From Adolescents’ Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok

Katrin Hartwig
Science and Technology for Peace and Security (PEASEC),
Technische Universität Darmstadt
Darmstadt, Germany
hartwig@peasec.tu-darmstadt.de

Tom Biselli
Science and Technology for Peace and Security (PEASEC),
Technische Universität Darmstadt
Darmstadt, Germany
biselli@peasec.tu-darmstadt.de

Franziska Schneider
Science and Technology for Peace and Security (PEASEC),
Technische Universität Darmstadt
Darmstadt, Germany
franziska.schneider@stud.tu-darmstadt.de

Christian Reuter
Science and Technology for Peace and Security (PEASEC),
Technische Universität Darmstadt
Darmstadt, Germany
reuter@peasec.tu-darmstadt.de

ABSTRACT

Misinformation poses a recurrent challenge for video-sharing platforms (VSPs) like TikTok. Obtaining user perspectives on digital interventions addressing the need for transparency (e.g., through indicators) is essential. This article offers a thorough examination of the comprehensibility, usefulness, and limitations of an indicator-based intervention from an adolescents’ perspective. This study ($N = 39$; aged 13-16 years) comprised two qualitative steps: (1) focus group discussions and (2) think-aloud sessions, where participants engaged with a smartphone-app for TikTok. The results offer new insights into how video-based indicators can assist adolescents’ assessments. The intervention received positive feedback, especially for its transparency, and could be applicable to new content. This paper sheds light on how adolescents are expected to be experts while also being prone to video-based misinformation, with limited understanding of an intervention’s limitations. By adopting teenagers’ perspectives, we contribute to HCI research and provide new insights into the chances and limitations of interventions for VSPs.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Social media**; *Empirical studies in collaborative and social computing*.

KEYWORDS

misinformation, disinformation, fake news, user intervention, teenagers, adolescents, social media, video-sharing platforms

ACM Reference Format:
Katrin Hartwig, Tom Biselli, Franziska Schneider, and Christian Reuter. 2024. From Adolescents’ Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642264

1 INTRODUCTION

Video-sharing platforms (VSPs) like TikTok, which was originally popular for content such as entertaining lip-sync and dance videos for teenagers, is now increasingly filled with questionable, sometimes dangerous and misleading content designed to appeal to and influence people of all ages. The serious consequences of misleading information on VSPs are particularly evident in crises such as the Russian war of aggression against Ukraine [72]. Not only intentionally misleading information can cause great damage, but also false information that is spread unintentionally. ‘Misinformation’ is often used as an umbrella term that includes both intentionally misleading information (i.e., “disinformation” or “fake news”) and unintentionally misleading information (i.e., “misinformation”) [1, 23, 91] - we will use the term accordingly.

Human-computer interaction (HCI) research has recently focused on VSPs such as TikTok, YouTube, and Twitch, not only in terms of misinformation, but also in contexts such as mental health or social participation [60]. The question of the transferability of existing findings regarding text-based platforms such as Twitter, where much of the research to date has taken place, to video-based platforms, is highly relevant from an HCI perspective. Misinformation and ways to support users in dealing with it have been researched for many years. More recently, first in-depth insights into misinformation on VSPs have been gained, revealing their relevance as a breeding ground for misinformation [2] and an extensive range of layers (e.g., video, audio, captions) offered by VSPs [62] having a potential to mislead. Research has delved into how adolescents assess credibility in video content, including considerations of comment sections on platforms like TikTok [40]. While existing HCI research emphasizes the relevance and challenges of misinformation on VSPs, we will fill the research gap of considering the user perspective on how to address that type of content with its multimodal potentials to mislead. Even after years of research on digital misinformation interventions and controversial discussions on their suitability, the field is far from a decisive

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

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642264

solution. Therefore, further user-centered exploration of specific needs, capabilities, and perceptions is relevant.

This paper aims to extend existing findings with qualitative user-centered insights to sharpen digital misinformation interventions as one building block for dealing with misinformation, to further delineate their possibilities and limitations, and thus to contribute to misinformation research from an HCI perspective. Previous research suggests that transparency plays a significant role in establishing trust among users in digital interventions [51] and minimizing reactance or other backfire effects [64]. To address the preference for transparency (i.e., for "approaches providing [users] with an opportunity for informed decisions" [51, p.14], for instance using explanations on why a content was labeled as misinformation), we focus on an indicator-based approach that presents comprehensible and user-centered cues for evaluating video-based information. We define an *indicator-based misinformation intervention* as a digital countermeasure that uses identifiable characteristics (indicators) of misinformation to assess and convey the credibility of online content. These indicators (e.g., attention-grabbing layouts or conspiracy theory endorsements) are presented to users for both immediate feedback and educational purposes, enabling them to develop skills in assessing the credibility of content. Typically, these interventions involve highlighting identified indicators within the content (e.g., using color-code [58]). Indicator-based approaches have been developed and controversially discussed in existing research, especially in relation to text [58]. Our aim is to extend these findings to video-based indicators such as edited images, emotion evoking sounds or facial expressions, and refuting reactions to the video. With VSPs such as TikTok as a central platform especially for adolescents, we are particularly focusing on them as potential users who are, on the one hand, regarded by many as highly capable digital natives, but, on the other hand, as particularly vulnerable by others. We conducted a twofold study with 39 adolescents between the ages of 13 and 16 years. First, we used a mixed-methods approach with focus groups and individual surveys to assess existing strategies and the comprehensibility of video-based indicators as a potential extension (Step 1). Based on the findings, an indicator-based prototype smartphone app was developed as a misinformation intervention for TikTok which was qualitatively evaluated in individual think-aloud sessions (Step 2). The sequential studies provide an opportunity for triangulation of data and in-depth findings.

We advance misinformation research by applying existing knowledge about indicator-based interventions to the very current modality of short-videos. In doing so, we adopt the user perspective of adolescents as a particularly relevant user group. Our core contributions (C) and findings (F) are first (C1) evaluating the comprehensibility and perceived usefulness of video-based indicators. In doing so, we (F1) found how indicators on diverse levels (e.g., layout, profile, interactions) expand and confirm the perceptions of adolescents, and (F2) how particular characteristics of manipulated videos are applied intuitively as indicators. We then (C2) implemented and qualitatively evaluated a smartphone app prototype as an indicator-based misinformation intervention and received (F3) an overall positive feedback on the approach and identified participants' substantial understanding of its features. Furthermore, we identified insights into (F4) potentials for transferability of the extended knowledge and skills. Finally (C3), we assessed opportunities, challenges, and limitations of the indicator-based approach and identified (F5) transparency of the indicators as central reason for the intervention's positive assessment, but also (F6) its limitations, such as adolescents' blind trust in the tool and the lack of realistic concerns.

2 RELATED WORK

Our work contributes to the design of user-centered digital misinformation interventions for VSPs like TikTok. We discuss related work, addressing how social media is faced with an overabundance of information, and the role of VSPs compared to primarily text- or image-based platforms (see Section 2.1). We shed light on digital misinformation interventions as one possibility to combat the effects of misinformation (see Section 2.2), delve into indicator-based misinformation interventions (see Section 2.3), and give an outline of the role of adolescents compared to others in terms of their susceptibility to misinformation (see Section 2.4). Finally, we summarize the resulting research gaps and present research questions to address them (see Section 2.5).

2.1 Misinformation on Video-Sharing Platforms

Misinformation research continuously generates new questions due to advancing technological possibilities, such as artificial intelligence and novel multimodal platforms. Social media in general has long been a central arena for the spread of misinformation in this regard, particularly in times of crisis [26, 85, 86]. While text-based platforms have dominated the social media landscape in the past, VSPs have recently gained relevance, transforming TikTok into one of the most successful platforms in the world, especially among younger people. HCI research on VSPs particularly addresses online communities like video game streamers and social participation [10]. This also extends to positive aspects of the platform, such as fostering creativity [66], inspiring playful technology [28], or connecting communities [60, 78]. However, recent events have demonstrated the vulnerability of VSPs to misinformation and other harmful content [2, 12, 55, 62]. In this regard, studies investigate how recommendations affect cross-platform sharing [20], who creates misinformation on TikTok and how users respond to them [2]. Beyond misinformation, research examines, e.g., how dark patterns are applied on TikTok and other social media platforms [59], how VSPs expose creators to hate and harassment [88], and how young people perceive digital safety, including harassment, financial fraud, but also misinformation and deep fakes [31]. Research emphasizes the extensive range of data (e.g., video, audio, descriptions, comments) offered by VSPs [62], which could potentially mislead users. Our work address these misleading potentials through the use of multimodal indicators. A specific relevance appears to apply to comment sections [62], especially among young users, when searching for and evaluating online information [40] - a finding that our work builds on. While there are different approaches to combating misinformation - for example, by teaching media literacy in schools or by supporting the work of professional journalists - one way to help deal with the overabundance of information online is to develop digital misinformation interventions.

