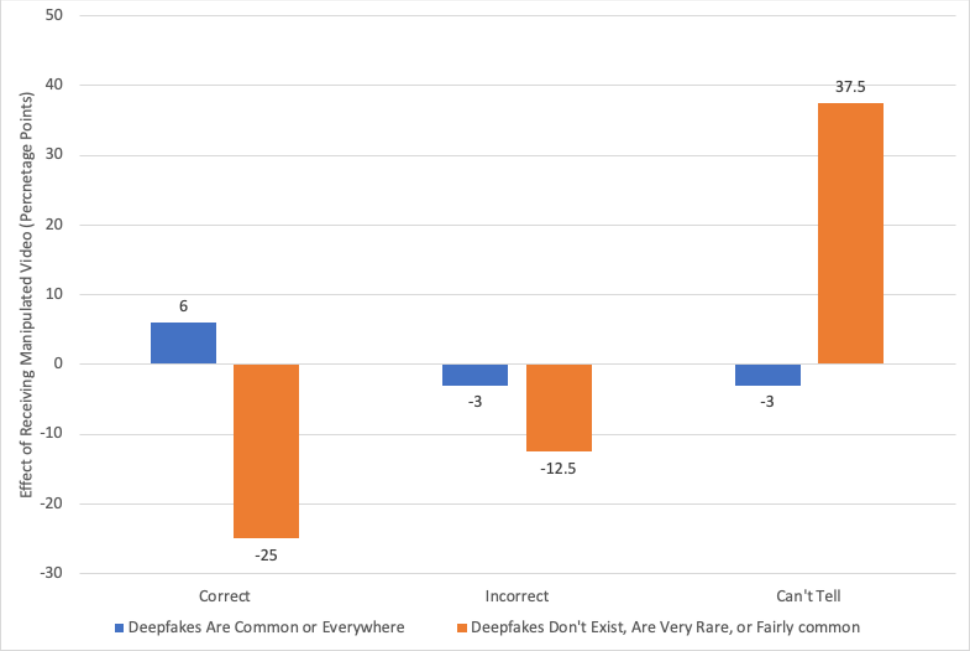



***Figure 5: Differences in Effect of Receiving Manipulated Videos by Perceived Prevalence of Deepfakes***

Notes: Results are derived from 37 middle school students judging the authenticity of 4 videos each. Vertical bars indicate the effect of receiving a manipulated video on the probability of choosing the relevant option regarding the video's authenticity in the subgroup of interest. No differences in effects are statistically significant.

## 5. Conclusion

In an era where learners are surrounded by abundant digital information and open-source AI models, the risk of STEM education being jeopardized by AI-generated misinformation has never been higher. In response to such risk, we launched this study to investigate the susceptibility of learners to AI-generated STEM content. A set of videos about climate change, including both authentic videos and videos manipulated by a state-of-the-art GNN model, were created and fielded to K-12 students and Carnegie Mellon University undergraduate students to test their ability to identify the authenticity of those videos. According to our survey results, the Carnegie Mellon University group has a higher success rate in identifying the video authenticity than the K-12 student group. Among both population groups, the manipulated videos are observed to increase the uncertainty of judgement. The survey results partially support previous finding that the (mis)alignment between a participants' beliefs and video content influences the ability to accurately discern between fake and authentic videos, but in some other aspects, our survey results also differ from previous works. Due to small sample size, we see the results as suggesting relationships which may or may not be confirmed after fielding a larger study. Nevertheless, a major contribution of this work is the design and development of a survey instrument that enables STEM researchers and educators to employ AI-generated technology to study the vulnerability of their learner populations to STEM misinformation. Furthermore, the

data acquired from the survey can help educators quantify the correlations that exist between student learners and STEM topics that are of great interest (e.g., climate change), but may be at risk of being controversial and hence manipulated. Finally, the proposed approach has the potential to assist STEM educators and AI-researchers to quantify the differences between how humans perceive fake/real information and how machines perceive real/fake information. Our future work will focus on creating and fielding a full survey containing more realistic manipulated videos to a much larger pool of middle school students and Carnegie Mellon University students, in addition to nationally representative samples of educators, principles, and US adults.

