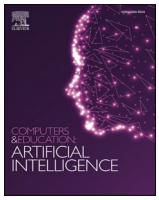
Computers and Education: Artificial Intelligence 2 (2021) 100040

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Children as creators, thinkers and citizens in an AI-driven future

Safinah Ali [a,*], Daniella DiPaola [a], Irene Lee [a], Victor Sindato [a], Grace Kim [a], Ryan Blumofe [b], Cynthia Breazeal [a]

[a] Massachusetts Institute of Technology, USA
[b] Cambridge Rindge and Latin School, USA

ARTICLE INFO

Keywords:
Misinformation
Deepfakes
Generative AI
Digital literacy
Media literacy
Social media

ABSTRACT

Generative Artificial Intelligence (AI) approaches open up new avenues of digital creation, and are simultaneously accompanied by societal and ethical implications such as the creation of Deepfakes and spread of misinformation, renewing our understanding of technical AI systems as socio-technical systems. Applications of, and media generated by generative AI techniques are abundantly present on social media platforms frequented by children, who are not yet aware of the existence of AI-manipulated media. Previous work has highlighted the importance of digital media literacy and AI literacy for children. In this work, we introduce middle school students to generative AI techniques as a tool for creation, while also focusing on critical discussion about their societal and ethical implications, and encouraging pro-activeness in being responsible consumers, creators and stakeholders of technology. We present learning activities that introduce 38 middle-school students to generative modeling, how it is used to generate Deepfakes, cues that help to recognize Deepfakes, and the spread and effects of misinformation. Students demonstrated an understanding that generative media may be believable, but not necessarily true, and can contribute to the spread of misinformation. They were also able to identify why misinformation may be harmful or lasting, drawing specific examples to social settings that indicate human-centered implications. Finally, students expressed opinions about policies surrounding the presence of Deepfakes on social media. This approach can be adopted to introduce students to other technical systems that constitute both productive applications and potential negative implications of technology.

CCS concepts: ·Applied computing → *Interactive learning environments;* ·Human-centered computing → Social media; Social networks; ·Social and professional topics → Computing literacy; *K-12 education*;

Additional key words and phrases: Misinformation, Deepfakes, digital literacy, media literacy, social media.

1. Introduction

In July of 2020, researchers at Massachusetts Institute of Technology released a never-before-seen video of Richard Nixon delivering a speech after the events of the Apollo Space Mission. Nixon, nervously holding a stack of papers, looks directly at the camera and says, *"Good evening my fellow Americans. Fate has ordained that the men who went to the moon to explore in peace will stay on the moon to rest in peace."* In reality, the astronauts on the Apollo mission had successfully landed on the moon and safely returned back to Earth. However, the video portrays a realistic press announcement the president may have given in an alternate history. The video is an example of a deepfake, or a fake video of one or more people that appears to be authentic. The researchers who created this video used a type of artificial intelligence called generative adversarial networks (GANs) to generate Nixon's voice and face. First introduced by Goodfellow in 2014, GANs are generative models that consist of two competing models - the generator model that generate new data instances, and the discriminator model that tries to classify instances as either real or fake (Goodfellow et al., 2014). Over multiple iterations of generating a new image and getting feedback on whether or not the image can pass as real, the system produces more and more realistic fakes. While GANs have several non-harmful applications such as artistic expression, they can also be used to create hyper-realistic Deepfakes, which have deep consequences such as aiding those who seek to rewrite history. With the advancements in generative AI algorithms and availability of large scale datasets, AI-generated media has become more realistic and indiscernible from real media. While GANs can create compelling artwork, or even used for imaging, healthcare,

Abbreviations: AI, Artificial Intelligence; GAN, Generative Adversarial Network.

* Corresponding author.
E-mail address: safinah@media.mit.edu (S. Ali).

https://doi.org/10.1016/j.caeai.2021.100040
Received 29 June 2021; Received in revised form 9 November 2021; Accepted 10 November 2021
Available online 22 November 2021

robotics and accessibility, they have a unique potential to generate and spread misinformation.

The potentially harmful use of technologies like Deepfakes to generate persuasive media is not a new phenomenon. People have doctored images and audio for years without the assistance of artificial intelligence. Yet, there are some key differences in the way that Deepfakes work that make them particularly dangerous tools for spreading of lies and misinformation. GANs, the technology used to produce Deepfakes, produce more realistic media than other types of image manipulation. In fact, Deepfakes were designed by computer scientists to give computers the capability to generate near-real images. Since then, new forms of GANs have been developed to produce even more realistic media. Media created by GANs have positive benefits such as clearer medical imaging, but also pose challenges in the case of Deepfakes (Alqahtani et al., 2021). Over the course of this year, we have seen dramatic changes in the quality of deepfake images, audio and video (Tolosana et al., 2020). Readily available applications on social media platforms like FakeApp,[1] enable users to manipulate their own digital media to their own purposes in short amounts of time.

Today's students are active users of social media platforms such as TikTok and Instagram that allow them to share information with one another. These platforms have been inundated with generative AI such as creative filters and Deepfakes (Dickson, 2020; Kaya, 2019). Research has shown that misinformation, or information that is false regardless of the intent, spreads farther and faster than true information (Vosoughi et al., 2018). Audio and video generated by AI are now present on social media, however, people stay unaware of them, and are likely to mistake them for real. This is especially a risk for children, who are more vulnerable due to their young age. Children begin to form opinions about social and political issues during their middle and high school, and consuming misinformation can perpetuate long-term harm. Digital media literacy efforts have been deployed to teach students how to think critically about the information they consume. However, no current efforts have been made to teach students about deepfake technology, despite its increasing accessibility to the public. Further, current aimed at teaching students about misinformation often focus on source-checking and looking for evidence. While that is an effective approach, the existence of Deepfakes lead to the ability of creating believable fake evidence, rendering the approach less effective. Advances in AI, specifically generative AI, can fuel the spread of misinformation, and in this work, we propose a literacy effort to inform children to better prepare them to encounter AI-generated fake media.

We present interactive classroom activities that aim to make students aware of generative AI, its ability to create Deepfakes, how Deepfakes can be spotted, how it can lead to the spread of misinformation, how misinformation spreads, and some ways to mitigate the spread. We present an exploratory study with 38 middle school students who participated in a week-long virtual workshop to learn about generative machine learning, deep fakes, and their implications. In this paper, we focus on the deepfake and misinformation activities to address the following research questions:

1. What do students already know about deepfake technology and misinformation?
2. Are students able to detect Deepfakes after learning about ways to spot them?
3. Do students understand how misinformation spreads online and are they able to apply that knowledge to Deepfakes?
4. After learning about Deepfakes and misinformation, what deepfake policies do students advocate for?

2. Background

2.1. Generative modeling

Generative models are a class of statistical machine learning models that are trained on large amounts of data (images, text or sounds) to create new data instances that resemble the training data. They are often viewed in contrast with classification algorithms that are trained to differentiate between different kinds of data. Commonly used generative networks include GANs (Goodfellow et al., 2014) and Variational Autoencoders models (VAEs) (Doersch, 2016). Generative models have positive applications such as in image processing, generating art, building AI writing assistants, medical imaging, and collaborative robots (Ali et al., 2020; Gatys et al., 2016; Vera et al., 2020). However, generative models can have malicious applications, such as generating misleading news articles that spread fake news, impersonating other people online, and generating content that reproduce historical bias in data sets.