2.2 Digital Misinformation Interventions

Developing user-centered approaches to combat misinformation poses a challenge within the HCI research community. To address this issue, digital misinformation interventions are employed to assist users in processing online misinformation, expanding on efforts in education and journalism. The term 'digital misinformation intervention' has already been established by several researchers [9, 75, 76]. These interventions vary widely in their primary goal and encompass, for instance, soliciting user feedback after automatic detection or nudges to reduce sharing. While many countermeasures target automatic detection [82, 94] (e.g., through machine learning), studies with a stronger HCI focus explore post-detection decisions or aspects detached from detection (e.g., default nudges for reflection). User-centered misinformation interventions "go beyond a purely algorithmic back-end solution and exert a direct influence on the user in the form of information presentation or withholding" [38, p. 2]. For example, there are approaches that provide a correction of misinformation by displaying a link to a fact-checking website, debunking videos, or corrections within the comment section [5, 14, 57]. Other work suggests providing a binary label to mark content as false [11], however that approach has shown to be less accepted by users when transparent explanations were missing [51]. Trust and distrust play a significant role when aiming to design effective interventions [5]. While some of the studies report promising initial results [22], e.g., to reduce sharing or engagement with misinformation, others demonstrate how common interventions have limited effects in isolation and are more successful when combined (e.g., suspending algorithmic amplification combined with a nudge) [9].

Efforts have been made to offer comprehensible indicators that guide the autonomous evaluation of problematic content, promoting critical thinking or media literacy. In this paper, in line with other researchers, we define media literacy as the ability to decode, evaluate, analyze, and produce both print and electronic media, that is, to have internalized a sense of "critical autonomy" in dealing with all media [6]. However, media literacy has received criticism for being only one of many components of the complex information space [19, 40], for relying on rationality [17], and for creating a false sense of confidence [19], prompting calls to "rethinking media literacy in the age of platforms" [19, p. 17]. Nonetheless, studies indicate that media literacy education can promote critical thinking and alter behavior [45, 48, 92] - also regarding misinformation [84] and especially among younger people [93]. In our study, we draw on this optimistic view.

2.3 Indicator-Based Misinformation Interventions

Studies highlight the significance of transparency and comprehensibility in interventions, as they foster trust [33, 51] and can thus be considered a crucial requirement when designing user-centered countermeasures [79] - an insight that guides our approach. Indeed, users prefer to understand why content is marked as misinformation as opposed to receiving binary classifications [51]. To address this, HCI research explores indicator-based interventions regarding misinformation and related phenomena (e.g., propaganda) [51, 58, 79]. For example, Bhuiyan et al. [16] examined how indicators such as information about the author, can be utilized to promote trust, from the perspectives of both news consumers and journalists. More generally, trustworthiness indicators have been evaluated in terms of perceived utility and visualization preferences in social media posts, revealing positive perceptions and a preference for simple visualizations of indicators to reduce the cognitive load [33]. Indicators as immediate user feedback enable users to understand characteristics of misinformation and encourage the development of own assessment skills [79]. While current research primarily focuses on textual content, some insights extend to images and videos. For instance, Sherman et al. [81] identified the source of information as key indicator when assessing different types of content, including fake videos. Although users generally appreciated displaying indicators as a message, favoring simplicity and clarity, the study identified a problem with the overgeneralization of indicators. Research indicates promising advantages (e.g., indicators reducing uncertainty [33], and aligning with users' preexisting mental models and practices [81]) and potential positive effects on trust [16, 51], perceived utility [33], and the development of autonomous assessment skills [79]. Nevertheless, concerns and challenges persist, with some users (and especially adolescents) still perceiving interventions as patronizing, underscoring the challenge of striking the right tone [40], and other users rejecting indicator-based interventions when their development was not transparent or their design was biased [33]. Our research delves into the nuances of misinformation mitigation and recognizes that while not universally applicable, indicator-based interventions are a promising step towards addressing user needs and preferences.

Identifying comprehensible and useful indicators for credibility assessment is crucial to the development of indicator-based misinformation interventions. Research has primarily explored indicators of misinformation or related phenomena in textual content, often with an emphasis on automatic detection (e.g., to use them as features for training misinformation detection algorithms) [82, 94] and in some cases with a user-centered or social science focus [58, 79, 89]. While text-based indicators (e.g., linguistic characteristics of propaganda techniques such as exaggeration or loaded language [58]) might be partly applicable to other modalities, user-centered indicators regarding video content still require more exhaustive investigation, complementing research that has gained initial insights. Some studies specifically identified or evaluated selected video indicators. This includes the relevance of comment sections with critical reactions of other users, or videos evoking strong emotions [40, 62]. Research emphasizes loaded language or attention-grabbing layout as indicators on social media, including VSPs [93], or highlights the potential of filters and voice changers to mislead [31]. Others more generally gained insights into how users navigate VSPs and tackled indicators for misinformation only briefly, for instance by taking a closer look at hashtag topics indicating misinformation on TikTok [55], or by taking very specific perspectives like characteristics of misinformation in videos on urological topics [65, 95]. Other research tackles misinformation and its indicators on VSPs indirectly, investigating aspects that are closely related to the modality like emotion recognition in videos [15]. Niu et al. [62] provide an overview on multiple components and layers of videos (e.g., sound, profile, reactions) with the potential to be misleading - a foundation we build upon to derive specific indicators for VSPs (e.g., outdated or mismatching sound, facial expression evokes strong feelings). VSPs like TikTok offer a variety of patterns and indicators that have not yet been thoroughly evaluated from a user perspective, hence the motivation for our work. The systematic overview of derived indicators and its sources can be found in Table 3 (Appendix). We further summarize the procedure of selecting our indicators in Section 3.6.

2.4 Adolescents as Vulnerable Users

Evidence indicates that susceptibility to misinformation on social media varies among users. Both individual attributes (e.g., decision-making style [8], cognitive ability [25]), and demographic factors (e.g., age [47], language skill [73]) play a role. Young users have been extensively studied due to their exposure to media literacy education, growing up with unique information sources while appearing to have a short attention span, and their status as 'digital natives', which grants them familiarity with new media such as VSPs and technologies like as deepfakes and filters.

Research diverges on the digital and media literacy of young people, with some demonstrating strong literacy of 18-24 year-olds in specific areas, such as health misinformation [30]. The fact, that young individuals belonging to 'Gen Z' generally have the capacity for a significant degree of digital literacy and a basic desire to seek information from different sources was supported by a WHO report based on a survey across 24 countries [90].

However, it is crucial to note that other studies display tendencies to prematurely share misinformation among adolescents under the age of 18 [41, 42]. Limited critical thinking skills and overconfidence contribute to their susceptibility [68]. Lower self-efficacy [67] and limited concern for information trustworthiness among 11-13 year-olds [27, 56] are associated with increased misinformation sharing. Adolescents tend to overestimate their ability to assess digital information accurately [63, 68, 70, 96]. In-depth qualitative work on the information practices of 13-24 year old adolescents and young adults suggests that they do not conceptualize information and online information processing as isolated and narrow processes, but as broad and situated in a social context [40]. This highlights the role of social cues and peer influence in adolescent's online information handling - a conceptualisation from which new vulnerabilities can potentially emerge.

These vulnerabilities, combined with high exposure to online misinformation due to excessive online activity, underscore the vulnerability of young adolescents. Their limited critical thinking skills, responsiveness to social cues, and limited digital literacy emphasize the special need for tailored attention and support in order to critically evaluate online information.

2.5 Research Gap

This study advances HCI research on misinformation interventions through the identification and assessment of user-centered indicators for misinformation on VSPs like TikTok. Based on this, we developed a prototype smartphone app designed as a digital misinformation intervention for adolescents. The study addresses gaps in the following areas:

1st gap: Focus on VSPs. The vast majority of research on social media misinformation centers on text-based content. However, recent crises such as the Russian war of aggression against Ukraine and the COVID-19 pandemic have shown that VSPs like TikTok are a highly relevant breeding ground for misinformation [2]. Existing HCI research highlights the significance of this challenge and the need for further research on how to address this specific type of content.

2nd gap: User-Centered Interventions. User-centered countermeasures that promote media literacy are better received by users when provided with comprehensible explanations [51]. Researchers have emphasized the importance of finding interventions that counteract reactance and take into account end-users' needs for comprehensibility [51, 64]. This can be achieved by providing comprehensible indicators of misinformation as user feedback - an approach that has already been partially addressed for text-based content [7, 32, 39, 58], but requires further development from a user perspective [34, 51]. Video-based content, with its multimodality, is interesting in that it presents its own unique opportunities for misinformation [62], which in turn require new strategies for user feedback (e.g., regarding emotional music, filters, and false context of sounds).

