# Design Insights for Developing Deepfakes Awareness Games for Adolescents

**Clara Maathuis, Samuel Derous and Ebrahim Rahimi**
Open University of the Netherlands, Heerlen, Netherlands.

clara.maathuis@ou.nl
samuel.derous@student.ou.nl
ebrahim.rahimi@ou.nl

**Abstract:** Around a decade ago, significant advancements in the deep learning domain like Generative Adversarial Networks (GANs) started to facilitate the creation of new synthetical content. The development of deepfake algorithms gained momentum due to the increase in computing power and open-source available resources which facilitated broader access to advanced deep learning methods. While initially deepfakes were created by creative pursuits in various media and marketing campaigns, they started to be misused driven by malicious purposes in various social media manipulation campaigns. These campaigns imply spreading disinformation and misinformation for discrediting individuals and creating confusion in relation to specific topics and events and are meant to altering users' beliefs and behaviours. Such facts raised important ethical and societal concerns. Although deepfakes still represent a recent research area, a large body of studies is dedicated to, on the one hand, generation, and detection of deepfakes, and on the other hand, to understanding their implications and consequences for society. At the same time, governmental and practitioner efforts are devoted to containing their use and impact on society by proposing various strategies and programs. Nevertheless, both academic and societal efforts are in an incipient stage and often adopt a generalist perspective. This represents a crucial point in relation to the fact that many users in various social media platforms like TikTok is represented by adolescents that are known to be unaware and more vulnerable to media content in general, and deepfakes content, in particular. This represents the knowledge gap that this research aims to tackle by synthesizing design insights as requirements and guidelines that can be used when developing and deploying deepfakes awareness games for producing and/or enhancing awareness of adolescents. Accordingly, a systematic literature review is conducted and merged with previous experience in this domain by adopting a transdisciplinary research approach that merges methods and techniques from the deepfakes, AI, social media, cyber security, and gamification domains.

**Keywords:** Deepfakes, Deepfakes Awareness, Social Media, Artificial Intelligence, Gamification.

## 1. Introduction

"The truth is rarely pure and never simple." (Oscar Wilde)

The advancements in the DL domain facilitated the creation of deepfakes, synthetic media content that convincingly replicates human actions and behaviors. While initially deepfakes were relatively simplistic, they evolved into sophisticated mechanisms that can generate highly realistic and nuanced synthetic media that became harder to distinguish from genuine content to the untrained eye (Shoaib et al., 2023). As believable media artificially generated, deepfakes aim to react or replace dimensions of reality focusing either on one or more aspects of individuals, e.g., voice alteration or mouth, gaze, and pose reenactment (Mirsky & Lee, 2021). This represents a significant threat to adolescents as they are frequent users of various social media platforms such as TikTok and Instagram, by undermining their ability to discern between real and manipulated content (Kilhoffer et al., 2023; Naffi et al., 2023; Maathuis et al., 2023), making them further exposed to threats like cyber bullying and online harassment, which have the potential to result in serious well-being and safety negative consequences. Hence, effective solutions and interventions that deal with the deepfakes threat by equipping adolescents with the necessary digital and human skills start with awareness as this implies and calls for understanding deepfakes, their possible consequences, and being able to further detect them and build resilience against their action (Taibi et al., 2023). Such initiatives play a crucial role in empowering adolescents to become responsible digital citizens in an increasingly complex, interconnected, and uncertain online world. Nevertheless, given these characteristics of cyberspace, an overall deepfakes awareness solution is impossible to build, as would require to be able to account and embed the needs, features, and skills of all possible target audiences. Hence, tailored solutions should be developed for adolescents so that their unique cognitive, emotional, and developmental characteristics are effectively addressed in corresponding deepfakes gaming awareness solutions (Ali et al., 2021). This would also maximize the impact on enhancing digital resilience and fostering a safe online environment for young users. On this behalf, a series of research, practitioner, and governmental efforts start to be developed, nevertheless, this requires a systematic consideration of the design choices and their underlying requirements and principles. To tackle this pressing demand and support the

development of tailored digital educational solutions and interventions, this research aims to capture design requirements and guidelines for building gaming solutions for deepfakes awareness of adolescents. To this end, the following research questions are formulated:

- **RQ 1:** Which deepfakes awareness solutions exist and are addressed to adolescents?
- **RQ 2:** What are the design principles and requirements useful for building deepfakes awareness gaming solutions for adolescents?

To attain this goal, this study proposes a comprehensive set of eighteen design requirements and guidelines for the development of deepfake awareness gaming solutions for adolescents. These insights are distilled from a systematic literature review that integrates perspectives from various disciplinary domains, encompassing artificial intelligence, social media studies, cybersecurity, and gamification, plus drawing upon prior expertise in the field.