2.2. Deepfakes

GANs have also been used to create Deepfakes, which are fake videos and audios of people that look real (Korshunov & Marcel, 2018). Deepfakes use machine learning techniques called generative neural network architectures, such as autoencoders or GANs to replace faces of one individual in a video with synthesized faces of another individual (Cao et al., 2019; Zhang et al., 2018). With continual improvements in generative modeling algorithms, Deepfakes become increasingly believable and have a high potential to deceive viewers (Nguyen et al., 2019, p. 11573). For example, in 2019, a deepfake video of Mark Zuckerberg (the founder of Facebook) emerged on social media declaring that he would be deleting Facebook due to security concerns. The video was believed by many, shared widely and gathered over 72 million views on various social media platforms.

In the past, the creation of these synthetic media was limited to individuals with the technical sophistication and computational resources to train large models. With the advent of mobile applications such as Reface,[2] and open source tools such as FakeApp anyone with a smartphone can swap faces to generate synthetic images and videos easily. Deepfakes have been used for several malicious purposes, such as swapping celebrity faces in pornographic videos, creating videos of world leaders in fake speeches (Bloomberg, 2018; Nguyen et al., 2019, p. 11573). Deepfakes have the potential of being misused to cause political or social tensions between countries, to affect results in election campaigns, or create chaos in financial markets by creating fake news (Chesney & Citron, 2019; Kaliyar et al., 2021). It can also be used to create false evidence affecting legal proceedings. There are also positive uses of Deepfakes, such as recreating voices for those who have lost their voice, or entertainment media, but the malicious uses outweigh the positives (Bernard, 2019).

2.3. Misinformation

Misinformation is one type of information disorder where fake news is shared but no harm was intended (Wardle & Derakhshan, 2017). Deepfake technologies are fully generated to misrepresent real images, video or audio, so they often fall under the category of misinformation. Misinformation may have harmful consequences for decision making, especially in democratic societies where governance relies on each person's opinions and independent decisions (Lewandowsky et al., 2017). Though not grounded in harmful intentions, misinformation has the potential to sway elections (Gunther et al., 2018), increase political polarization (Flaxman et al., 2016), and incite violence (Haag & Salam,

[1] https://www.fakeapp.org/.

[2] https://hey.reface.ai/.

2017). Social media platforms allow misinformation to be shared across social networks, often targeted at political institutions (Phartiyal et al., 2018; StarbirdJim Maddock et al., 2014; Van der LindenAnthony et al., 2017). Prior work in the UK found that Deepfakes of political content contributed towards disinformation. The uncertainty caused by Deepfakes may contribute toward generalized indeterminacy and cynicism, further intensifying recent challenges to online civic engagement in democratic societies (Vaccari & Chadwick, 2020). Children are considered vulnerable populations (Mechanic & Tanner, 2007) and are more at risk of believing Deepfakes given their lack of exposure to such manipulated media and their lack of knowledge to contextualize new information. Furthermore, children form several social and political opinions in their formative years and can become targets of disinformation spread by convincing Deepfakes (Torney-Purta, 2017). Thus far, there is also little research suggesting whether children are aware of the existence of AI-generated media that are not real, and an understanding of how they are generated.

2.4. Countering Deepfakes

As Deepfakes become more common, efforts to detect them are also being developed (Nguyen et al., 2019, p. 11573). In previous work, a deepfake detection technique that uses convolutional neural network (CNN) to extract frame-level features from the video and train a recurrent neural network (RNN) was used to classify if a video has been subject to manipulation or not (Güera & Delp, 2018). In 2018, researchers at SUNY developed a technique that detected a lack of eye-blinking in videos to detect Deepfakes (Li et al., 2018). Soon after, Deepfakes emerged that started incorporating blinking, rendering the eye-blinking detection technique obsolete (Agarwal et al., 2019). Hence, with Deepfakes evolving and improving so rapidly, it becomes increasingly difficult to detect them. Furthermore, these detection techniques are not accessible to the public. Researchers compiled signs to look for in media to detect Deepfakes that are understandable by laypersons, such as, unnatural eye movement, or lack of emotions, or teeth that merge together (Johansen, 2020). The R.E.A.L. framework was suggested to counter deepfake risks: Record original content to assure deniability, Expose Deepfakes early, Advocate for legal protection, and Leverage trust to counter credulity (Jan et al., 2020).

Government agencies and corporate entities have also increasingly incorporated policies to detect and notify their users of occurrences of Deepfakes. In 2019, Congress introduced the Defending Each and Every Person from False Appearances by Keeping Exploitation Subject (Deepfakes) to Accountability Act. It proposes that harmful Deepfakes must be labeled if potentially harmful, unless they're intended to help public safety (Coldewey, 2019). Big Tech companies have started proposing their own policies as well. Twitter's policy requires tagging all synthetic media and warns users before they share any fake information (Perez, 2019). Facebook's policy promotes artificial intelligence to detect Deepfakes and deletes them, unless they are used as satire or entertainment (Bickert, 2020). YouTube has banned all political Deepfakes, but allows other types of Deepfakes to stay on the platform (Coble, 2020). However, Big Tech companies and AI researchers continue to discuss the implications of the potential of using Deepfakes generated for malicious intents, and innovating on ways to mitigate them, such as reverse engineering the generative model used from the generative media (Asnani et al., 2021). While *algorithmic* approaches to mitigate misinformation and Deepfakes are in progress, we take an *educational* approach, where we aim to inform potentially vulnerable content consumers about the existence of Deepfakes and how it can be used to spread misinformation.

2.5. Digital media literacy

Digital literacy, the ability to use information and communication technologies to find, evaluate, create, and communicate information, continues to evolve and change in step with digital technology (Buckingham, 2016). In today's era of abundant misinformation, the ability to analyze and validate articles and other forms of media has become an increasingly essential skill for any who use the Internet (Joseph et al., 2012; Paul & Benjamin, 2013). Previous work, however, has demonstrated that teenagers are not proficient at recognizing fake news. A study used a fake website to demonstrate that only 11% school children recognized its hoax source as fake (Leu et al., 2007). The same website was used to demonstrate how 65% students claimed to trust the website (Pilgrim et al., 2019). Methods of checking a source's reliability such as a "credibility checklist" or the CRAAP test to check for the Currency, Relevance, Authority, Accuracy, and Purpose of a website seem appealing, but with the current state of technology and freely-accessible tools, it is far too easy to make believable web content (Blakeslee, 2004). Current media literacy approaches focus on media bias, but not enough on online misinformation (Polizzi & Taylor, 2019). Given that a staggering 75% of students get their news online, new approaches to teaching digital literacy skills that are responsive to a rapidly evolving technology landscape and also address critical civic reasoning skills are very critical (Joel Breakstone et al., 2018).

In response to the growing popularity of the internet, smartphones, social media, video conferencing, and other "disruptive technologies", teachers are opting to integrate these devices into their classes to teach digital literacy skills (Nowell, 2014). Prior works have focused on the spread of misinformation and disinformation. Agosto and Abbas developed tips for helping students become safer and more critical social media users, which included good information sharing practices (Denise & Abbas, 2016). Several education researchers have called for developing digital media literacy tools and curricula that focus on fake news consumption and spread (Lee, 2018; Polizzi & Taylor, 2019). The News Literacy Project's digital learning modules and fact-checking platform *Checkology* aims to teach students how to discern fake news from factual news.[3] Literat, Chang, and Hsu used participatory game design as a media literacy tool. They engaged youth in the design of news literacy games and studied how children understand and engage with fake news (Literat et al., 2020). Roozenbeek and Linden developed a psychological intervention in the form of an online browser game in which players take on the role of a fake news producer and master six common techniques to enhance the spread of misinformation: polarization, invoking emotions, spreading conspiracy theories, trolling people online, deflecting blame, and impersonating fake accounts (RoozenbeekSander van der Linden, 2019). They found that the game improved people's ability to spot and resist misinformation.