3rd gap: Adolescent Vulnerability. Adolescents are particularly susceptible to misinformation, and research suggests potentially harmful overconfidence [19, 77]. Addressing the needs of adolescents in dealing with misinformation on VSPs like TikTok requires further investigation aligning with their tone and style preferences [40], building on HCI research that has already produced significant findings [40, 62] and by involving adolescents in a design process from the beginning.

Considering related studies and combining the resulting gaps, our overarching goal is to answer the following research questions:

RQ1: *How do adolescents evaluate misinformation indicators in terms of comprehensibility and usefulness?*
RQ2: *How can indicators be applied to a smartphone app as a digital misinformation intervention to assist adolescents on TikTok?*
RQ3: *How do adolescents perceive chances and limitations of an indicator-based digital misinformation intervention?*

3 METHODOLOGY

In the following, we provide details on our participants, ethical considerations, research method, analysis, and stimuli. To answer our research questions, we conducted a twofold study with 39 adolescents aged 13 to 16 in Germany in the summer of 2023. Our study consists of two steps of data collection (see Figure 1): focus groups (Step 1) and individual think-aloud Sessions (Step 2), involving a different set of participants.

In Step 1, our goal was to gain insight into which indicators adolescents autonomously use to assess misinformation on TikTok, and into how they evaluate established indicators in terms of comprehensibility and perceived usefulness. This allowed us to identify a set of potential user-centered indicators for VSPs, and for TikTok in particular, thus primarily targeting RQ1. We chose to implement focus groups in order to gain a variety of insights from the interactions and discussions between participants [53]. Interactive focus groups facilitate the exchange of ideas and the exploration of different viewpoints, which corresponds to our overall aim of identifying the most promising indicators for adolescent users.

For Step 2, we developed a smartphone app prototype as a digital misinformation intervention based on the previously in Step 1 established indicators, and evaluated how adolescents interacted with it, rated its usefulness, and identified challenges and opportunities – primarily to address RQ2 and RQ3. We thereby employed the think-aloud method [54] to gain rich insights into user perceptions during real-time use, rather than retrospectively [74], which is a method widely used in HCI and misinformation studies [61, 68, 74]. In contrast to focus groups, it provides a more individualised and introspective insight into participants' thought processes and decision making as they interact with the prototype independently. By combining both focus group (Step 1) and think-aloud (Step 2) methods, our aim was to gain both depth and breadth of insight and to answer the research questions from a variety of perspectives. Overall, both methods provide rich insights and complement each other to facilitate an understanding of users' needs, behavior, and experiences.

3.1 Recruitment and Participants

We carried out Step 1 with an entire school class and Step 2 with a separate school class in addition to a cohort of four additional female participants from a youth center to ensure reasonable gender parity. Recruitment was mediated through contacts with a local digital media education center - which, however, had not previously conducted any thematically relevant training courses with the specific pupils in our study. The majority ($N = 38$) of the participants were students at an integrated comprehensive school, i.e., they are mixed in terms of their intended school leaving qualifications (e.g. Abitur, Mittlere Reife, Hauptschulabschluss in Germany), and one additional student came from a secondary high-school ('Hauptschule'). The overarching inclusion criterion was the age range of 13-16 years to gain insights specifically from young adolescents. In addition, the choice of school for the recruitment of participants ensured the coverage of diverse levels of education and learning abilities. The specific classes (one for Step 1, another for Step 2) were chosen on the basis of time availability within a project week at the date of data collection. All pupils attending class at the time of data collection were included in the study and no specific exclusion criteria were applied. For the focus groups (Step 1), the participants were divided into three age- and gender-balanced groups with the help of the teacher. This arrangement was designed to break up cliques and ensure an open and productive discussion atmosphere. The groups as a wholes were rather homogeneous, as they all consisted of pupils from the same class. No further steps were taken to make the groups more homogeneous, e.g. in terms of social media usage, since the remaining degree of heterogeneity was considered useful for stimulating interesting discussions among the participants.

For Step 1 ($N = 21$), participants were between 14 and 16 years old (*median* = 15). Ten were male, ten were female, and one was non-binary. For Step 2 ($N = 18$), participants were between 13 and 16 years old (*median* = 13.5), seven were female and eleven were male. Since participants interacted with a smartphone in Step 2, we also documented their usual operating system. Six reported using Android, seven iOS, four both, and one was unsure. Participants also reported using TikTok daily ($N = 25$), several times a week ($N = 5$), or never ($N = 9$), but all had already used it in the past.

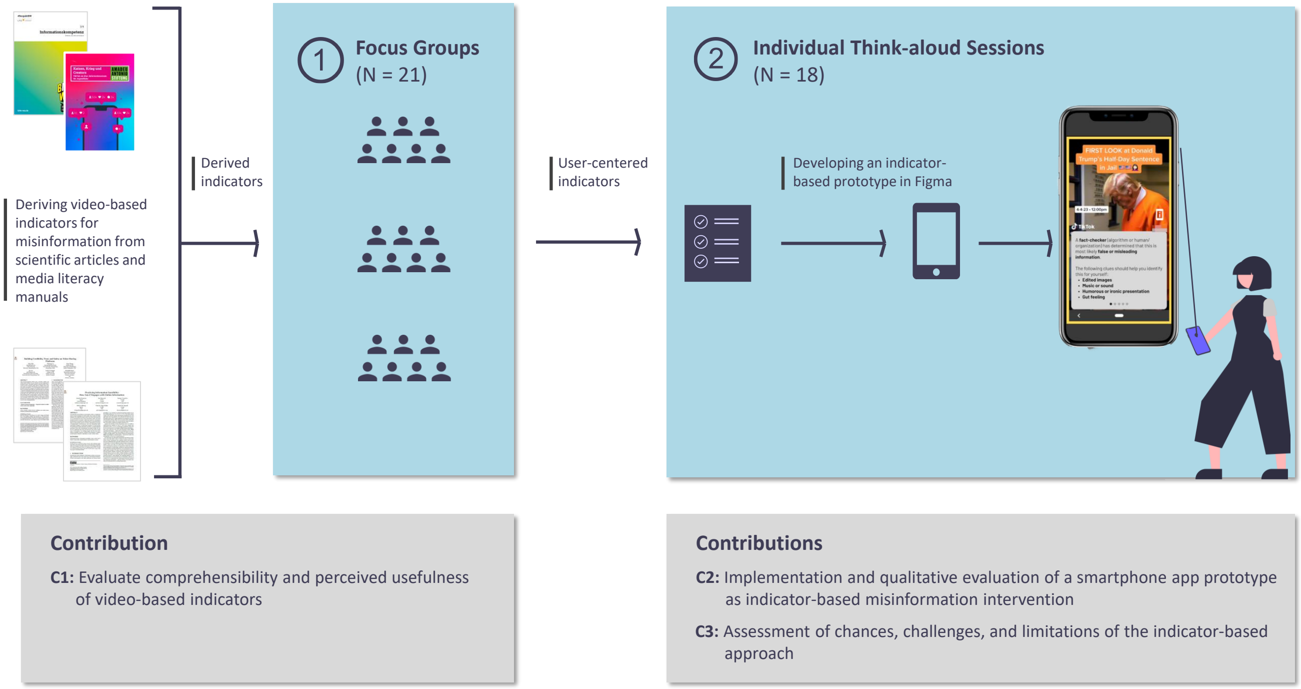
3.2 Ethical Considerations

When involving minors, ethical planning and conduct are crucial. Accordingly, the positive voting of the university's ethics committee (IRB Number EK 47/2023) was obtained in advance. Stimuli and study procedures were coordinated with a local media education center and teachers to include interdisciplinary expertise. The content shown was adapted to the age of the participants, excluding any violent or disturbing images (e.g. war, explosions) and politically biased messages. The videos were presented within a TikTok simulation to control the selected content. The truthfulness of the stimuli was clarified immediately after the study and the adolescents were provided with additional written information. Explicit informed consent (in age-appropriate language) for participation and data use was obtained in advance from both parents and youth, in accordance with the permission of our institutional ethics review board. Participants received a €15 book voucher as compensation.

3.3 Step 1: Mixed-Methods User Study with Focus Group Discussions

3.3.1 Mixed-Methods Approach. Individual paper-based surveys provided insights into autonomous assessment strategies immediately after watching a TikTok video. Focus groups were chosen as they enable discussions among participants and generally provide a broad range of opinions [53]. They are frequently applied in HCI in general [53] and particularly in the context of misinformation [31]. Thereby, in-depth explanations were generated, which was crucial for assessing participants' comprehension of certain indicators of misinformation, for identifying which indicators require improved explanations, and for understanding our participants' general needs. This mixed-methods approach allowed participants the necessary space to think individually before engaging in the group discussions. By triangulating the data, we gained deep insights into our young participants' responses.