The remainder of the article is structured as follows. Section 2 discusses relevant studies that tackle various dimensions of the context of this research. Section 3 presents the methodological approach considered in this article. Section 4 elaborates on characteristics identified on existing deepfakes-related awareness studies. Section 5 captures design requirements and guidelines to be considered when building gaming awareness solutions for adolescents. At the end, the last section discusses concluding remarks and future research perspectives.

## 2. Research Background

Social media platforms became integral to the lives of adolescents, particularly in recent years, with a significant surge observed during the COVID-19 lockdown. Platforms like YouTube, TikTok, and Instagram serve as vital channels for social interaction, entertainment, and self-expression among adolescents. These platforms employ advanced DL-based algorithms to personalize content delivery, aiming to enhance user experience by presenting material tailored to individual preferences. Nevertheless, while these algorithms are effective in catering to user interests, they also pose various risks, including the formation of echo chambers that reinforce unrealistic and biased perceptions of reality while favouring building a positive attitude and enhancing trust in AI algorithms used, often without fully comprehending their functions or potential biases (Fard & Maathuis, 2021; Theophilou et al., 2023). This lack of awareness leaves them vulnerable to accepting algorithm-generated recommendations uncritically, which may perpetuate systemic biases and hinder the development of critical thinking skills (Ask et al., 2023). Addressing these issues requires interventions to challenge adolescents' blind trust in algorithms, raise awareness of potential consequences, and equip them with strategies to navigate social media platforms responsibly.

As deepfakes' complexity advances, it becomes increasingly challenging to detect them due to the intricate combination of diverse features and mechanisms used in their creation process. Deepfakes typically involve the synthesis of facial expressions, gestures, voice patterns, and even contextual backgrounds to create convincing replicas of individuals in video or audio recordings (Naitali et al., 2023; Masood et al., 2023). Deepfakes rely on deep learning techniques like a) GAN that consist of two neural networks (generator and discriminator) engaged in a competitive process where the generator generates fake content while the discriminator evaluates them to distinguish between real and fake, and b) AU that encode an input image into a latent representation and then decodes it to reconstruct the original image (Ali et al., 2021). Moreover, the proliferation of large datasets and computing power fuels the refinement of deepfake algorithms, enabling them to produce highly realistic and nuanced simulations (Ali et al., 2021; Mirsky & Lee, 2021). Detecting such complex deepfakes necessitates the integration of diverse detection methods, including forensic analysis, anomaly detection, and AI-based algorithms trained to identify subtle inconsistencies. As the literature study conducted in this research shows, this represents a challenging goal that needs to be tackled in a systematic and multidisciplinary manner in an educational setting while adopting a human-centred approach that accounts the needs, goals, and skills of adolescents when building effective and responsible deepfakes awareness gaming solutions for them.

**Table 1: Abbreviations**

| Abbreviation | Definition |
|---|---|
| AI | Artificial Intelligence |
| AE | Autoencoders |
| BIPOC | Black, Indigenous, and People of Colour |
| DL | Deep Learning |
| GAN | Generative Adversarial Networks |
| GBL | Game Based Learning |
| HLT | Hypothetical Learning Trajectory |
| HCI | Human Computer Interaction |
| ML | Machine Learning |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| RQ | Research Question |

## 3. Research Methodology

As this research aims to capture design insights for building deepfakes awareness gaming solutions for adolescents, the following research questions are formulated:

- **RQ 1:** Which deepfakes awareness solutions exist and are addressed to adolescents?
- **RQ 2:** What are the design principles and requirements useful for building deepfakes awareness gaming solutions for adolescents?

To achieve this aim, a systematic and transparent approach is taken as outlined by Denyer & Tranfield (2009) which encompasses several key stages, including the establishment of clear objectives, the identification of relevant studies, the meticulous selection of pertinent literature, thorough analysis and synthesis of findings, and the comprehensive presentation of results. Adhering to the guidelines set forth by the PRISMA methodology, data collection is conducted using the PRISMA 2020's stage flow diagram for meticulous data collection, screening, and analysis (Page et al., 2021).

In the context of building deepfakes awareness efforts for adolescents, systematic literature reviews serve as essential tools for consolidating and summarizing relevant evidence accurately and reliably. This allows gathering empirical evidence that meets established criteria, enabling researchers to address key questions effectively and make informed decisions and draw meaningful conclusions about deepfake threats. The efficacy of the PRISMA methodology was demonstrated across diverse domains, underscoring its value in informing educational efforts about the deepfakes threat among adolescents and empowering them to navigate digital environments safely.