The use of artificial intelligence in apps and online media platforms makes digital literacy landscape even more complex. Modern artificial intelligence techniques can be used to create "hyper-realistic" misinformation such as Deepfakes (De Vries, 2020), making it even harder to find credible sources on the Internet. Given the threat that realistic and manipulated media poses to the spread of misinformation, there is an urgent need to develop approaches that augment digital media literacy with deepfake literacy.

While the topic of generative media, Deepfakes and misinformation remain technically complex, we made use of example-based and simulation game-based approaches to make these concepts accessible. Game-based learning has been identified as an emerging topic and effective approach within educational technology research (Chen et al., 2020a, 2020b). In the next section, we discuss the design of three learning activities and outline our approach to combining Deepfakes literacy with digital media literacy., 2020). The knowledge of how to use these applications is important because digital writing is very different from traditional print because it is collaborative, participatory, and rapidly shared through means like social media. In addition to all of these skills, learning acceptable internet behavior is also a crucial part of digital

[3] https://get.checkology.org/.

literacy.

3. Learning activities

We begin by introducing students to generative modeling techniques through examples of generative media, including Deepfakes. Students get a brief introduction to the two neural networks that make up a GAN, a commonly used generative algorithm, and how they are used to create Deepfakes. They also practice techniques to detect Deepfakes. Then, through a news sharing simulation application, students witness what misinformation is, how misinformation spreads, and how Deepfakes can fuel that spread. Finally, students voice their opinions on what policies should be in place to regulate the presence of Deepfakes on social media platforms.

3.1. Activity 1: generative modeling (created by AI or not)

In our first activity, students are introduced to the concept of generative AI through examples of machine-generated media. Students learn that AI can be used to generate synthetic media such as images, text, music, colors, paintings, digits or videos. The goal of the activity is to expose students to applications of generative models as well as to discuss how realistic some machine-generated media can be. We begin the activity by playing a game called "Created by AI or Not". Students are shown 14 examples of machine-generated media (4 style transfer artworks including portraits, 2 deepfake images, 1 deepfake video, 1 color compilation, 1 digits compilation, 1 handwriting generator, 2 text paragraphs and 2 music tracks) in editable Google Slides. Media was chosen to represent multiple modalities and to reflect commonly shared online content, such as a photograph altered using a popular filter application Prisma,[4] a joke generated in the knock-knock joke style, or a generative "Happy Birthday" song. Examples of media can be found in Fig. 1. Students are *not* told that all of the media is created by AI and are given three options to choose from: "Created by AI", "Not created by AI", or "I am not sure". After completing the activity, students discuss which media they thought was created by AI and provide their reasoning. Students are then told that all of the media examples, in fact, were created using generative AI models. We then discuss whether students found this surprising and which artworks seemed particularly unlikely to be machine-generated. We support learning by eliciting an emotional response and triggering cognitive dissonance, or contradicting their conceptualizations of "what is real".

3.2. Activity 2: Spot the Deepfakes

In Activity 1, students familiarize themselves with deepfake media from an assortment of examples (like AI News Anchor [5] and This Person Does Not Exist [6]). In this activity, we define what Deepfakes are by analyzing the examples from Activity 1. Students fill out a questionnaire with 10 different videos (5 real and 5 Deepfakes) and are asked to try to identify the Deepfakes. The videos were taken from the public dataset released by the Kaggle Deepfake Detection Challenge.[7] This activity was followed by a classroom viewing of another series of video clips featuring particularly evocative examples of Deepfakes. We showed a range of videos including deepfake Mona Lisas and fake recordings of opinions voiced by political figures.

Students then discussed their own ideas about how to spot Deepfakes on social media feeds or news sources. We engaged students in open-ended discussion on these topics to suggest to students that these issues are part of an ongoing discourse of collaborative problem-solving. There is no set-in-stone method to identify GAN-generated media and students have a role to play in the broader conversation about the technology. Additionally, we shared with students already established techniques for identifying Deepfakes such as blurry backgrounds or asymmetry in faces (Johansen, 2020). Examples of other components that can help identify Deepfakes can be found in Fig. 2. Finally, students orally discussed the possible societal implications of Deepfakes and how they think harmful uses of Deepfakes should be countered.

With this new knowledge of how to spot Deepfakes, students were asked to complete a follow-up activity where they could put their new knowledge to the test. Students were once again presented with 5 real and 5 deepfake videos. For a given video, students were asked to guess whether a video was real or fake and write down the reasoning behind their assessment. In this way, we provided a framework with which they could critically think about the validity of media, and we gave them the opportunity to practice observing these components in real life examples.

This exercise helps to prepare students to spot synthetic media in the future, as they will continue to be exposed to Deepfakes online. It also serves to help them build general intuition around what specific features GANs can manipulate in an image or video. By breaking down the flaws and gaps in GAN-produced media, students gain a deeper understanding of what they should look for when questioning online content as well as why they should be questioning what they see on social media.

3.3. Activity 3: school-book app: simulating the spread of misinformation

With the goal of getting students familiarized with what misinformation means, how it spreads, and how Deepfakes can fuel the spread of misinformation, we had students participate in an activity with a mock-up of a news sharing platform. We designed a web application called School-Book that students interacted with synchronously. The application simulated a Twitter-like interface, where students could see a set of posts emulating school-related news items that students could choose to share. To begin, students logged in with their name and were presented with six one-liner headlines, three of which had characteristics of true information and three of which had characteristics of misinformation. From the given set of headlines, students could choose to share one headline with their classmates. In the next round, the page refreshed to display the headlines shared by others. The number of headlines in round 2 equalled the number of students. Next, students were again invited to share one piece of news with their network. The items shared the most (top 50%) stayed visible in the game for subsequent rounds while the other half was discarded. Students continued playing until three or fewer headlines remained in the game.

All headlines used in our School-Book app were related to school uniforms being introduced into their school. Half of the posts were neutral conveying rules on school policy or clearly articulated opinions on said policies. For example, "Hey everyone! Friendly reminder that school uniforms must be worn on 4 days of every school week". The other half of the posts contained features of misinformation such as polarization, invoking emotions, spreading conspiracy theories, trolling people online, deflecting blame, and impersonating fake accounts. For example, "Word has it that we must pay for school uniforms made out of real animal fur." They often framed opinionated statements as fact (see Fig. 3).

After the game ends, students viewed an online visualization of the spread of shared messages in the School-Book App. This online visualization was structured like a flow chart moving horizontally from left to right (Fig. 5). The first layer of the flow chart was comprised of all possible feed updates. Each subsequent layer consisted of only the social media updates that the students had chosen to move forward. In the visualization at the end of the game, each of the social media posts was color-coded to indicate whether or not they had characteristics of misinformation. We used this feature so that students could easily see

[4] https://prisma-ai.com/.
[5] https://www.theguardian.com/world/2018/nov/09/worlds-first-ai-news-anchor-unveiled-in-china.
[6] https://thispersondoesnotexist.com/.
[7] https://www.kaggle.com/c/deepfake-detection-challenge.

Fig. 1. Students see examples of media created by generative models and try to guess if they are created by AI or not. Left to right: a. Image style transfer photo filter. b. Deepfake image from This Person Does Not Exist. c. Generated digits.

Fig. 2. Example techniques of how to recognize Deepfakes in deepfake images (Ali et al., 2021a).

what type of social media posts propagated the farthest and how the ratios between misinformation and real posts changed from round to round.

Students then discussed their interactions with the application along the following prompts presented by the instructors:

- How did you choose what information to pass along?
- What do you think is the difference between the red and green items?
- What are the qualities of the final three headlines? How are they different from the other headlines?
- What do you notice about the headlines that get passed forward vs. the ones that did not?