3.3.2 Study Procedure. Three parallel focus group sessions of approximately equal size were each led by a member of the research team, following a structured guideline with standardized language. Each session lasted 120 minutes. After a brief introduction, written informed consent was obtained and demographic information was collected before audio-recording began. Participants sequentially viewed five TikTok videos, each following a systematic procedure (see Section B.1, App. for details): (1) Participants answered general questions individually on a paper questionnaire regarding the familiarity of the video and the credibility assessment. (2) The focus group received multiple indicators for misinformation and discussed their understanding and usefulness (see Table 3 for sources and representation of indicators). Finally, researchers provided clarification on the videos' truthfulness. One of the five videos contained only factual content. Thus, no cues for misinformation were displayed. (3) After following the procedure for all five videos, we presented three additional indicators that were not included in the selected videos but still seemed promising to evaluate based on our


Deriving video-based indicators for misinformation from scientific articles and media literacy manuals
Derived indicators
1 Focus Groups
(N = 21)
User-centered indicators
2 Individual Think-aloud Sessions
(N = 18)
Developing an indicator-based prototype in Figma
FIRST LOOK at Donald Trump's Half-Day Sentence in Jail
4-4-23 • 12:00pm
TikTok

A **fact-checker** (algorithm or human/ organization) has determined that this is most likely **false or misleading information**.

The following clues should help you identify this for yourself:

- **Edited images**
- **Music or sound**
- **Humorous or ironic presentation**
- **Gut feeling**

Contribution

C1: Evaluate comprehensibility and perceived usefulness of video-based indicators

Contributions

C2: Implementation and qualitative evaluation of a smartphone app prototype as indicator-based misinformation intervention

C3: Assessment of chances, challenges, and limitations of the indicator-based approach

Figure 1: Flow of our twofold study design demonstrating key contributions of each step.

literature review. Again, participants discussed their comprehensibility and usefulness. (4) Afterwards, the focus groups engaged in general questions about the indicators overall. The study procedure was identical for all three groups, i.e. all participants watched the same videos (in a different order) and were asked the same questions. The general aim of the focus groups was to obtain a variety of perspectives. For this reason, we decided to conduct three separate focus group discussions rather than one large group in order to create a constructive atmosphere and reduce the influence of very dominant individuals. The range of three focus groups – also in contrast to only one large focus group – allowed us to have in-depth discussions that provided different insights from a variety of participants while maintaining a sufficient level of control over the dynamics of the conversation. Subsequently, all three researchers reported similar experiences in their focus groups.

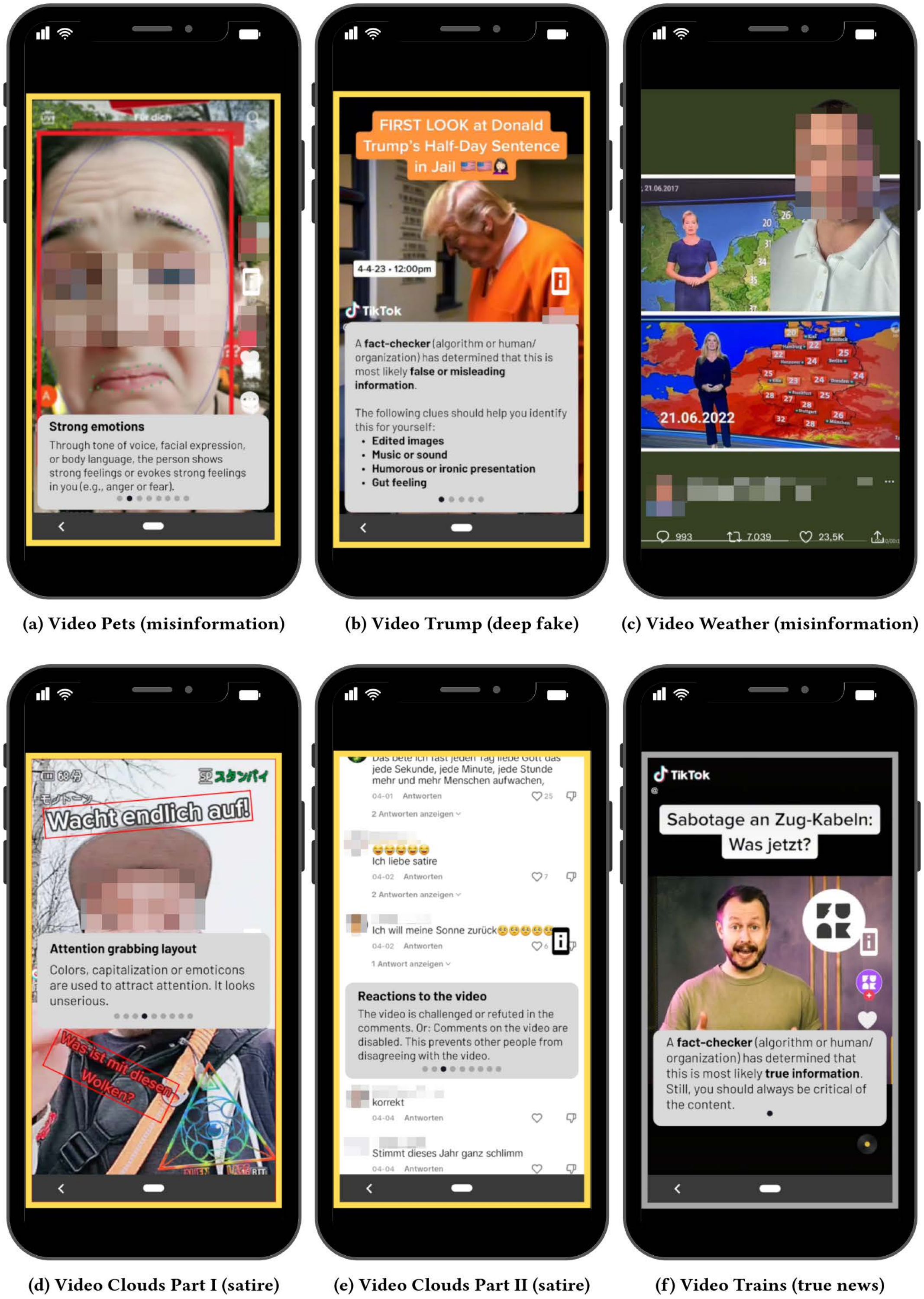
3.4 Step 2: Individual Think-aloud Sessions

We designed and evaluated a smartphone app prototype (see Figure 2) as a digital misinformation intervention for TikTok, incorporating and refining the indicators from Step 1 (see Table 3 for representations in Step 1 and 2).

3.4.1 Developing a Figma Prototype. We designed the smartphone app (= 'Misinfo-App') as a prototype using *Figma*[1], where TikTok videos are embedded in a simulated 'ForYou page' as can be found on TikTok. Different views of our Figma prototype design can be seen in Figure 2. The prototype corresponds to a realistic click dummy with a simulated detection. To disrupt the everyday user experience, the app was integrated into the prototype as an *active intervention*: From the simulated ForYou page, a video can be shared with the Misinfo-App using the share button built into TikTok. The user is then taken to an overview of the detection results (see Figure 2b and 2f). If misinformation was successfully detected, the corresponding indicators are listed there. By swiping, the user can switch to the detailed descriptions (see Figure 2a, 2d, and 2e). Some indicators are associated with appropriate highlighting of the referenced items in the video (e.g., display of the comments (Figure 2e) or the profile for the video). An arrow leads the user back to the ForYou page.

3.4.2 Study Procedure. The one-on-one think-aloud sessions were also conducted by three researchers, following a pre-defined systematic guideline. Each session lasted 40 minutes. (1) After giving a brief introduction and obtaining written informed consent, a short simulation of the think-aloud method was demonstrated. Participants were generally instructed to "think aloud" while they were using the smartphone. We recorded audio and screen touch to evaluate our participants' interaction with the app. (2) Participants were instructed to imagine being on their ForYou page on TikTok, using our smartphone prototype. They were able to click and swipe everywhere while watching the videos. (3) They were asked to share each video with our Misinfo-App to check its credibility and to look at the app's output for as long as they wanted. All participants thoroughly swiped through the detailed indicator descriptions. (4) After watching and sharing all four videos, the participants switched

[1] https://figma.com


Für dich
FIRST LOOK at Donald Trump’s Half-Day Sentence in Jail 🇺🇸🇺🇸🤦‍♀️
4-4-23 • 12:00pm
TikTok
21.06.2017
21.06.2022
993
7.039
23,5K
Wacht endlich auf!
Was ist mit diesen Wolken?
Sabotage an Zug-Kabeln: Was jetzt?