Accordingly, the core stages taken in this research are further discussed and depicted in Figure 1:

- **Identification**: To find the initial set of studies written in English, search strings were formed based on *deep fakes, deepfakes, awareness, training, security, prevent, prevention, detection, detection, game, gaming, social media, and TikTok* keywords. The search strings were used to query the following seven scientific databases: ACM, IEEE Digital Library, Google Scholar (first 10 pages), SpringerLink, Taylor & Francis, Web of Science, and Wiley. The timeline considered was 01.09.2016 to 31.12.2023. The initial search provided 481 studies (conference and journal articles), out of which 2 duplicates were removed.
- **Screening:** In this phase, the studies were screened based on their Abstract, title, and keywords used and excluding criteria like being irrelevance to the topic or not being addressed to deepfakes. This resulted in the final set of 22 studies to be considered for review.
- **Included:** In this phase, the 22 studies were fully read and analysed, and their scope is captured in Table 2.

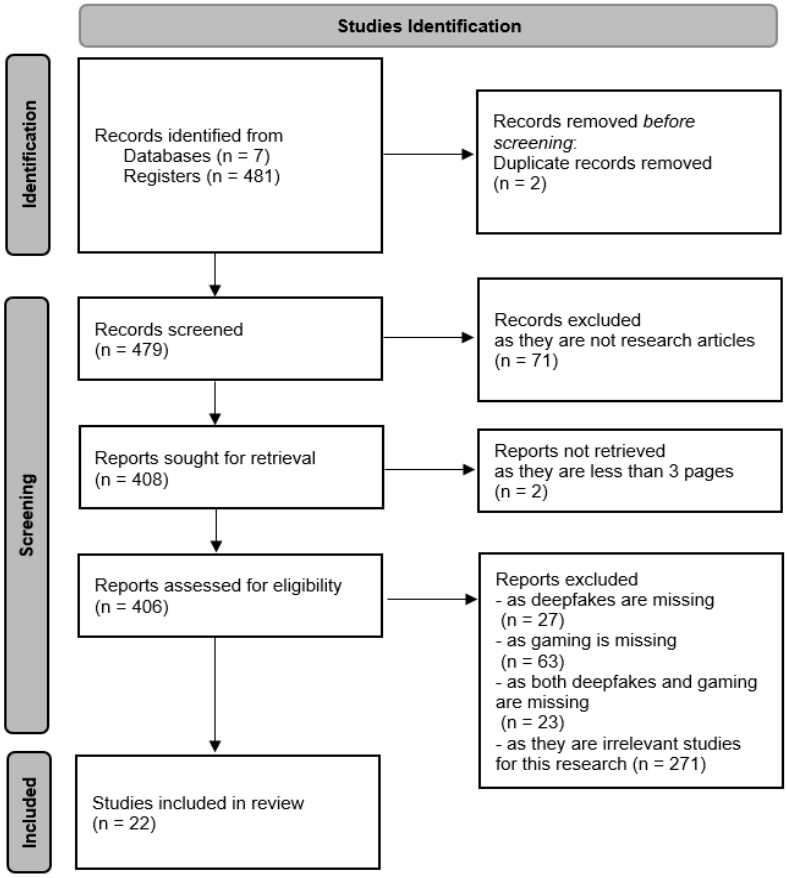


**Figure 1: Review process**

**Table 2: Specifications of the articles selected**

| No | Title Study | Year | Research Aim | Target Audience | Artefact Type | Reference |
|---|---|---|---|---|---|---|
| 1 | "How technical do you get? I'm an English teacher": Teaching and Learning Cybersecurity and AI Ethics in High School | 2023 | To understand how AI ethics and cyber security topics are taught in high school. | K-12 students and teachers | Interview insights | Kilhoffer et al., (2023) |
| 2 | The Creation and Detection of Deepfakes: A Survey | 2021 | To explore the creation and detection of deepfakes. | Review article | Reflection | Mirsky & Lee (2021) |
| 3 | What to do about deepfakes | 2021 | To understand how to minimize or prevent the negative impact of deepfakes. | Position article | Reflection | Johnson & Diakopoulos (2021) |
| 4 | Exploring Generative Models with Middle School Students | 2021 | To design learning activities for non-technical students for understanding the technical composition, applications, and ethical issues of generative models. | K-12 students | Interview insights | Ali et al., (2021) |
| 5 | How Do Users Perceive Deepfake Personas? Investigating the Deepfake User Perception and Its Implications for Human-Computer Interaction | 2023 | To understand human interactions between humans and deepfakes in the HCI context. | General | Experiment Insights based on content created using Synthesia tool | Kaate et al., (2023) |
| 6 | Cybersecurity Education in the Age of AI: Integrating AI Learning into Cybersecurity High School Curricula | 2023 | To design an exploratory curricular initiative that introduces students AI/ML topics in the context of cyber security. | K-12 students | Experiment Insights based on applications created in NetsBlox programming language | Grover, Broll & Babb (2023) |