Post discussion, the students compiled their reflections on their consumption and sharing of news in the real-world on an online forum called Padlet [8] with the following prompts:

- Where do you typically get your news from?
- How do you know when information is not true?
- What percentage of your news feeds is true information?
- Is it difficult to tell the difference between true and misinformation? Why or why not?
- Do you think Deepfakes will travel farther than true videos?

Students discussed the answers as a class and ended with a discussion about how Deepfakes may impact the spread of misinformation.

4. Methods

4.1. Study setting

We tested these activities as part of a larger 5-day virtual workshop to teach middle school students about generative machine learning. Due to COVID-19 emergency restrictions, the workshop study was conducted remotely using synchronous online learning on Zoom with other digital tools. All activities were facilitated by one instructor on Zoom and supported by three teaching assistants. Students participated in class by speaking out loud or writing in Zoom's chat window.

4.2. Participants

A total of 38 middle school and high school students (ages 10–15 years old, 18 female and 20 male) from five states across the United States participated in the Amazon Future Engineers program that consisted of two separate but identical summer workshops. 37 students attended Title-1 public schools. One student did not report their school. 28 students (68.42%) came from demographic groups that are under-

[8] https://padlet.com/.

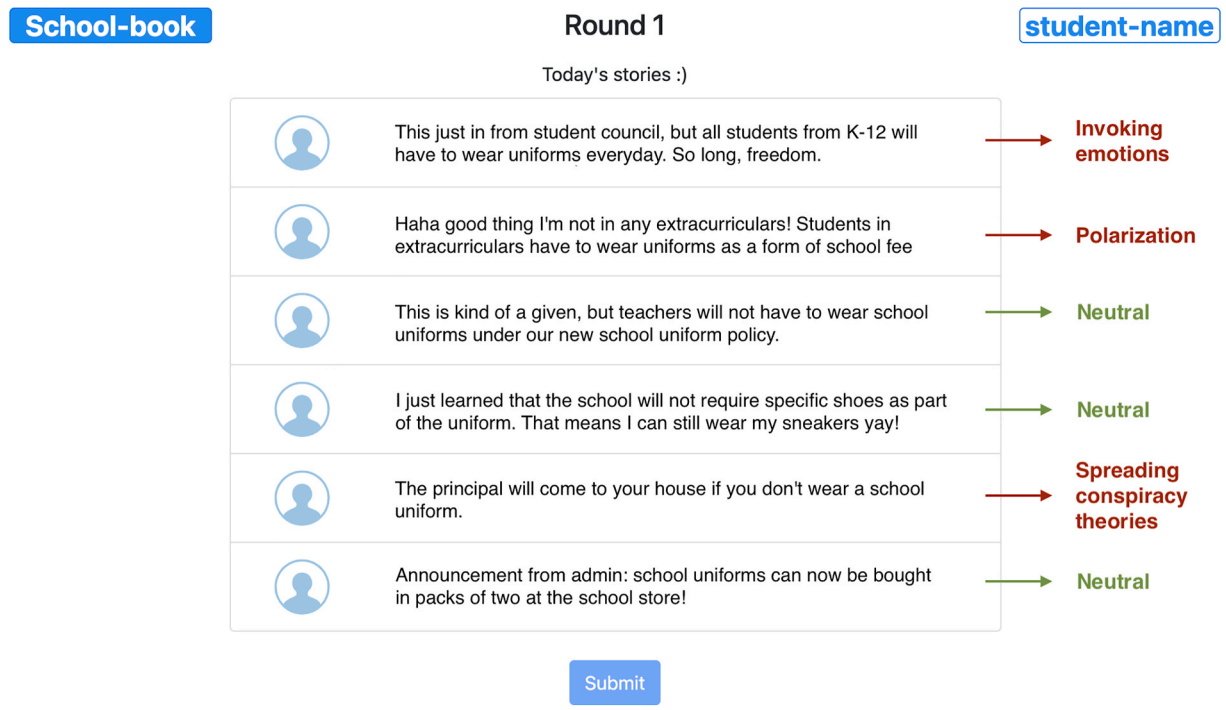


Fig. 3. School-book web application view (overlaid with the type of information).

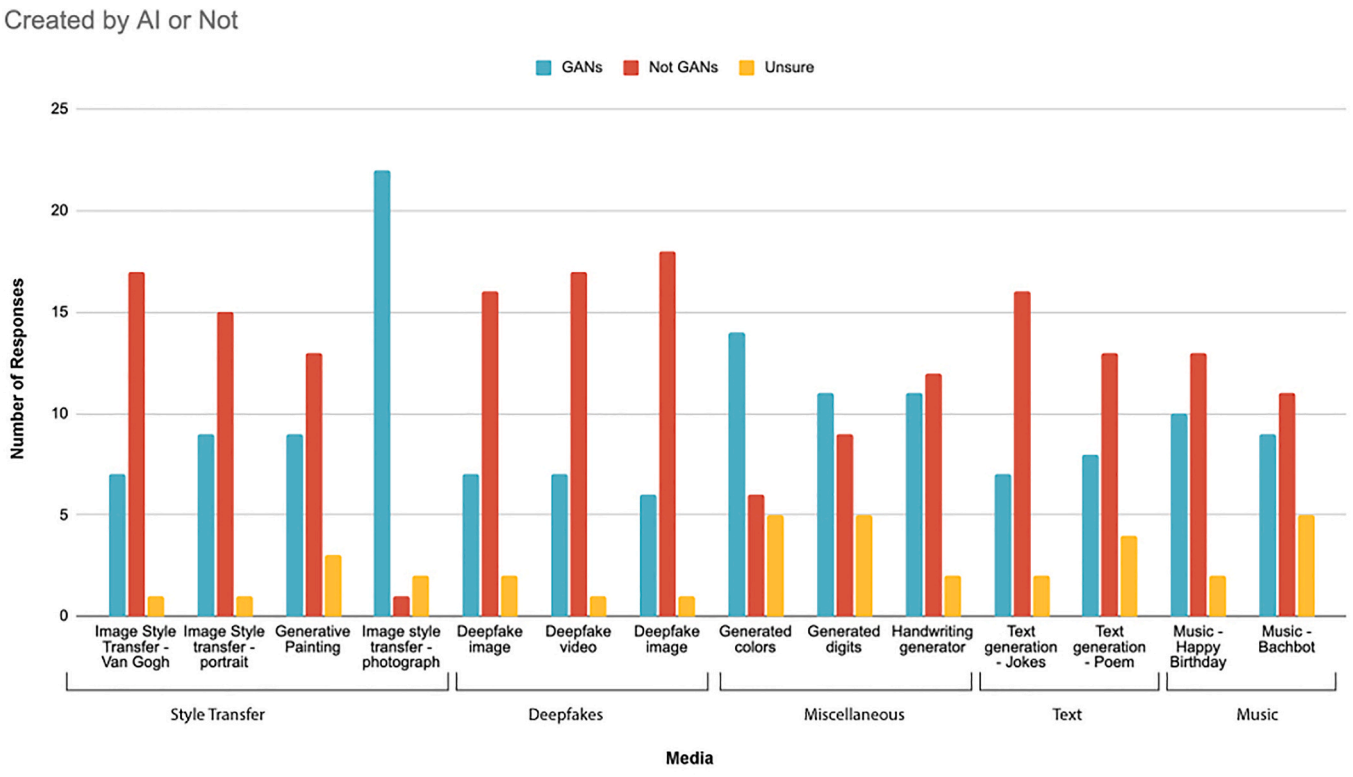


Fig. 4. Students' responses to the *Created by AI or Not* activity indicated that most students could not identify Deepfakes as created by AI but were better at recognizing image style transfer photographs and generated colors. (Ali et al., 2021a) (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

represented in STEM (female and BIPOC youth). Interested students signed up for this extracurricular summer workshop through their teachers who participated in the program. Sixteen students participated in the first workshop, and 22 students participated in the second workshop. Our protocol was IRB approved, and all participants and their parents consented to participating in the study, permitting us to collect their video, audio, and activity participation data.