**Acknowledgement**

This work is supported by the National Science Foundation grant # 2039613. Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

**References**

[1] N. Carbonara, "Information and communication technology and geographical clusters: opportunities and spread," *Technovation*, vol. 25, no. 3, pp. 213–222, 2005.
[2] O. B. Adedoyin and E. Soykan, "Covid-19 pandemic and online learning: the challenges and opportunities," *Interactive Learning Environments*, pp. 1–13, 2020.
[3] G. E. Prestera and L. A. Moller, "Facilitating Asynchronous Distance Learning: Exploiting Opportunities for Knowledge Building in Asynchronous Distance Learning Environments.," 2001.
[4] R. Blair and T. M. Serafini, "Integration of education: Using social media networks to engage students," *Systemics. Cybernetics, and Informatics*, vol. 6, no. 12, pp. 28–31, 2014.
[5] N. Buzzetto-More, "Student attitudes towards the integration of YouTube in online, hybrid, and web-assisted courses: An examination of the impact of course modality on perception," *Journal of Online Learning and Teaching*, vol. 11, no. 1, p. 55, 2015.
[6] "How many videos have been uploaded to YouTube? - Quora." https://www.quora.com/How-many-videos-have-been-uploaded-to-YouTube (accessed Jun. 12, 2017).
[7] "Einstein's Theory Of Relativity Made Easy - YouTube." https://www.youtube.com/ (accessed Jun. 12, 2017).
[8] "Dog Training 101: How to Train ANY DOG the Basics - YouTube." https://www.youtube.com/ (accessed Jun. 12, 2017).
[9] "knowledge: definition of knowledge in Oxford dictionary (American English) (US)." http://www.oxforddictionaries.com/us/definition/american_english/knowledge (accessed Jun. 24, 2015).
[10] R. N. Carney and J. R. Levin, "Pictorial illustrations still improve students' learning from text," *Educational psychology review*, vol. 14, no. 1, pp. 5–26, 2002.
[11] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "ArtGAN: Artwork synthesis with conditional categorical GANs," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3760–3764.
[12] D. Shu *et al.*, "3D Design Using Generative Adversarial Networks and Physics-based Validation," 2019.

[13] C. E. Lopez, J. Cunningham, O. Ashour, and C. S. Tucker, "Deep Reinforcement Learning for Procedural Content Generation of 3D Virtual Environments," *Journal of Computing and Information Science in Engineering*, pp. 1–33, 2020.
[14] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, no. 2002, p. 841, 2002.
[15] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.
[16] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.
[17] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.
[18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *arXiv preprint arXiv:1912.04958*, 2019.
[19] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2e-try on net: Fashion from model to everyone," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 293–301.
[20] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *arXiv preprint arXiv:2003.00196*, 2020.
[21] D. A. Scheufele and N. M. Krause, "Science audiences, misinformation, and fake news," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7662–7669, 2019.
[22] R. A. Nash, "Changing beliefs about past public events with believable and unbelievable doctored photographs," *Memory*, vol. 26, no. 4, pp. 439–450, 2018.
[23] J. Cook, S. Lewandowsky, and U. K. Ecker, "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence," *PloS one*, vol. 12, no. 5, p. e0175799, 2017.
[24] K. L. Einstein and D. M. Glick, "Do I think BLS data are BS? The consequences of conspiracy theories," *Political Behavior*, vol. 37, no. 3, pp. 679–701, 2015.
[25] A. S. Tseng, "Students and evaluation of web-based misinformation about vaccination: critical reading or passive acceptance of claims?," *International Journal of Science Education, Part B*, vol. 8, no. 3, pp. 250–265, 2018.
[26] L. Schaewitz, J. P. Kluck, L. Klösters, and N. C. Krämer, "When is Disinformation (In) Credible? Experimental Findings on Message Characteristics and Individual Differences," *Mass Communication and Society*, vol. 23, no. 4, pp. 484–509, 2020.
[27] S. J. Nightingale, K. A. Wade, and D. G. Watson, "Can people identify original and manipulated photos of real-world scenes?," *Cognitive research: principles and implications*, vol. 2, no. 1, pp. 1–21, 2017.
[28] D. J. Robertson, A. Mungall, D. G. Watson, K. A. Wade, S. J. Nightingale, and S. Butler, "Detecting morphed passport photos: a training and individual differences approach," *Cognitive research: principles and implications*, vol. 3, no. 1, pp. 1–11, 2018.
[29] S. Lewandowsky, U. K. Ecker, and J. Cook, "Beyond misinformation: Understanding and coping with the 'post-truth' era," *Journal of applied research in memory and cognition*, vol. 6, no. 4, pp. 353–369, 2017.
[30] N. I. Brown, "Deepfakes and the Weaponization of Disinformation," *Va. JL & Tech.*, vol. 23, p. 1, 2020.