| No | Title Study | Year | Research Aim | Target Audience | Artefact Type | Reference |
|---|---|---|---|---|---|---|
| 7 | Understanding Digital-Safety Experiences of Youth in the U.S | 2023 | To understand the variety of digital threats that youth experiences. | K-12 students and their parents | Interview and focus groups insights | Freed et al., (2023) |
| 8 | Research on Deepfake Technology and Its Application | 2023 | To analyse security risks of deep forgery for creating public awareness. | Review article | Reflection | Bu, Jiang & Zheng (2023) |
| 9 | It's About Time: Attending to Temporality in Misinformation Interventions | 2023 | To build educational methods for enhancing digital literacy in respect to misinformation engaging BIPOC and rural communities. | K-16 students | Workshop insights | Wilmer et al., (2023) |
| 10 | Developing Middle School Students' AI Literacy | 2021 | To design a workshop on developing AI literacy for middle school students. | K-12 students | Curricula | Lee et al., (2021) |
| 11 | The use of an escape room as an immersive learning environment for building resilience to misinformation | 2023 | To design an immersive environment escape room game for building resilience to misinformation using mechanisms like deepfakes. | General | Immersive learning game | Cho et al., (2023) |
| 12 | Empowering Youth to Combat Malicious Deepfakes and Disinformation: An Experiential and Reflective Learning Experience Informed by Personal Construct Theory | 2023 | To design an educational method that produces awareness to young individuals to deepfakes and disinformation. | University students | Experiment insights using FaceApp, Reface, Impressions, and TikTok | Naffi et al., (2023) |
| 13 | Deepfake detection using deep learning methods: A systematic and comprehensive review | 2023 | To provide an overview of deep learning-based deepfakes detection solutions. | Review article | Reflection | Heidari et al., (2023) |
| 14 | Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward | 2023 | To provide an overview of existing challenges depicted from of state-of-the-art deepfakes generation and detection studies and corresponding solutions. | Review article | Reflection | Masood et al., (2023) |
| 15 | A literature review and perspectives in deepfakes: generation, detection, and applications | 2022 | To provide an overview of deepfakes generation and detection applications. | Review article | Reflection | Dagar & Vishwakarma (2022) |
| 16 | The profile: unleashing your deepfake self | 2023 | To design a game that generates ego-based deepfakes to produce awareness and reflection on relevant ethical dilemmas. | General | Multimedia software application | Cheres & Groza (2023) |
| 17 | AI Ethics Curricula for Middle School Youth: Lessons Learned from Three Project-Based Curricula | 2023 | To design a curriculum that builds awareness on AI ethics topics to middle school students. | K-12 students | Curricula | Williams et al., (2023) |
| 18 | Integrating Ethics and Career Futures with Technical Learning to Promote AI Literacy for Middle School Students: An Exploratory Study | 2023 | To design an AI literacy workshop that covers both ethical issues and their potential implications. | K-12 students | Workshop curricula | Zhang et al., (2023) |
| 19 | Is This Fake or Credible? A Virtual Learning Companion Supporting the Judgment of Young Learners Facing Social Media Content | 2021 | To design a web-based learning environment for creating awareness on social media threats. | K-12 students | Game environment | Aprin et al., (2021) |
| 20 | The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media. | 2021 | To reflect on existing dilemma points in respect to deepfakes detection solutions. | General | Workshop insights | Leibowicz, McGregor & Ovadya (2021) |
| 21 | Children as creators, thinkers, and citizens in an AI-driven future | 2021 | To design a social media game for deepfakes detection and analysis of misinformation effects. | K-12 students | Workshop curricula | Ali et al., (2021) |
| 22 | Fakey: A game intervention to improve news literacy on social media | 2021 | To design a game that enhances news literacy and supports countering social media misinformation. | General | Digital game | Micallef et al., (2021) |

## 4. Awareness Solutions

Based on the systematic literature review conducted, it is evident that research efforts have been directed towards several key audiences, as illustrated in Figure 2. Studies have notably focused on K-12 students, K-16 students, and university students recognizing the importance of educating younger generations about importance of building digital literacy skills together with awareness and resilience against the deepfakes threats and its implications. Additionally, studies have extended to encompass K-12 students alongside their teachers or parents, acknowledging the crucial role of guardians and teachers in safeguarding children and young individuals against potential (digital) harm. Moreover, there has been a notable emphasis on individuals actively developing deepfake detection and awareness solutions, reflecting the proactive approach of researchers and technologists in mitigating the adverse effects of synthetic media manipulation. Furthermore, research was directed towards disseminating knowledge about deepfakes to a wider audience by raising general awareness and fostering digital literacy to empower individuals in discerning manipulated content. Overall, this study reveals the multifaceted efforts aimed at addressing the challenges posed by deepfakes across various societal sectors.