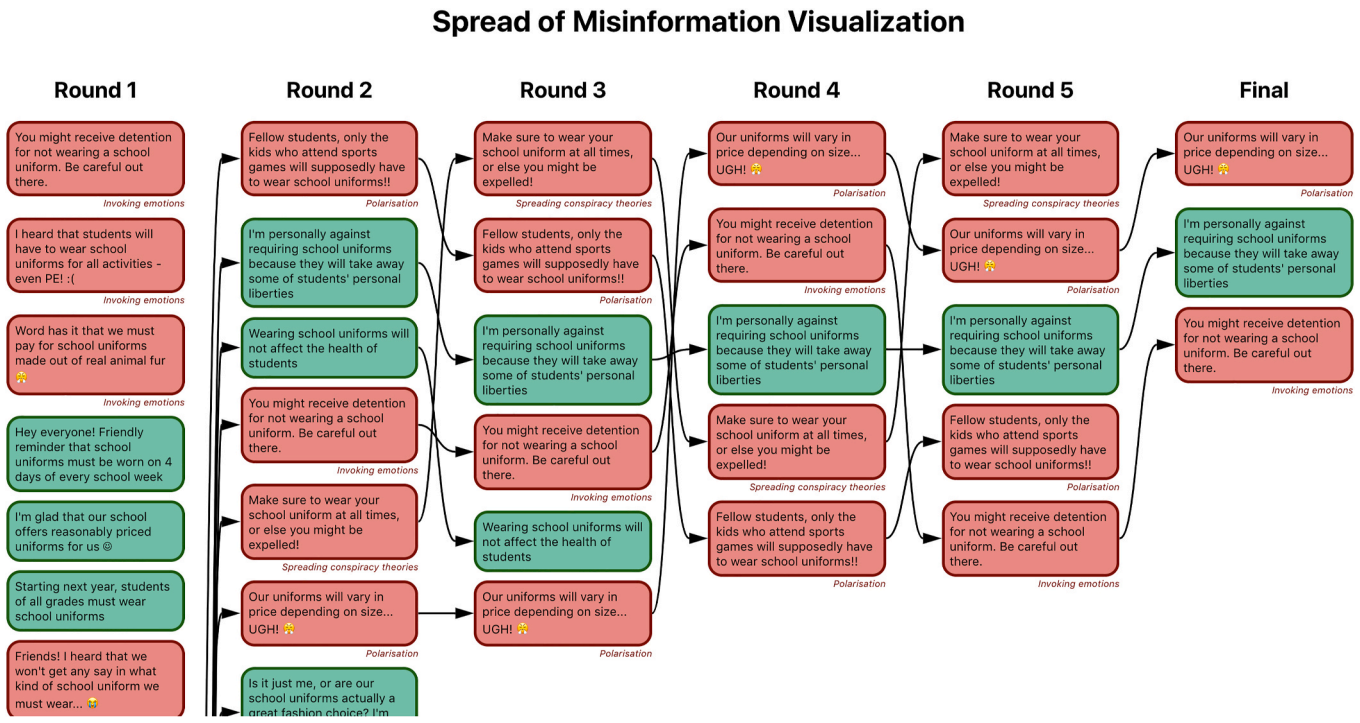


Fig. 5. School-book app message sharing: tree-map visualization of one full round of interaction. Misinformation tended to spread more than neutral information. (Image cropped at bottom).

4.3. Data collection and measures

4.3.1. Generative modeling (created by AI or not)

At the beginning of the workshop, students were asked to identify whether different types of media were generated by AI or Not. A piece of media was shared on the students' individual Google Slides and students marked it as "Created by AI", "Not created by AI", or "Not sure". Responses were aggregated across all students. Media was divided into the following categories: style transfer, Deepfakes, writing generation, text generation, and audio generation. Responses were analyzed by the each media, and each category. After the activity, students discussed which media they were surprised by. Student discussion was recorded through classroom audio and chat. Researchers transcribed the classroom audio and picked out all quotes discussing students' reasoning for their classification of media, and their emotional response after knowing the correct answers.

4.3.2. Spot the deepfakes

Students watched 10 videos and marked each one as a deepfake or not in a Google Form. Students then learned about common tactics to spot Deepfakes and took a second quiz with 10 new videos. Each quiz was scored for percentage correct response in the pre-test and post-test. We first analyzed the scores for normal distribution, and then conducted a paired *t*-test to analyze whether there was a significant improvement in children's identification of Deepfakes after the lesson. During class, students discussed ways to detect Deepfakes, possible societal implications of Deepfakes, and how we can counter these societal implications. Student discussion was recorded through classroom audio and chat. Researchers transcribed all students quotes corresponding to (1) their experience with the activity, and (2) their ideas about the implications of Deepfakes.

4.3.3. Spread of misinformation

In this activity, the messages shared in the School-Book app were saved. A flow diagram for how the messages survived from round to round was generated at the end of the game. Further, we recorded participants' responses to open-ended questions about misinformation through in-class audio recordings, and Padlet file. Researchers transcribed all student responses to the open-ended questions.

4.3.4. Questionnaire

Students completed a pre and post questionnaire to capture their attitudes toward AI and knowledge of generative machine learning. In the post-test, students in the second cohort were given five policies about the existence of Deepfakes and were asked which would be best to handle Deepfakes. These policies were inspired by the current policies around manipulated media at YouTube, Facebook, Twitter, and the Deepfakes Accountability Act introduced in Congress.

5. Results

5.1. Generative modeling

A total of 25 students across both cohorts completed the *Created by AI or Not* survey. We observed that for 11 out of 14 generated media, more students believed the media was real (or created by humans). Media were divided in four categories - style transfer, Deepfakes, text, music and miscellaneous (colors, digits and hand-writing). We observed that students were overall better at recognizing images that used artistic style transfer (46% thought they were real) and miscellaneous media (36%) (see Fig. 4). They could not recognize Deepfakes (68%) and generative text (58%) very well. For the deepfake image shown in Fig. 1 and 72% participants believed it was a photo of a real person.

After the survey, instructors informed students that all media they viewed was generated using AI techniques. Students expressed surprise out loud and in chat. One student wrote,

> "Some of the media looked human-made and others looked AI-made. When [instructor] told us that all the media were AI-made, it was shocking."

5.2. Spot the Deepfakes

A total of 31 students from both cohorts responded to the *Spot the Deepfakes* survey before and after the activity. The maximum score was 100 when they got all correct, and minimum was 0, when they were all incorrect. A *t*-test revealed that on average, students had the same number of correct answers from pre-test (M = 54.68, SD = 15.37) to post-test (M = 53.87, SD = 15.2). This lack of change from pre to post indicates that the activity had no significant effect on children's ability to spot Deepfakes, $t(61) = -0.0294$, $p = .488$. After the activity, multiple students reported in class that it was very difficult to spot the Deepfakes and that they answered randomly. One student said,

> "I found the deepfake exercise really hard, this is because AI videos can be very convincing."

During their in-classroom discussions while looking at examples of Deepfakes, some students could use the techniques they learned to spot Deepfakes. For instance, one student look at the deefape video of Mark Zuckerberg and said,

> "I can tell it is fake because he would not be saying those things about Facebook."