Strong emotions

Through tone of voice, facial expression, or body language, the person shows strong feelings or evokes strong feelings in you (e.g., anger or fear).

A **fact-checker** (algorithm or human/ organization) has determined that this is most likely **false or misleading information**.

The following clues should help you identify this for yourself:

- **Edited images**
- **Music or sound**
- **Humorous or ironic presentation**
- **Gut feeling**

(a) Video Pets (misinformation) (b) Video Trump (deep fake) (c) Video Weather (misinformation)

Attention grabbing layout

Colors, capitalization or emoticons are used to attract attention. It looks unserious.

Das bete ich fast jeden Tag liebe Gott das jede Sekunde, jede Minute, jede Stunde mehr und mehr Menschen aufwachen,
04-01 Antworten 25
2 Antworten anzeigen

😂😂😂😂😂
Ich liebe satire
04-02 Antworten 7
2 Antworten anzeigen

Ich will meine Sonne zurück🥺🥺🥺🥺🥺
04-02 Antworten 6
1 Antwort anzeigen

Reactions to the video

The video is challenged or refuted in the comments. Or: Comments on the video are disabled. This prevents other people from disagreeing with the video.

korrekt
04-04 Antworten

Stimmt dieses Jahr ganz schlimm
04-04 Antworten

A **fact-checker** (algorithm or human/ organization) has determined that this is most likely **true information**. Still, you should always be critical of the content.

(d) Video Clouds Part I (satire) (e) Video Clouds Part II (satire) (f) Video Trains (true news)

Figure 2: Overview on the stimuli videos we used in Step 1 (all) and Step 2 (all but c) demonstrating several views of the smartphone app’s UI within our Figma Prototype. Faces of private persons were pixelated for publication. UI elements were translated from German to English.

from our prototype to an online survey to fill out the SUS and a demographics questionnaire. To accommodate the young age of the participants, the language of the SUS was slightly modified. We adopted the wording of the study by Putnam et al. [71] and slightly reformulated all items that have a grade level >6 in the original (see Section B.3, App.). (5) Finally, we asked open-ended questions to gain deeper insights into the overall indicator-based approach, its usefulness, and any concerns about an underlying algorithm. A detailed description of the study procedure can be found in Section B.2.

3.5 Analysis

Steps 1 and 2 generated rich qualitative data from the focus groups and individual think-aloud sessions, as well as additional quantitative data from the individual survey items. We locally transcribed the audio material with *Whisper*[2], followed by a thorough manual revision. Participants' responses were anonymized.

We then (a) analyzed the quantitative survey items calculating descriptive statistics ($mean, median, frequencies$) and (b) clustered the free-text question thematically according to the individually mentioned assessment strategies from Step 1. We (c) employed thematic analysis for the focus group discussions (Step 1) and think-aloud sessions (Step 2), which is common for this type of study [18, 35] and which serves to identify and interpret thematic patterns within qualitative data [18]. Overarching themes were thus derived from a set of pre-defined codes that were created prior to the analysis and were iteratively added as new topics emerged during coding. Two coders independently applied the codebook (see Table 4, App.) to the data and any ambiguities were discussed to achieve a consensus. Inter-coder reliability, measured by Cohen's kappa score, demonstrated substantial agreement ($k > 0.61$) across all iterations. The themes as thematic clusters derived from our codes form the basis for our results reports in Sections 4.1, 4.2, and 4.3.

3.6 Stimuli

The stimuli for Steps 1 and 2 were selected and revised in consultation with the local digital media education center, teachers (especially with regard to simple language, age appropriateness, and representative mapping of content on the adolescents' ForYou page), and the ethics committee. They comprise real-world TikTok videos and indicators of misinformation that could help assess content on VSPs. The videos and indicators are similar in Step 1 and Step 2 (see Table 3 for representations in both steps).

Videos. An overview of all included videos can be found in Figure 2 and Table 5 (Appendix) provides a detailed description of the content of the videos. TikTok videos were selected as stimuli following a systematic process. (1) We consulted official fact-checking websites to identify current topics that have been officially debunked which covered a wide range of misinformation. Either these sites already linked to specific recent TikTok videos, or we manually searched TikTok for these videos, referring back to the officially debunked content. (2) We excluded videos that were not age appropriate. For all videos, we chose those that were most suitable for the age group and were approved by the digital media education center and teachers involved in the study. (3) We made sure to include diverse forms of misinformation. While misinformation types can take various forms [49], we chose our selection of videos to represent a realistic sampling of diverse forms of misinformation on VSPs without a claim to completeness: deep fakes (Figure 2b; manipulated video showing Donald Trump in prison), satire with potential to mislead, as well as conspiracy theories (Figure 2d/2e; a TikToker making fun of common conspiracy theories by pointing to clouds that block out the sun and referring to the pharmaceutical industry being responsible), and commenting on and sharing misinformation from other users that the person sharing it believes to be true (Figure 2a; a TikToker expresses anger at the alleged call by Fridays for Future to ban pets over 10kg). We included one video containing factual information (Figure 2f; a presenter reports on the sabotage of railway cables). The same videos were used in Steps 1 and 2 with the exception of Video Weather (Figure 2c; a TikToker comments on an alleged manipulative attempt by media to overstate global warming) which was omitted in Step 2, due to time constraints and because the video was found to be too complex for the age group to understand the content in Step 1.

Indicators. In accordance with other research [62], we differentiate the following layers of short-videos on VSPs: spoken words, video, sound and music, layout, content and message, profile, reactions and interactions, and the overall impression. Characteristics within these layers might represent valuable indicators for misinformation, e.g., a sound that evokes strong feelings. To systematically identify a comprehensive selection of indicators for misinformation on TikTok, we (1) used these layers as reference. We (2) conducted a literature review on misinformation in VSPs to identify indicators within the layers, examining which characteristics have already been captured and/or scientifically investigated as potential indicators of misinformation in videos (see Section 2.3). To compensate for the gap of scientific publications, we (3) consulted non-scientific official media literacy manuals [13, 44, 80] and fact-checking websites [24] that specifically address TikTok and its particular features. This was (4) supplemented by literature on mainly text-based indicators that may also be suitable for video-based platforms (e.g., attention-grabbing layout) and have been evaluated in the context of misinformation and other related concepts like propaganda [58], conspiracy theories [37, 89], or political extremism [36]. An overview of indicators integrated in our approach can be found in Table 3, App.

4 RESULTS

In Section 4.1, we report how adolescents assess the comprehensibility and usefulness of indicators as cues for video-based misinformation on TikTok derived from Step 1. Building on those findings, we developed a Figma prototype of a smartphone app in Step 2 to evaluate how indicators can be applied as part of a digital misinformation intervention (see Section 4.2). Thereby, we gained insights into how adolescents perceive opportunities for and limitations of the proposed indicator-based approach (see Section 4.3). We present the results as themes obtained from our thematic analysis. Our core contributions and findings are summarized in Table 1.

[2] https://github.com/openai/whisper

4.1 Assessment of Indicators for Video-based Misinformation

For all indicators derived in Section 3.6, we evaluated their comprehensibility and perceived usefulness through focus group discussions in Step 1.

4.1.1 Video. Adolescents are familiar with AI-generated *manipulated images* (see Figure 2b), enabling them to identify indicative details and features such as inconsistent lighting conditions. While the information is considered easily understandable, its usefulness is considered limited due to perceived obviousness

> *"I think you can see it anyway, because the photos that are just generated, there is just such an artificial intelligence [...]. There were also photos with the Pope where he was wearing MontClaire jackets and things like that. And these are pictures like that."* (focus group 2)

Similarly, *using filters* (e.g., on voice or face) are a common feature on TikTok and well-known. Opinions on the necessity of acknowledging filter usage vary: *"Many older people are just fooled. I know it myself, I can give examples, but anyway, I think it makes sense, it's good."* (focus group 1) Although participants were able to explain the meaning of *facial expressions and body language evoking strong feelings* (see Figure 2a), its usefulness to assess a TikTok's credibility received mixed feedback. *"So I don't think it's very useful, because you can be calm, you can be cool, but you can still lie. "* (focus group 2) When presented the rather complex indicator of the *origin or time stamp of a video being outdated*, participants struggled with the understanding of its meaning regarding potential misleading content. However, this indicator was included in Video Weather (see Figure 2c), which participants struggled to understand overall due to its complexity. Hence, this might have strongly affected the comprehensibility of the indicator in this context.