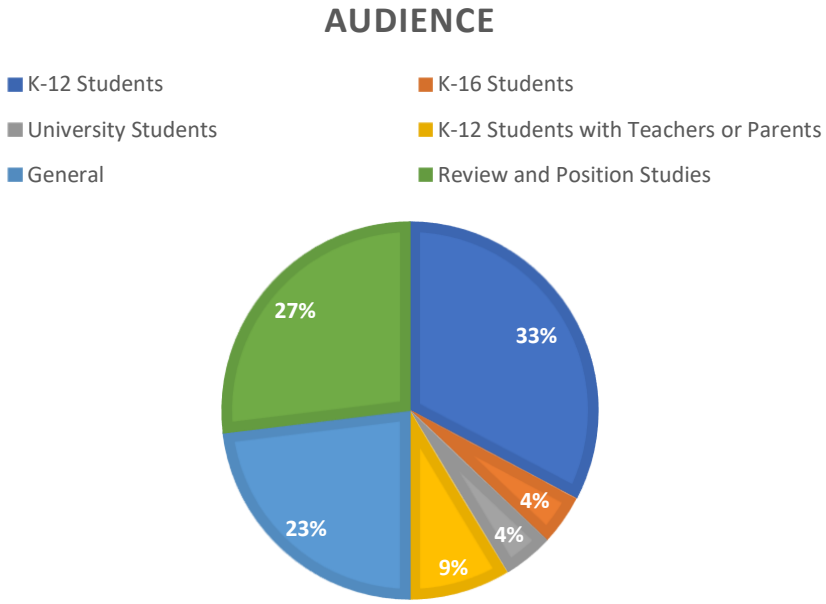


**Figure 2: Target audience of the studies analysed**

Given their scope and target audience, the application-based studies analysed can be classified in three clusters, as follows:

### 4.1 Cluster 1: Deepfakes

Kaate et al., (2023), Cho et al., (2023), Naffi et al., (2023), Cheres & Groza (2023), Leibowicz, McGregor & Ovadya (2021), and Ali et al., (2021). This cluster captures important elements of generating and detecting deepfakes, investigates challenges posed by this manipulation mechanism in the digital context, and presents valuable methods for building and supporting awareness and resilience against deepfakes. Specifically, Kaate et al., (2023) delves into user perceptions of deepfakes in design tasks providing insights on how the users' experience could be improved on this behalf. Aligned with this, Cho et al., (2023) emphasizes the pivotal role of addressing misinformation challenges through innovative approaches like immersive games learning by developing an escape room that proves to be effective for increasing awareness of misinformation tactics, including deepfakes, and fostering critical reflection among participants. Moreover, Naffi et al., (2023) and Ali et al., (2021) focus on educational interventions aimed at empowering youth to be aware and counter deepfakes. By providing experiential learning experiences and virtual workshops, these studies equip participants with the skills to discern the validity of online information and resist manipulation, irrespective of its source. Additionally, Leibowicz, McGregor & Ovadya (2021) highlight the importance of considering not only the technical dimensions but also the psychological and artistic aspects of deepfakes when building solutions for deepfakes awareness and mitigation strategies.

### 4.2 Cluster 2: AI Ethics and Cyber Security

Kilhoffer et al., (2023), Ali et al., (2021), Grover, Broll & Babb (2023), Freed et al., (2023), Williams et al., (2023), Zhang et al., (2023), and Aprin et al., (2021). This cluster focuses on the intersection of AI ethics and cyber security by exploring the meaning and implications of AI technologies for ensuring digital security. Kilhoffer et al., (2023), Williams et al., (2023), Zhang et al., (2023) stress the importance of AI literacy curricula, such as Creative AI and Developing AI Literacy (DAILy), designed to equip students with technical knowledge, ethical understanding, and awareness of AI's societal implications. These curricula employ active learning methodologies, embedding ethics into technical lessons to enhance critical thinking and prepare students for a future where AI is going to play an important role in their activities. Grover, Broll & Babb (2023) introduces the "AI & Cybersecurity for Teens" (ACT) initiative, aiming to integrate AI/ML concepts into high school cybersecurity curricula delving into critical issues like safety and bias, essential for meaningful interrogation of cybersecurity challenges. In this process, Freed et al., (2023) emphasize the importance of accounting the youth voices when building protective digital mechanisms against potential risks, and Aprin et al., (2021) build a web-based learning environment as a dedicated companion that supports students' understanding, critical thinking, and resilience against media threats that include AI-related cyber security risks like deepfakes.

### 4.3 Cluster 3: Digital Literacy

Wilmer et al., (2023), Lee et al., (2021), and Micallef et al., (2021). This cluster examines the knowledge, skills, and competencies required for individuals to effectively navigate and critically engage with AI-based technologies. While Lee et al., (2021) outline a comprehensive curriculum designed to educate students on various topics like bias and the impact on society and career by incorporating hands-on activities and discussions, Wilmer et al., (2023) emphasizes the importance of community-engaged projects in promoting digital literacy, particularly in rural and BIPOC communities, by augmenting existing temporal structures to foster circumstances conducive to repairing the information environment. Additionally, Micallef et al., (2021) presents the development and deployment of Fakey, a game designed to improve news literacy by simulating social media feeds and encouraging players to discern between mainstream and low-credibility content. Through analytics and user interviews, the study demonstrates that engagement with the game leads to improved recognition of mainstream content without compromising the ability to identify misinformation.