Another student looked at a deepfake image of a woman and said,

> "The earrings look different from each other."

However, there were some images and videos that none of the students could detect manipulations in, and had to be assisted by the instructor.

While discussing potential harms of Deepfakes, students identified impersonation crimes and spread of misinformation as harms. One student said,

> "It can make people say things they didn't say and spread rumors about them."

One student also said "politics" but did not explain their rationale. When asked about what can be done to counter Deepfakes, one student said,

> "Having government control on all AI products"

5.3. Spread of misinformation

5.3.1. School-Book App simulation

All 38 students in both the workshops participated in the school-book activity immediately after learning about Deepfakes. Sixteen students in first workshop played for six rounds and the 22 students in the second workshop played for nine rounds. The game went until three or fewer news items remained. The visualization generated by the first group can be seen in Fig. 5. The headlines in green are neutral (have characteristics of factual information) and convey school rules or clear opinions, while those in red have features of misinformation intended to incite emotions, polarize groups, or spread conspiracy theories. The first round ended up spreading misinformation and neutral messages equally, which already indicates how misinformation is appealing to share. After only 2 more rounds, misinformation was spread 4 times more often than neutral headlines. In the end, of the 3 final headlines, 2 were misinformation and 1 was neutral. In general, headlines that invoked fear of a consequence spread the furthest, whether it was factual or not. Similar patterns of news message spread were observed for the other cohort. We noticed that more students tend to require a higher number of rounds until completion of the activity, since the group starts with more posts.

The School-Book app is intended to simulate how messages about news spreads in the real world. Post activity, students viewed the visualization tree-map of information spread during their rounds. In the instructor-led discussion, when students were asked how they passed information along, one student said,

> "I just shared what I found interesting and wanted other people to know."

None of the students spoke about considering the authenticity of items before sharing them.

Students also shared their thoughts on what information seemed to be shared the most. They saw a difference between the information in red and the information in green, in both the content and the way it travelled,

> "The red ones are more critical than the green ones…. the red ones [got shared more], probably because there were rumors."

5.3.2. Reflection

For the second workshop, in the post-task reflection on Padlet, 75% of the 16 students who participated reported using trusted news publications (such as CNN, Fox News, New York Times) as their source of news and information, 37.5% students reported getting news from their family, 31.25% got news from others their school environment like their friends, and 25% relied on social media (such as TikTok and Snapchat) for news. Students reported that they know that information is true when they can verify it from many sources (53.33%), if it comes from a source that they trust (46.66%), or if they know the author (26.66%). One student reported using personal discretion,

> "I look at other sources to verify and i think about if it makes sense."

Students had a varied sense of what percentage of their newsfeed is true information ($M = 61.18 \pm 17.95\%$). None of the students reported that all the news they consume is real or factual.

To connect the topics of misinformation and Deepfakes, we asked students if they thought that Deepfakes would travel faster in a social network than true videos. Out of the 14 students who responded to this question, 12 said "yes", 1 said "maybe", and 1 believed that we will get better at recognizing Deepfakes. One student explained why they thought it would travel faster,

> "Yes, that is because Deepfakes are sometimes over exaggerated and people find them more interesting than real news."

5.4. Policy decisions

A total of 16 students from the second workshop answered the policy question as a part of their post-test questionnaire where they chose which policy they would support from existing legal policies around Deepfakes (see Table 1). We calculated the percentage of students who responded with each of the presented options. 43.75% students chose the policy stated by the Congress Deepfakes Accountability Act: "All Deepfakes must be labeled if they could potentially be harmful. If they are not, it is considered a crime. Public officials and employees can create Deepfakes to aid in public safety." Students were not aware of the source of the policy. Students' responses to the policy questionnaire are summarized below.

6. Discussion

In this work, we took a three-part approach to teach students about a complex socio-technical system (comprised of algorithms, information and social media) – the evolution of generative modeling techniques, their ethical implications (e.g., Deepfakes), and how they affect the consumption and spread of information via social media. Each part feeds into the next. In the first part, students gain an understanding of a complex technical system, namely GANs, using accessible approaches. In the second part, they expand on this knowledge to become aware of a potentially harmful application of the said technical system (i.e.,

Table 1
Students' responses to the policy questionnaire.

Policy	Source	Percent of students	Example comments
All Deepfakes must be labeled if they could potentially be harmful. If they are not, it is considered a crime. Public officials and employees can create Deepfakes to aid in public safety	Congress - Deepfakes Accountability Act	43.75%	"Anything that is harmful to anyone, or frames anyone for something they did not say or do should be taken down, or flagged fake immediately because it can really affect a person's life."
All manipulated media (including Deepfakes) will be noted as being false. People will be notified if the information they share is marked as manipulated, and they will be able to read more about why it was detected as such	Twitter	37.5%	"I see it as too authoritarian to completely ban them, but marking them as manipulated is reasonable."
All Deepfakes must be deleted, unless used for entertainment. Other fake manipulated media (that doesn't use AI technology) can stay on the web, though it will be tagged as false	Facebook	18.75%	"I believe it can stay on there for entertainment, but it can be harmful."
All political Deepfakes are banned, but all other Deepfakes can stay	YouTube	0	
All people should learn more about Deepfakes, but they are still allowed on the Internet	No policy, digital media literacy	0	

Deepfakes) then critically reflect on the implications of this technology. The third part involves students learning about misinformation and how it spreads. Then, students apply their knowledge of Deepfakes to reckon how these realistic manipulated media can fuel the spread of misinformation. Finally, students take the role of policy-makers and provide their opinions on regulatory policies for Deepfakes on social media.

6.1. Introducing complex technical concepts to novice learners

Our goal was to introduce students to fake news and the implications that AI-generated Deepfakes have for the spread of fake news. A lack of access to technical pre-requisites and computational resources often make CS and AI education less accessible to k12 students. Previous work has found how socio-technical systems such as Deepfakes often involve technical concepts that children are not familiar with Ali et al. (2021a, 2021b). While generative AI might be a novel concept for many middle and high school students, their applications are certainly not. Media generated by popular generative modeling applications such as Prisma (Prisma) have been widely used by youth and shared widely on social media. In order to understand the social and political implications of these systems, students must be able to anchor their opinions in an understanding of technical capabilities.

To accomplish this goal, we gave students an understanding of Deepfakes and the generative AI used to create them. Historically, to understand the technical composition of GANs learners need experience with machine learning and high level mathematical concepts, along with expensive computing resources. In our work, we make use of accessible media examples, interactive activities and games to learn about generative AI. In the *Created by AI or Not* activity, we introduce the concept of generative AI to students through examples of media that they are already familiar with, thereby building upon existing knowledge. Our lessons include both positive and potentially harmful applications to give students the full breadth of what is possible.

In the activity, we observed that students believed that several AI-created media examples, including Deepfakes, were real or factual. When students were told that the media was generated by AI, they expressed surprise which led to a poignant revelation about potentially negative applications of generative AI. The *Created by AI or Not* activity facilitates learning by eliciting an emotional response from students through cognitive dissonance, or challenging their assumptions of what is *real*. One student wrote,

> "The AI or not activity was hard because I didn't have a single clue about what was human or AI made. I (now) know AI can be really good at generating things."

Through showing them what AI was capable of creating, they learned to both think critically about media they encounter, and not assume that they are real even when they seem so. It also piqued their interest in learning about how these realistic media are generated. This introduction to generative AI enabled us to subsequently discuss Deepfakes with students, since they now understood what generative media meant. We also had students play a role-playing game to understand the underlying algorithm that makes up a GAN (Ali et al., 2021a, 2021b), which added to their technical understanding. This technical knowledge of how media can be manipulated arms students with the ability to be skeptic of believing the authenticity of digital media, in turn, making them responsible consumers and sharers of information. In this work we demonstrated how they could gain this conceptual understanding without having access to the mathematical or computational pre-requisite knowledge.