[31] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media+ Society*, vol. 6, no. 1, p. 2056305120903408, 2020.
[32] T. Dobber, N. Metoui, D. Trilling, N. Helberger, and C. de Vreese, "Do (microtargeted) deepfakes have real effects on political attitudes?," *The International Journal of Press/Politics*, p. 1940161220944364, 2020.
[33] N. C. for Education Statistics, "The NCES Fast Facts Tool provides quick answers to many education questions (national center for education statistics)," 2016.
[34] "The Condition of Education - Postsecondary Education - Postsecondary Students - Undergraduate Enrollment - Indicator May (2020)." https://nces.ed.gov/programs/coe/indicator_cha.asp (accessed Mar. 08, 2021).
[35] "2019 State of the States." https://stateofthestates.educationsuperhighway.org/#future. (accessed Mar. 08, 2021).
[36] 1615 L. St NW, Suite 800Washington, and D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, "Demographics of Internet and Home Broadband Usage in the United States," *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ (accessed Mar. 08, 2021).
[37] "University of Phoenix Survey Finds That American K-12 Teachers Assign Less Homework Than Often Perceived | University of Phoenix." https://news.phoenix.edu/press-release/university-of-phoenix-survey-finds-that-american-k-12-teachers-assign-less-homework-than-often-perceived/ (accessed Mar. 08, 2021).
[38] S. J. Barnes, "Information management research and practice in the post-COVID-19 world," *International Journal of Information Management*, vol. 55, p. 102175, 2020.
[39] D. L. LLC EdTech Strategies, "Year in Review – The K-12 Cybersecurity Resource Center." https://k12cybersecure.com/year-in-review/ (accessed Mar. 08, 2021).
[40] A. Klein, "Schools say no to cellphones in class. but is it a smart move?" Education Week, vol. 39, no. 4, pp.1–10, 2019.
[41] S. Fan, R. Wang, T.-T. Ng, C. Y.-C. Tan, J. S. Herberg, and B. L. Koenig, "Human perception of visual realism for photo and computer-generated face images," *ACM Transactions on Applied Perception (TAP)*, vol. 11, no. 2, pp. 1–21, 2014.
[42] A. Leiserowitz, N. Smith, and J. R. Marlon, "American teens' knowledge of climate change," *Yale University. New Haven, CT: Yale Project on Climate Change Communication*, vol. 5, 2011.
[43] S. K. Yeo, M. A. Xenos, D. Brossard, and D. A. Scheufele, "Selecting our own science: How communication contexts and individual traits shape information seeking," *The ANNALS of the American Academy of Political and Social Science*, vol. 658, no. 1, pp. 172–191, 2015.
[44] H. Cheng and J. Gonzalez-Ramirez, "Trust and the Media: Perceptions of Climate Change News Sources Among US College Students," *Postdigital Science and Education*, pp. 1–24, 2020.
[45] K. Gay, J. Torous, A. Joseph, A. Pandya, and K. Duckworth, "Digital technology use among individuals with schizophrenia: results of an online survey," *JMIR mental health*, vol. 3, no. 2, p. e15, 2016.

[46] K.-S. Kim, S.-C. J. Sin, and T.-I. Tsai, “Individual differences in social media use for information seeking,” *The journal of academic librarianship*, vol. 40, no. 2, pp. 171–178, 2014.
[47] J. Ternovski, J. Kalla, and P. M. Aronow, “Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments,” 2021.
[48] D. J. Flynn, B. Nyhan, and J. Reifler, “The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics,” *Political Psychology*, vol. 38, pp. 127–150, 2017.
[49] C. Shen, M. Kasra, W. Pan, G. A. Bassett, Y. Malloch, and J. F. O’Brien, “Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online,” *New media & society*, vol. 21, no. 2, pp. 438–463, 2019.