## 5. Design

Given the scope and the type of artefacts/solutions considered or proposed in the studies, these can be classified in the following three categories, as discussed below, and depicted in Figure 3:

- **Category 1 - Concept:** These studies imply reflecting on various concepts, methods, and techniques that are relevant to building deepfakes awareness gaming solutions either at a theoretical level or by means of aggregating this knowledge with valuable insights gathered from interviews, focus groups, and workshops (Kilhoffer et al., 2023; Mirsky & Lee, 2021; Johnson & Diakopoulos, 2021; Ali et al., 2021; Freed et al., 2023; Bu, Jiang & Zheng, 2023; Wilmer et al., 2023; Heidari et al., 2023; Masood et al., 2023; Dagar & Vishwakarma, 2022; Leibowicz, McGregor & Ovadya, 2021).
- **Category 2 - Tool:** These studies include building software artefacts or instantiating existing ones, as in the form of a digital game, multimedia software application, or directly creating deepfakes using an existing social media platform (Kaate et al., 2023; Grover, Broll & Babb, 2023; Cho et al., 2023; Naffi et al., 2023; Cheres & Groza, 2023; Aprin et al., 2021; Micallef et al., 2021).
- **Category 3 - Curricula:** These studies suppose building a curriculum for understanding, dealing, and countering digital threats like deepfakes in various scenarios (Lee et al., 2021; Williams et al., 2023; Zhang et al., 2023; Ali et al., 2021).

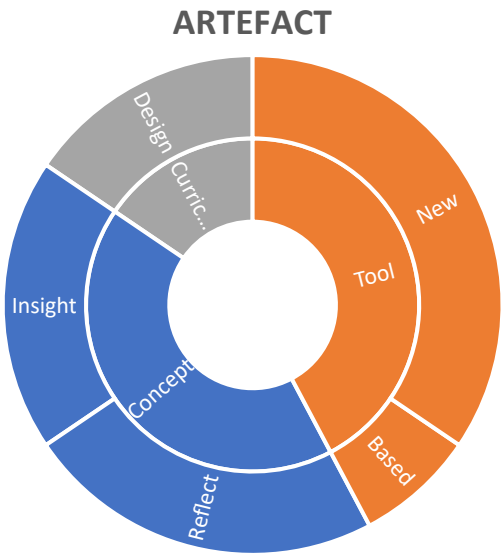


**Figure 3: Artefact types in the studies analysed**

Furthermore, a set of eighteen technical and educational design guidelines and requirements for building deepfakes awareness gaming solutions are gathered and further proposed. These capture elements that refer not only to the way of building and presenting content to the players, but also methodological aspects referring to the approach taken to develop and deploy the game, as presented in Figure 4 below:

| No | Design Guidelines and Requirements | Source |
|---|---|---|
| 1 | Take into consideration the multidisciplinary nature of the deepfakes phenomenon. | Kilhoffer et al., (2023), Bu, Jiang & Zheng (2023), Aprin et al., (2021) |
| 2 | Consider a collaborative, iterative, and progressive design process when multiple stakeholders are involved in building and deploying the game. | Cho et al., (2023), Masood et al., (2023), Micallef et al., (2021) |
| 3 | Focus on more commonly used social media platforms, e.g., TikTok, Snapchat, or Instagram. | Kilhoffer et al., (2023) |
| 4 | Ensure that the game is accessible to students with no prior STEM knowledge. | Ali et al., (2021) |
| 5 | Establish clear legal, ethical, and safety boundaries in respect to the deepfakes content used inside the game. | Freed et al., (2023), Bu, Jiang & Zheng (2023), Lee et al., (2021), Cheres & Groza (2023) |
| 6 | Consider an HLT approach to develop the game's structure and progression. | Ali et al., (2021) |
| 7 | Consider a GBL approach to engage students and enhance their understanding of deepfakes. | Ali et al., (2021), Grover, Broll & Babb (2023), Wilmer et al., (2023) |
| 8 | Reduce the cognitive load during gameplay by considering tailored and accessible learning activities and simplifying game mechanics and feedback mechanism. | Ali et al., (2021), Kaate et al., (2023), Grover, Broll & Babb (2023), Ali et al., (2021), Micallef et al., (2021) |
| 9 | Integrate educational components into the game to explain how deepfakes work as well as how they can be generated and detected in a way that encourages critical exploration, experimentation, and critical thinking. | Mirsky & Lee (2021), Johnson & Diakopoulos (2021), Williams et al., (2023), Zhang et al., (2023) |
| 10 | Use ready-to-use resources that are simplified, clear, and easily understandable by users. | Kilhoffer et al., (2023), Ali et al., (2021), Bu, Jiang & Zheng (2023), Naffi et al., (2023), Aprin et al., (2021), Ali et al., (2021) |
| 11 | Adopt a gamified approach that offers challenges, hints, rankings, or tests to users. | Kilhoffer et al., (2023), Mirsky & Lee (2021), Wilmer et al., (2023), Ali et al., (2021), Micallef et al., (2021) |
| 12 | Consider the use of multiple scenarios inside the game and realistic narratives that match current trends and anticipate future trends. | Mirsky & Lee (2021), Johnson & Diakopoulos (2021), Grover, Broll & Babb (2023), Cho et al., (2023), Masood et al., (2023), Dagar & Vishwakarma (2022), Cheres & Groza (2023), Zhang et al., (2023) |
| 13 | Showcase examples of technical solutions that could be used to detect and counter deepfakes | Johnson & Diakopoulos (2021), Naffi et al., (2023), Dagar & Vishwakarma (2022) |
| 14 | Incorporate high-quality interactive, multimedia elements, or mini games that allow players to experience firsthand how deepfakes are created and/or detected. | Bu, Jiang & Zheng (2023), Lee et al., (2021), Masood et al., (2023), Cheres & Groza (2023), Williams et al., (2023), Leibowicz, McGregor & Ovadya (2021), Micallef et al., (2021) |
| 15 | Include game elements that refer to the potential implications, consequences, and harm of deepfakes to individuals and society. | Johnson & Diakopoulos (2021), Lee et al., (2021), Williams et al., (2023), Zhang et al., (2023), Ali et al., (2021) |
| 16 | Reflect on diverse types and sources of synthetical content, e.g., audio, and visual-based. | Masood et al., (2023), Micallef et al., (2021) |
| 17 | Integrate assessment and verification mechanisms into the game when players need to distinguish between authentic and fake content and include post-play assessment. | Johnson & Diakopoulos (2021), Lee et al., (2021), Cho et al., (2023), Leibowicz, McGregor & Ovadya (2021) |
| 18 | Implement reporting and help systems that allows users to report their observations and findings and use guidance when needed. | Freed et al., (2023), Naffi et al., (2023), Dagar & Vishwakarma (2022) |

**Figure 4: Design Requirements and Guideline**

## 6. Conclusions

The emergence of deepfakes which are built using advanced deep learning techniques, has revolutionized synthetic media creation. Initially conceived for artistic and commercial pursuits, the widespread accessibility of advanced deep learning methodologies has democratized the production of deepfakes, particularly evident on social media platforms like TikTok, Instagram, and Snapchat. Nevertheless, the malevolent exploitation of deepfakes for disseminating disinformation and misinformation and inciting societal discord has precipitated profound ethical and socio-political concerns. Of particular significance is the vulnerability and susceptibility of adolescents, constituting a sizable demographic within the social media realm, to the manipulative influence of deepfakes, owing to their limited media literacy and heightened vulnerability to digital content. Despite scholarly, practitioner, and governmental efforts aimed at understanding, detecting, and mitigating the adverse effects of deepfakes, existent research and practical initiatives find themselves in an incipient phase of development and would further benefit for adopting a systematic approach to their design. Consequently, there is a need to develop tailored educational interventions, such as deepfake awareness games, to empower adolescents in discerning and counteracting the (harmful) effects of deepfake manipulation. To this end, this research proposes a series of eighteen design guidelines and recommendations for building deepfakes awareness solutions captured based on a systematic literature review drawing from multidisciplinary perspectives encompassing AI, social media, cyber security, and gamification as well as previous field expertise. This research continues by building and evaluating a gaming solution for deepfakes awareness of adolescents and aims to set a stone when building a digitally aware, safe, and robust future (digital) generation.

## References


Ali, S., DiPaola, D., Lee, I., Hong, J., & Breazeal, C. (2021). 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., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers, and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence*, no. *2*, pp. 100040.

Aprin, F., Manske, S., Chounta, I. A., & Hoppe, H. U. (2021, November). Is this fake or credible? a virtual learning companion supporting the judgment of young learners facing social media content. In *International Conference on Web-Based Learning*, pp. 52-60.

Ask, T. F., Lugo, R., Fritsch, J., Veng, K., Eck, J., Özmen, M. T., ... & Sütterlin, S. (2023). Cognitive flexibility but not cognitive styles influence deepfake detection skills and metacognitive accuracy.

Bu, J., Jiang, R. L., & Zheng, B. (2023, May). Research on Deepfake Technology and Its Application. In *Proceedings of the 2023 4th International Conference on Computing, Networks, and Internet of Things*, pp. 47-51.