6.2. Critical thinking about implications

Advances in generative machine learning have made it easier to create and share Deepfakes on social networks, which have a heavy influence on socio-political discourse through disinformation (Chesney & Citron, 2019). Children are especially vulnerable to convincing misinformation since they form social and political opinions in their formative years and can become targets of convincing Deepfakes (Torney-Purta, 2017). Therefore, it is imperative for children to think about Deepfakes in the context of complex social networks.

Firstly, in the *Spot the Deepfakes* activity, students found it challenging to discern factual versus fake media. This is in line with existing research that shows that it is difficult for humans to perceive Deepfakes (Korshunov & Marcel, 2020). This gave students first-hand experiences of how Deepfakes can convincingly deceive users. Secondly, the *School-Book* activity allowed students to contribute to and witness the spread of misinformation. The visualization at the end of the activity was a key part of demonstrating how information spreads. Students noticed that misinformation spreads faster and farther than true information, and even compared the activity to the ways rumors spread in school. Finally, students bridged together their knowledge of Deepfakes and the spread of information. After the *School-Book* activity, when asked how Deepfakes will proliferate online, 85.7% students responded that Deepfakes will become more widespread than true videos. One student said,

> "There are many people who either want to get views or trick people, or they might mishear something. many websites like that are getting more and more popular."

In this activity, we made use of a previously acquired technical knowledge, and encouraged students to expand the learned technical concept to think about its ethical implications. While we focus on Deepfakes and misinformation, this approach of reflecting on implications also motivates students to expand this practice to other technical concepts.

6.3. Empowerment and agency: taking action

Socio-technical systems, like Deepfakes, require input from multiple stakeholders to encourage its responsible use (Costanza-Chock, 2020). These stakeholders not only include the designers and engineers of the systems, but policy makers and users. It was important to have students understand that they, as current users and future developers and policy makers, can make an impact on how Deepfakes are implemented. At the end of the workshop, students applied their understanding of generated media and its implications to real-world policies. Students deliberated on policies similar to those recently put forth by the government and tech companies.

All students selected a policy that took some action to regulate Deepfakes, signaling that they all thought this application of technology needs oversight. This may be due to the fact that all students had a difficult time detecting which media was fake in the *Spot the Deepfakes* activity and were vocal about these frustrations in class. In the larger context of GAN development, the technology is getting more sophisticated and harder to discern from true media. The strategies that students learned in class will likely be irrelevant in the coming years, which means it will be important to deploy different strategies to combat the spread of misinformation.

Over 75% of students chose policies that identified and labeled Deepfakes, but did not ban them (Deepfakes Accountability Act and Twitter). Students did not believe that Deepfakes should be banned fully, and understood that there is some creative merit to creating and sharing Deepfakes. They advocated for knowing whether something was a deepfake or not instead of having to decide on their own. No students chose the policy inspired by Digital Media Literacy, showing that students believed that they could not be taught how to spot manipulated media.

As students continue to navigate the digital world, it is important that they are not only are aware of new types of technologies and their implications, but that they understand how harmful outcomes of the technology can be mitigated. In the case of our workshop, students had a chance to explore policy-driven solutions. As future makers of AI technologies and citizens in an increasingly digital society, being able to apply technical knowledge to policy-making will be an important skill.

7. Conclusion

In this work, we outline an approach to introduce students to Deepfakes which includes understanding the technical systems that create Deepfakes, their implications on spreading misinformation, and potential policies to mitigate their potential harm. We confirm that middle school students could successfully conceptualize what Deepfakes are, what misinformation is and how it spreads, and had opinions about policy surrounding Deepfakes - making these appropriate learning goals for this age group. This work adds to existing literature on digital media literacy by equipping students with the knowledge to think critically about what they see online. However, to our knowledge, this is the only work that introduces students to Deepfakes within a media literacy context. Due to the low barrier to create high quality Deepfakes, this is an especially timely topic with deep societal implications. This is an especially critical curriculum for this age group because of the presence of such media on social media apps frequented by middle school students. This work also contributes to the growing literature around advances in middle school AI literacy (Lee, 2020), specifically ethical AI literacy (Ali et al., 2019; An Ethics of Artificial Intelligence Curriculum for Middle School Students) and creative AI literacy (Ali et al., 2021a, 2021b). We propose the concepts of generative manipulated media that have potential harmful implications to be included in digital media literacy curricula. We also discuss different alternatives to just media literacy, such as considering stakeholders outside of creators and consumers of technology, specifically, government or corporate policies. In our work, we observed that even though students gained an understanding of how Deepfakes are created and how to spot them, they still expressed the need to have policy regulating the presence of Deepfakes on social media. Through these activities, they understood the role of various stakeholders in this socio-technical system: creators of technology, consumers of media and policy makers. These activities also add to work in K-12 AI and CS literacy, specifically around the ethical implications of generative machine learning. We hope that this work inspires educators and curriculum designers to consider both a technical and political approach to digital media literacy. This literacy approach towards socio-technical concepts can be expanded to other technical topics. Finally, while Deepfakes will continue to exist, and get better, children are now better prepared to challenge their authenticity and hinder their spread as misinformation when they encounter them online.

7.1. Limitations and future work

Information, and as a result, misinformation is grounded in cultural and political contexts. While this work is situated in the context of the United States of America (USA), it would be beneficial for researchers and educators in other countries to adapt these lessons to their socio-political contexts. For instance, we consider policies from US congressional bodies and major technology corporations based in the USA. These might not be relevant in other countries. We used a social media simulation to make students witness how misinformation spreads wider than true information. Future work could also discuss how social media algorithms lead to positive feedback loops, making popular content and people even more visible. There is also an opportunity to discuss filter bubbles, where social media algorithms isolate opinions that feed into each other and can amplify false information.

We situate this work as a digital media literacy lesson, but it can be used as part of K-12 AI literacy curricula to discuss the ethical implications of AI. Digital literacy is a complex topic to teach and learn, and needs to incorporate different subject areas, such as sociology, politics, and technology. While we introduce a novel approach to discuss one technical application (Deepfakes created by generative AI) that can influence one kind of digital media, future work could tie in other relevant topics, such as social-network bots, filter bubble and algorithmic bias. Existing media literacy approaches discuss media bias and trustworthy sources, and there is a need to include awareness about how evolving technology influences the spread of information - which this work aims to do. In the future, we hope this work inspires symbiotic opportunities to discuss media literacy in computer science curricula and technical literacy in digital media lessons for K-12 classrooms.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Cc Song, Helen Zhang and Yihong Cheng for their teaching assistance and feedback on the activity. Thank you to our participating students and teachers for all of your help in organizing the workshop and giving us feedback on the activity. Lastly, thank you to National Science Foundation under grant #2022502/#2048746 and the Amazon Future Engineers program for funding this work.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting world leaders against deep fakes. In *CVPR workshops* (pp. 38–45).

An Ethics of Artificial Intelligence Curriculum for Middle School Students. https://thecenter.mit.edu/wp-content/uploads/2020/07/MIT-AI-Ethics-Education-Curriculum.pdf, (2020)–. (Accessed 23 November 2021).

Ali, et al. (2019). Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *International Workshop on Education in Artificial Intelligence K-12 (EDUAI'19)* (pp. 1–4).