4.1.2 Spoken words. Among the young participants, understanding how *tone of voice can evoke strong feelings* and can signal misinformation was present: *"So he raised his voice a little bit to make you think that he was scared himself. I think he wanted that to be more believable."* (focus group 2) Nonetheless, as with facial expressions, participants rated this as not necessarily a useful cue since truth on an emotional topic can be told with a raised voice, too.

4.1.3 Sound & Music. Interestingly, while evoking strong feelings through facial expressions or a speaker's voice received mixed feedback, *emotional sound or music* were deemed comprehensible and useful indicators. Participants successfully applied the indicator to both funny / entertaining music, identifying satirical videos, and sad music, identifying an intention to mislead. Regarding sound, the *outdated origin of sound* received positive feedback with participants correctly grasping its role as an indicator:

> *"I also find it quite helpful that when you hear gunshots or something in a video. And then the video is not about gunshots, I think it is right that you know that it is not part of the video. For example, if I'm watching a video about the Paris riots and suddenly I hear a bomb explosion, that doesn't fit in either."* (focus group 2)

4.1.4 Layout. Regarding layout as an indicator for misinformation, *attention-grabbing layout through color, capitalization, emoticons* (Figure 2d) is viewed as readily understandable indicator of unseriousness, up to the creator's discretion. *Spelling or grammatical errors* on the other hand, also suggesting being nonserious, are dependent on creators' language skills, if not from official sources. These are not necessarily evaluated as related to the credibility of the content, but by some participants considered relevant:

> *"Well, I think it comes across as really nonserious when you don't have punctuation and a comma and nothing at all in that whole sentence. I think a serious source would never want to just run off in a sentence like that, so fast that everybody gets kind of upset about it."* (focus group 2)

More critically, two out of three focus groups explicitly addressed the issue of dyslexia: *"So, rather useful [...] but that could be that someone with dyslexia wrote this."* (focus group 3)

4.1.5 Content & Message. Some potential indicators refer to specific characteristics of a video's modality. Others, regarding content and message, are applicable to all types of information. For example, any social media post containing *conspiracy theories* (e.g. using hashtags like #simulationtheory) has the potential to mislead. For TikTok videos, our participants were able to explain what this indicator means and rated its usefulness rather positively. Similarly, stating an *opinion without sources* was assessed as comprehensible and useful indicator. It is important to note that adolescents are particularly familiar with this type of indicator as it is a common way of teaching how to assess information in schools.

4.1.6 Profile. A hint to the profile of a creator was perceived as useful means to assess if a video might contain misinformation and was actively applied after watching the videos. Particularly, participants applied this measure to check if a video was meant to be satirical, originating from a comedy account. However, they were not aware of satirical content's potential for misdirection (e.g., hiding radical right-wing messages through satire). Overall, themes, questionable content patterns, and the creator's apparent expertise were reviewed by examining the profile, with familiarity fostering trust in known sources (e.g., official news profiles) and skepticism toward unfamiliar ones.

4.1.7 Reactions & Interactions. Reactions and interactions, including if *the content is questioned or refuted in the comment section* (see Figure 2e), are pivotal for adolescents when assessing video credibility. This strategy, has been highlighted in previous research [40] and was reaffirmed in our study. Participants proactively examined comments to assess overall sentiment and to identify potential misinformation or satire. They viewed comments critically, particularly when other users used irony and did not explicitly voice their disagreement: *"They don't really explain it to him, they just make him look stupid."* (focus group 3) Conversely, *deactivation of comments* was considered a useful indicator of misinformation: *"Where it's so blatant or something, and the comments are off. Then people usually don't accept any criticism at all. And then you can't see in the comments if it's true or not. [...] I always find the comments disabled to be such a red flag."* (focus group 2)

4.1.8 Overall Impression. Encouraging critical reflection of a video's credibility is advocated by media literacy resources and applies to

VSPs as well. Unlike objective indicators, this is subjective and based on individual reflection. The idea of this individual indicator to exhibit it as a default nudge: "Listen to yourself. What does your gut say?" and it yielded differing participant responses. Whether a video *evoked strong positive or negative feelings* was rated as rather not useful, especially because the theme of evoking feelings had already been addressed more specifically regarding body language and tonality, and was perceived somewhat redundant. Nonetheless, participants rated this nudge to self-reflection positively when it was phrased more generically as *gut feeling: Overall, the story is hard to believe*: *"I think you can tell right away that it's a little bit nonserious, and that automatically makes you think that the person might be talking garbage."* (focus group 1)

4.1.9 Autonomous Assessment Strategies. In the vast majority of cases in Step 1 (86%), participants stated that they had not previously known the videos which they were shown. Nonetheless, they often correctly assessed their truthful content before the video-specific indicators were discussed. In the best case (regarding Video Clouds, see 2d), 17 out of 21 participants correctly identified a video as containing misinformation, the remaining four at least identified it correctly in tendency ("rather not true"). In other cases (regarding Video Weather, see 2c), however, the assessment was more difficult: only nine participants tended to assess the content correctly, six tended to assess the content incorrectly and an additional six made no statement.

Despite these varying results, it is evident that the participants are, to some extent, capable of making their own correct assessments. In our study, we found that they did so by applying several assessment strategies autonomously. The adolescents were very successful in recognizing if a video was meant to be *humorous or ironic* and emphasized that potentially harmless satire and irony may be misleading as well, when not identified as such. Further, some participants particularly recognized sarcasm or irony as means to mislead:

> *"The irony, as if it is completely clear that all the animals are being brought in and so it somehow shows that she believes all the crap and thinks that she is the enlightened one who knows everything and is so much smarter than us because she knows all the conspiracy theories."* (focus group 2 regarding Video Pet, see Figure 2a)

While this type of indicator was originally not included in our study, we included it in Step 2 as part of 'Content & Message' as it was mentioned autonomously multiple times (see Table 3). Moreover, participants applied a strategy of *comparison with existing knowledge*, where they contrasted the given information with what they had already known about the topic, for instance, based on news reports or common sense. Similarly, they frequently displayed attempts to autonomously *identify reasoning or alternative logical explanations*, as in: *"So rather not, because where should these pictures come from, because in prison you're not allowed to have mobile phones. That's why no."* (focus group 1 regarding Video Trump, see Figure 2b).

Besides, participants autonomously paid attention to the content of the *video description*, in particular with regard to the hashtags therein or whether any sources were provided. They also rated the *number of likes and followers*, as well as the *general appearance* of a person in the video as a sign of the reliability. Conversely, participants displayed a variety of strategies to recognize credible information, ranging from the naming of sources and provision of further video evidence, to their general assessment of an account's appearance. Particularly the notoriety or the presence of a verification badge were perceived as useful indicators.

4.2 Applicability of the Indicator-based Approach

We evaluated how the indicators from Step 1 can be applied to a smartphone app as a digital misinformation intervention. We incorporated feedback from Step 1 to refine the indicators in Step 2, where we optimized their arrangement based on the input from adolescents, resulting in a more streamlined and less redundant set of indicators (see Table 3 for details). Specifically, an indicator related to humorous or ironic content was introduced based on this feedback, and indicators (e.g., related to strong emotions or wrong context) were merged to eliminate redundancy. These refined indicators then informed the development and evaluation of a TikTok-based smartphone app prototype aimed at mitigating digital misinformation. We derived findings regarding the usability and suitability of our indicator-based approach in Step 2.

4.2.1 Usability of the Misinformation Intervention. **SUS and qualitative results on perceived usefulness.** Our smartphone app as misinformation intervention yielded a positive average SUS score of 81.7. This favorable outcome aligns with the qualitative feedback on the usefulness and handling of the app. One of the most frequent positive comments was that the use of the familiar share button and the swiping within the results display is intuitive and very straightforward. The recording of screen touches also confirms this: our participants were able to independently navigate from the ForYou page to the Misinfo-app and within it right from the start. There were critical complaints about occasional loading problems of the app, which can be attributed to Figma's limited possibilities for prototype display with a lot of video material an overlays. In addition, several suggestions for improvement were made, which are presented below.