Cheres, I., & Groza, A. (2023). The Profile: unleashing your deepfake self. *Multimedia Tools and Applications*, *82*(20), 31839-31854.

Cho, Y., Coward, C., Lackner, J., Windleharth, T. W., & Lee, J. H. (2023). The use of an escape room as an immersive learning environment for building resilience to misinformation. *Journal of Librarianship and Information Science*, 09610006231208027.

Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International journal of multimedia information retrieval*, vol. *11*, no. 3, pp. 219-289.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In The SAGE Handbook of Organizational Research Methods.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Freed, D., Bazarova, N. N., Consolvo, S., Han, E., Kelley, P. G., Thomas, K., & Cosley, D. (2023). Understanding digital-safety experiences of youth in the US In Proceedings of the 2023 CHI conference on human factors in computing systems, pp. 1–15.

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1520.

Grover, S., Broll, B., & Babb, D. (2023, March). Cybersecurity Education in the Age of AI: Integrating AI Learning into Cybersecurity High School Curricula. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education,* pp. 980-986.

Johnson, D. G., & Diakopoulos, N. (2021). What to do about deepfakes. *Communications of the ACM*, vol. *64*, no. 3, pp. 33-35.

Kaate, I., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2023, September). How Do Users Perceive Deepfake Personas? Investigating the Deepfake User Perception and Its Implications for Human-Computer Interaction. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter,* pp. 1-12.

Kilhoffer, Z., Zhou, Z., Wang, F., Tamton, F., Huang, Y., Kim, P., ... & Wang, Y. (2023). "How technical do you get? I'm an English teacher": Teaching and Learning Cybersecurity and AI Ethics in High School. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 2032-2032, IEEE.

Leibowicz, C. R., McGregor, S., & Ovadya, A. (2021, July). The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 736-744.

Lee, I., Ali, S., Zhang, H., DiPaola, D., & Breazeal, C. (2021, March). Developing middle school students' AI literacy. In *Proceedings of the 52nd ACM technical symposium on computer science education*, pp. 191-197.

Maathuis, C., Kerkhof, I., Godschalk, R., & Passier, H. (2023). Design Lessons from Building Deep Learning Disinformation Generation and Detection Solutions. In *ECCWS 2023 22nd European Conference on Cyber Warfare and Security* (No. 1). Academic Conferences and publishing limited.

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, vol. *53*, no. 4, pp. 3974-4026.

Micallef, N., Avram, M., Menczer, F., & Patil, S. (2021). Fakey: A game intervention to improve news literacy on social media. *Proceedings of the ACM on Human-Computer Interaction*,, pp. 1-27.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, vol. *54*, no. 1, pp. 1-41.

Naffi, N., Charest, M., Danis, S., Pique, L., Davidson, A. L., Brault, N., ... & Barma, S. (2023). Empowering Youth to Combat Malicious Deepfakes and Disinformation: An Experiential and Reflective Learning Experience Informed by Personal Construct Theory. *Journal of Constructivist Psychology*, pp. 1-22.

Naitali, A., Ridouani, M., Salahdine, F., & Kaabouch, N. (2023). Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Computers*, vol. *12*, no. 10, pp. 216.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Alonso-Fernández, S. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista española de cardiología*, vol. *74*, no. 9, pp. 790-799.

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023, November). Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large ai models. In *2023 International Conference on Computer and Applications (ICCA)* pp. 1-7, IEEE.

Taibi, D., Scifo, L., Bruno, N., & Fulantelli, G. (2023). Social Media Literacy to Support a Conscious Use of Social Media in Adolescents and Improve Their Psychological Well-Being: A Pilot Study. *Sustainability*, vol. *15*, no. 17, pp. 12726.

Theophilou, E., Lomonaco, F., Donabauer, G., Ognibene, D., Sánchez-Reina, R. J., & Hernàndez-Leo, D. (2023). AI and narrative scripts to educate adolescents about social media algorithms: insights about AI overdependence, trust and awareness. In *European Conference on Technology Enhanced Learning*, pp. 415-429.

Williams, R., Ali, S., Devasia, N., DiPaola, D., Hong, J., Kaputsos, S. P., ... & Breazeal, C. (2023). AI+ ethics curricula for middle school youth: Lessons learned from three project-based curricula. *International Journal of Artificial Intelligence in Education*, vol. *33*, no. 2, 325-383.

Wilner, T., Mimizuka, K., Bhimdiwala, A., Young, J. C., & Arif, A. (2023, April). It's About Time: Attending to Temporality in Misinformation Interventions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1-19.

Zhang, H., Lee, I., Ali, S., DiPaola, D., Cheng, Y., & Breazeal, C. (2023). Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of Artificial Intelligence in Education*, vol. *33*, no. 2, pp. 290-324.