Ali, S., DiPaola, D., Lee, I., Hong, J., & Breazeal, C. (2021a). Exploring Generative Models with Middle School Students. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Ali, S., DiPaola, D., & Breazeal, C. (2021b). What are GANs?: Introducing Generative Adversarial Networks to Middle School Students. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35* pp. 15472–15479). No. 17.

Ali, S., Park, H. W., & Breazeal, C. (2020). Can children emulate a robotic non-player character's figural creativity?. In *Proceedings of the annual symposium on computer-human interaction in play* (pp. 499–509).

Johansen, Alison Grace (2020). *How to spot deepfake videos — 15 signs to watch for*. https://us.norton.com/internetsecurity-emerging-threats-how-to-spot-deepfakes.html.

Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering, 28*(2), 525–552, 2021.

Asnani, V., Yin, X., Hassner, T., & Liu, X. (2021). *Reverse engineering of generative models: Inferring model hyperparameters from generated images*. arXiv:2106.07873 [cs.CV].

Bernard, M. (2019). The best (and scariest) examples of AI-enabled deepfakes, 2019, Retrieved from https://www.forbes.com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples- of-aienabled-deepfakes.

Bickert, M. (2020). *Enforcing against manipulated media*. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

Blakeslee, S. (2004). The CRAAP test. *LOEX Quarterly, 31*(3), 4, 2004.

Bloomberg. (2018). *How faking videos became easy and why that's so scary*. https://fortune.com/2018/09/11/deepfakes-obama-video/.

Buckingham, D. (2016). Defining digital literacy. *Nordic Journal of Digital Literacy*, 21–34, 2016.

Cao, J., Hu, Y., Yu, B., He, R., & Sun, Z. (2019). 3D aided duet GANs for multi-view face image synthesis. *IEEE Transactions on Information Forensics and Security, 14*(8), 2028–2042, 2019.

Chen, X., Zou, Di, Cheng, G., & Xie, H. (2020a). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education. *Computers & Education, 151* (2020), 103855.

Chen, X., Zou, Di, & Xie, H. (2020b). Fifty years of British journal of educational technology: A topic modeling based bibliometric perspective. *British Journal of Educational Technology, 51*(3), 692–708, 2020.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs, 98*, 147, 2019.

Coble, S. (2020). *YouTube issues deepfake ban reminder*. https://www.infosecurity-magazine.com/news/youtube-issues-deepfake-ban/.

Coldewey, D. (2019). *DEEPFAKES Accountability Act would impose unenforceable rules - but it's a start*. https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

De Vries, K. (2020). You never fake alone. Creative AI in action. *Information, Communication & Society, 23*(14), 2110–2127, 2020.

Denise, E. A., & Abbas, J. (2016). Simple tips for helping students become safer, smarter social media users. *Knowledge Quest, 44*(4), 42–47, 2016.

Dickson, E. J. (2020). *TikTok stars are being turned into deepfake porn without their consent*. https://www.rollingstone.com/culture/culture-features/tiktok-creators-deepfake-pornography-discord-pornhub-1078859/.

Doersch, C. (2016). *Tutorial on variational autoencoders*. arXiv preprint arXiv:1606.05908, 2016.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly, 80*(S1), 298–320, 2016.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414–2423).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

Gunther, R., Beck, P. A., & Nisbet, E. C. (2018). *Fake news did have a significant impact on the vote in the 2016 election: Original full-length version with methodological appendix*. Unpublished manuscript. Columbus, OH: Ohio State University, 2018.

Haag, M., & Salam, M. (2017). Gunman in pizzagate shooting is sentenced to 4 years in prison. *New York Times, 23*, 2017.

Jan, K., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146, 2020.

Joel Breakstone, McGrew, S., Smith, M., Ortega, T., & Sam, W. (2018). Why we need a new approach to teaching digital literacy. *Phi Delta Kappan, 99*(6), 27–32, 2018.

Joseph, K., Lee, N.-J., & Feezell, J. T. (2012). Digital media literacy education and online civic and political participation. *International Journal of Communication, 6*, 24, 2012.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). DeepFakE: Improving fake news detection using tensor decomposition-based deep neural network. *The Journal of Supercomputing, 77*(2), 1015–1037, 2021.

Kaya, Y. (2019). *Instagram head says company is evaluating how to handle deepfakes*. https://www.cnn.com/2019/06/25/tech/instagram-deepfakes/index.html.

Korshunov, P., & Marcel, S. (2018). *Deepfakes: A new threat to face recognition? Assessment and detection*. arXiv preprint arXiv:1812.08685 (2018).

Korshunov, P., & Marcel, S. (2020). *Deepfake detection: Humans vs. machines*. arXiv preprint arXiv:2009.03155, 2020.

Lee, N. M. (2018). Fake news, phishing, and fraud: A call for research on digital media literacy education beyond the classroom. *Communication Education, 67*(4), 460–466, 2018.

Lee. (2020). Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, 52* pp. 191–197). https://doi.org/10.1145/3408877.3432513

Leu, D. J., Reinking, D., Carter, A., Castek, J., Coiro, J., Henry, L. A., & Zawilinski, L. (2007). Defining online reading comprehension: Using think aloud verbal protocols to refine a preliminary model of Internet reading comprehension processes. In *Annual meeting of the*. Chicago, IL: American Educational Research Association.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353–369, 2017.

Li, Y., Chang, M.-C., & Lyu, S. (2018), 2018. In *Ictu oculi: Exposing ai generated fake face videos by detecting eye blinking*. arXiv preprint arXiv:1806.02877.

Literat, I., Chang, Y. K., & Hsu, S.-Y. (2020). Gamifying fake news: Engaging youth in the participatory design of news literacy games. *Convergence, 26*(3), 503–516, 2020.

Mechanic, D., & Tanner, J. (2007). Vulnerable people, groups, and populations: Societal view. *Health Affairs, 26*(5), 1220–1230, 2007.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). *Deep learning for deepfakes creation and detection: A survey*. arXiv preprint arXiv:1909, 2019.

Nowell, S. D. (2014). Using disruptive technologies to make digital connections: Stories of media use and digital literacy in secondary classrooms. *Educational Media International, 51*(2), 109–123, 2014.

Paul, M., & Benjamin, T. (2013). Media literacy as a core competency for engaged citizenship in participatory democracy. *American Behavioral Scientist, 57*(11), 1611–1622, 2013.

Perez, S. (2019). *Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media*. https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/.

Phartiyal, S., Patnaik, S., & Ingram, D. (2018). When a text can trigger a lynching: WhatsApp struggles with incendiary messages in India. *Reuters UK, 25*, 2018.

Pilgrim, J., Vasinda, S., Bledsoe, C., & Martinez, E. (2019). Critical thinking is critical: Octopuses, online sources, and reliability reasoning. *The Reading Teacher, 73*(1), 85–93, 2019.

Polizzi, G., & Taylor, R. (2019). *Misinformation, digital literacy and the school curriculum*, 2019.

Prisma [n.d.]. Prisma. prisma-ai.com.

Roozenbeek, J., & Sander van der Linden. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(1), 1–10, 2019.

Starbird, K., Jim Maddock, Orand, M., Achterman, P., & Mason, R. M. (2014). In *Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. IConference 2014 Proceedings*, 2014.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148, 2020.

Torney-Purta, J. V. (2017). *The development of political attitudes in children*. Routledge.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, 2056305120903408 *Social Media+ Society, 6*(1), 2020.

Van der Linden, S., Anthony, L., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), 1600008, 2017.

Vera, S., Barash, Y., Konen, E., & Klang, E. (2020). Creating artificial images for radiology applications using generative adversarial networks (GANs)–a systematic review. *Academic Radiology, 27*(8), 1175–1185, 2020.

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

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe report. *27* pp. 1–107), 2017.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1947–1962, 2018.