Barriers of usage and mitigation. We collected insights into what hinders adolescents from using the evaluated Misinfo-app and found that factors strongly differ between individuals. While the majority supported the use of the share button and the individual and detailed display of the indicators, three out of 18 participants found it too cumbersome and wished for a more compact display or a direct integration as a plugin for TikTok to reduce the *effort*: *"Maybe by not having to share it so much and then it just being there, for example, by me being able to press anything on TikTok and then it being displayed that way."* (P16, female, aged 13-14). Moreover, two criticized the app for being too *complex* for some users, as the feedback texts might be overwhelming for users who have a short concentration span, dislike reading or are illiterate: *"But for people like me who don't like to read, it's just rather more convenient to have: 'This is fake. Period.' [...] Young peoples' patience is also very short. [...] Or even my [anonymized], when she downloads the app, she can't read."* (P03, male, aged 15-16). This highlights the diverse experiences of young people in their immediate environment and their ability to empathize with potential other users. Conversely, two participants stressed that, for many videos on their ForYou page, checking for credibility is just *not relevant* because it mainly consists of entertainment videos, and they are rarely confronted with more serious content. Besides, especially the detailed feedback on single indicators using mainly text-based explanations was criticized by a small number of participants and might be a relevant factor when applied on a larger sample of adolescents. Some participants independently proposed a potential mitigation strategy using *personalization.*

4.2.2 Proposed user groups. We asked our young participants why or why not they would like to use the Misinfo-App on a regular basis and who else they would consider as potential users.

Adolescents as users. Participants liked the idea of the app having a means to transparently check for misinformation in occasions where they are uncertain about the veracity of the content. They especially emphasized positive social effects of being well informed: *"So I don't go to school and say, 'Hey, have you heard this?' And it's not even true."* (P12, male, aged 13-14) or *"I would use it. Because I didn't go to school once because of some weather thing [...] on Tiktok and that was a lie."* (P09, male, aged 13-14). Moreover, it became clear that receiving feedback from the app would not be relevant for the majority of videos, but rather for specific cases: *"So an app like this would actually be helpful. I don't see these videos too often, but when I do, I always research them and want to know if they're true or not."* (P09, male, aged 13-14)

Other users and other platforms. We were additionally interested in whom they thought would benefit from the app and found that our participants thought that older or very young people would benefit the most. Through personal experiences, the adolescents exhibited a heightened awareness of older adults' susceptibility to misinformation. Several participants stated, for example, that they take on the role of fact-checkers for older relatives in their families, especially regarding video-based content. These aspects were brought up in both the individual think-aloud sessions and in the focus group discussions: *"Yes, definitely. I would recommend it to my grandma, my little sister. Those are the people that should be using it."* (P02, female, aged 15-16) or *"Yes, especially with my grandparents, who very often, very, very often fall for fake news."* (P03, male, aged 15-16) or

> *"I can imagine that for Facebook because there are a lot of older people. For example, when the earthquake happened in Turkey, there was a lot of fake news, my [anonymized] told me that this earthquake was made by America and the country is about to be invaded. [...] As an example, I've seen it, but a lot of people believe it, I know it.* (Focus group 1)

The emphasis on relevance, especially for older individuals, also suggests the potential applicability of the indicator-based approach to other platforms, such as Facebook, where older people spend more time on social media.

Responsibility and confidence. The high responsibility of adolescents as fact checkers, e.g., for older relatives, regarding VSPs is certainly understandable and often justified based on the accumulated experiences and intuitive upbringing with these technologies. However, our results regarding how participants rated the truthfulness of TikTok videos also show limitations of those capabilities. For example, the adolescents initially made incorrect judgments in some cases without the support of the app. Six out of 18 participants assessed the truthfulness of at least one of the videos incorrectly before being exposed to the Misinfo-App results. This is not surprising, as VSP's misinformation can be misleading in multiple and often intricate ways. Some level of overconfidence when evaluating one's own credibility assessment skills is very common and reflected in our study.

4.3 Perceived Chances and Limitations of the Indicator-based Approach

Our results of Step 2 revealed several insights into how adolescents perceive benefits and limitations of the proposed indicator-based intervention for VSPs. Below, we outline our main findings.

4.3.1 Expected learning and applicability. The goal of our qualitative study design was to collect in-depth insights into adolescents' perceptions and needs. The effectiveness of digital misinformation interventions, e.g., to reduce sharing of social media posts, has been investigated in other research [9]. In contrast, our findings provide further qualitative outlooks on expected learning effects or applicability of knowledge and behavior learned when introducing an indicator-based intervention.

Applying indicators. We observed how participants quickly adapted the displayed indicators to novel TikTok videos: *"Okay, now that I read through this [the indicators], I realize that his arguments don't make any sense. So no sources either, just babbling something."* (P01, male, aged 15-16). In the further course of the study with the same participant, it became apparent that this indicator is now autonomously added for critical reflection: *"Above all, the cases are sources, I've already thought of that in my head, so it's just kind of written in random text, without any sense or anything, without sources or at least anything. So that's just obviously fake."* (P01, male, aged 15-16).

Changes in credibility assessment. Moreover, two aspects could be identified with regard to changes in participants' assessment of information credibility. On the one hand, there were several instances in which participants changed or consolidated their assessment of the information presented with the help of the indicators in Step 2. Referring to the attention-grabbing layout as an indicator, for example, one participant commented as follows:

> *"I didn't understand why the colors and such have something to do with it, but now when I think about it, that everything is written so big and especially in red, because red is such a... Red is such a sign of... I don't know how to explain it, for something important. And until now I never noticed it before, but now it makes sense somehow."* (P07, female, aged 13-14).

On the other hand, participants also found the indicators useful for practice and as a reminder to be more able to identify disinformation on their own in the future, even without app support: *"I would say that these tips are needed. So that in the future, even without this app, or so that even without this app, people can really identify for themselves that it's fake news."* (P05, male, aged 13-14)

4.3.2 *Transparency.*

The proposed approach of the Misinfo-App is based on the concept of transparently giving insights via user-centered indicators. We were interested in the adolescents' perceptions and (mis)conceptions of this approach and gained in-depth findings.

(Blind) trust in the algorithm. Out of 18 participants in Step 2, twelve explicitly stated to fully trust in the (simulated) algorithm's decision on whether a TikTok video is misleading or not. When asked for reasons, some participants expressed a general belief in computers being more intelligent or reliable than humans:

> *"Yes, so of course we trust because we are on the smartphone a lot and it has everything to do with computers. And that's why I trust it because computers are usually smarter than people. So of course I trust, I use my cell phone 24/7 and it has everything to do with a computer. And that's why, yes, of course."* (P07, female, aged 13-14)

While misconceptions of algorithms and AI-systems in particular have already been studied (e.g., revealing the *automation bias* [87]), our participants' trust in the application is also influenced by how stimuli and detection results were presented. To focus on our main goals and adhere to strict time constraints, we decided to not intentionally include false positives or false negatives, which would inevitably arise in a realistic environment of a detection app from time to time. The provision of indicators was an important element in building trust and giving adolescents a sense of autonomy in their assessments, demonstrating the great potential of our indicator-based approach: *"And there are also reasons. If you just say, yeah, it's fake, you wouldn't believe it so quickly because there are no reasons and I need reasons to know whether it's fake or not."* (P11, female, aged 13-14) Indeed, many adolescents preferred independent thinking and viewed the app as a supplementary aid: *"But for me, I would also rely more on myself and just search it again to see if I really know. After a while I would trust the app more."* (P03, male, aged 15-16)

Envisioning consequences of mistakes. As a related concept to trust in the (simulated) algorithm, we gained insight into how participants were able to envision the consequences of the Misinfo-App making mistakes. Again, this was biased by our study design which did not include mistakes of the app. Nevertheless, we can derive a general impression that our adolescent participants were overall aware of the app, algorithms, or more specifically AI systems making mistakes: *"Yes, every AI makes a few mistakes. ChatGPT has also made a few mistakes."* (P03, male, aged 15-16) In contrast, four out of 18 participants were sure that such systems would never make mistakes but would always function perfectly and a lot more precise than humans. When brainstorming about possible consequences of the app making errors, all 18 adolescents were confident that it would not have severe consequences. We then informed our participants about the possibility of the system to make incorrect decisions and the resulting consequences. All in all, this again demonstrates the advantage of the indicator-based approach over purely algorithmic detection, where errors would be even more serious.

Indicators versus binary feedback. In general, providing indicators as comprehensible and useful explanations why a video contains misinformation has been received largely positively and seems to encourage trust and knowledge applicability to other content or platforms. Indeed, when asked more specifically, the majority of participants liked the approach and considered it as a key component of the digital intervention. Their overall impression can be summarized as: *"I would like to know why it is Fake News, that's why I would have some tips about it."* (P10, male, aged 13-14) Only one participant did not like the effort and text-heavy presentation and argued for binary labeling as fake or not, or at least for an option to toggle additional information.

5 DISCUSSION

In this study, we evaluated how comprehensible and useful indicators for video-based misinformation can be applied to a smartphone app as a digital intervention to assist adolescents. In a twofold study design, we first conducted a mixed-method user study including focus group discussions with adolescents and individual surveys to examine existing strategies in dealing with video-based misinformation, complementing existing research [40], and to gain novel insights into the comprehensibility and perceived usefulness of indicators derived from literature. Building on those findings, we then developed an indicator-based smartphone app prototype, and evaluated the applicability, chances and limitations of this approach in individual think-aloud sessions with adolescents. This allowed for the examination of in-depth findings on adolescents' needs and expectations. In addition, we derived overarching design implications for user-centered, transparent digital misinformation interventions for adolescents from the following detailed responses to the research questions (see Table 2).

5.1 RQ1: How do adolescents evaluate misinformation indicators in terms of comprehensibility and usefulness?

We found that our young participants were very capable of comprehending various types of indicators regarding different components of a TikTok video. They assessed the video itself, the tonality of spoken words, emotional sound and music of a video, the layout, content and message, the profile, reactions and interactions with a video, and the overall impression which is mainly based on the adolescents' individual gut feeling (design implication (1), Table 2). It was noteworthy how adeptly the adolescents were able to identify satirical content by referencing the comment section, the creator's profile, or by relating the video to the music accompanying it. As individuals who have grown up with advances in AI, our participants were very intuitive and natural in handling and recognizing manipulated images. Indeed, they were often able to autonomously name characteristics of manipulation such as video exposure or wrong details which they used as heuristics to assess a video's credibility. These findings support the notion that adolescents and young adults, as digital natives, have a certain level of digital literacy and ability to use information competently, as suggested by some studies [30, 90].

Table 1: Core contributions and findings of the focus group discussions (Step 1) and think-aloud sessions (Step 2).

Core Contributions	Core Findings
C1: Evaluate comprehensibility and perceived usefulness of video-based indicators	F1: Indicators on diverse levels (the video itself, layout, content and message, profiles, reactions and interactions, and overall impression) expand and confirm the perceptions of adolescents F2: Particularly characteristics of manipulated videos are used intuitively as indicators by adolescents
C2: Implementation and qualitative evaluation of a smartphone app prototype as indicator-based misinformation intervention	F3: Overall positive feedback on the indicator-based application and good understanding of its features F4: Potentials for transferability of the extended knowledge and skills
C3: Assessment of chances, challenges, and limitations of the indicator-based approach	F5: Transparency of the indicators for misinformation are a central reason for the intervention's positive assessment F6: The indicator-based approach comes with limitations like adolescents blindly trusting in the simulated algorithm and a lack of realistic concerns about the robustness of the intervention

Table 2: Design implications for user-centered countermeasures targeting adolescents in dealing with misinformation.

Design Implication	Explanation
(1) Promote self-reflection	Encourage users to critically evaluate content based on their own gut instincts and indicators. Encourage self-reflection by including prompts such as "What does your gut say?" to motivate users to assess their feelings in terms of a stimuli's credibility
(2) Use comprehensive, transparent misinformation indicators	Design interventions that provide comprehensive indicators across different components of video content (sound, layout, interactions, etc.) to allow for a well-founded assessment
(3) Allow for personalization	Allow users to customize the intervention according to their preferences to allow for individual circumstances, abilities and needs. This includes users who might be overwhelmed by text-heavy explanations and who require simple visual cues as well
(4) Mitigate overconfidence	Recognize that young users may be overconfident in their ability to identify misinformation and design interventions to promote self-awareness and humility when in content evaluation
(5) Consider user trust and clarify algorithm limitations	Mitigate blind trust in algorithms by establishing a clear understanding of an algorithm's limitations and potential errors. Encourage users to use the indicators as a tool to develop their own skills in evaluating information, thereby promoting a sense of autonomy

Our findings extend and confirm existing research on trust heuristics regarding VSPs content that have been clustered as "convenience, aesthetics, and tone" by Hassoun et al. [40]. Further, similarly to the findings of previous research [40], our participants particularly emphasized a high relevance of familiar creators for credibility assessments, and in general revealed some socially-oriented concerns on what would happen if they did not recognize misinformation and shared it among their peers, reporting individual experiences. The potential of using source reliability as a key indicator for evaluating online information was also demonstrated in a randomized controlled trial, where an intervention aimed at enhancing this ability improved the students' capacity to distinguish between reliable and less reliable sources [69]. Thus, it is not surprising that our participants regularly opened the comments section of TikTok videos before deciding if the content is misleading or not.

While existing adolescent evaluation strategies in relation to the content of VSPs have been studied qualitatively [40], we build on these findings by supporting the need for shared sense-making through indicators in the domain of 'reactions & interactions'. However, we stressed that these indicators should always be used in conjunction but with others that extend the view on characteristics which are detached from other users' assessments and potential manipulations.

5.2 RQ2: How can indicators be applied to a smartphone app as a digital misinformation intervention to assist adolescents on TikTok?

Building on those findings, we developed a prototype smartphone app as an indicator-based misinformation intervention and gained in-depth insights in individual user study sessions. Overall, the comprehensive approach of integrating misinformation indicators across different components of a video was found useful by our participants (design implication (2), Table 2). The majority of participants saw value in the approach for themselves or other TikTok users of their age, especially because of the transparent indicators and the freedom of choice that this preserves. In contrast to more technical approaches that use machine learning methods to automatically detect misinformation [e.g., 82, 94], this approach enables transparency and allows for self-deliberation, and may be a suitable extension to follow after automatic detection via machine learning. The positive assessment of the adolescents is consistent with other research, suggesting that users generally prefer comprehensible explanations to binary labelling approaches [51]. In this context, comprehensibility is thought to play an important role in building trust [16]. The positive assessment of this approach aiming at improving media literacy is consistent with other studies that display the potential of media literacy as a means of combating misinformation on social media [84, 93]. However, when aiming to improve media literacy, it is important to remember that this relies on the rationality of the individual [17]. Particularly for younger people, "online information processing is fundamentally a social practice" [40], which can potentially conflict with rational analysis of information accuracy.

Furthermore, our approach is not entirely immune to the problem that young people may feel patronized by interventions aiming at enhancing media literacy and that their digital skills are not taken into account - an issue that has been raised in previous research [40]. In our findings, this is most evident in the additional user groups suggested. For example, some of our participants see older or very young people as users who could particularly benefit from the approach, rather than themselves because of their already developed skills in identifying misinformation. The varying skill levels observed in adolescents, alongside their reported different preferences (e.g., regarding the amount of text-heavy explanations of the indicators) also suggest that incorporating personalization into digital interventions may alleviate such concerns (design implication (3), Table 2) – an approach that has shown promise in other areas of critical information handling [3, 29].

Moreover, our results displayed the balancing act that young people face in their role as judges of misinformation, especially with regard to video-based content. On the one hand, because they are 'digital natives', they are called upon as experts, especially by older relatives, to clarify the trustworthiness of videos. Our qualitative insights regarding the evaluation of TikTok videos indeed reflect a certain competence, especially with regard to manipulated videos. On the other hand, our study shows that in some cases even young people who are confident in their abilities are not able to recognize misinformation on their own and may benefit from (technical) support due to the complex and multifaceted misleading potential of the videos. The concept of 'overconfidence' in assessing media content may therefore also apply to adolescents in our context, as they repeatedly stated that they saw potential in such an indicator-based app, especially for other people (design implication (4), Table 2). Such overconfidence and general overestimation of digital skills in adolescents and young adults has been explored in other studies [19, 63, 70, 96] and is partly explained by a lack of critical thinking [68]. Furthermore, the level of self-efficacy in adolescents was suggested as a crucial factor for both checking sources and refraining from sharing unreliable information [67]. Overall, our study indicates significant potential for mutual learning by co-designing studies with young people, which can also benefit other user groups such as older adults.

5.3 RQ3: How do adolescents perceive chances and limitations of an indicator-based digital misinformation intervention?

Our qualitative findings reveal opportunities for an indicator-based approach to provide useful support in dealing with misinformation. However, they also reveal limitations and challenges. For example, our approach shows potential for an automation bias that encourages blind trust in (e.g., AI-based) algorithms [87]. Only few of our participants showed healthy skepticism about the capabilities and limitations of the technical tool. In particular, it was noted that there was little understanding of the serious consequences of an incorrect output from the application. While this potential for blind trust in automation has been suggested [87], other studies have also shown skepticism towards automated systems [4, 52]. Thus, the clear lack of skepticism in the sample of adolescents in our study suggests a potentially high level of vulnerability (design implication (5), Table 2). This makes it all the more important to use transparent approaches when helping adolescents to deal with misinformation, rather than relying solely on automatic detection with subsequent binary labelling [e.g., 11]. An indicator-based approach can, to some extent, counteract this risk by further challenging the participants' own thinking. Our qualitative findings offer initial insights into potential opportunities for the applicability of the indicators to new videos and the transferability of knowledge and skills to platforms beyond TikTok.

5.4 Limitations and Potential for Future Work

Our study has some limitations and potential for future work. *First*, quantitative evaluation of the effectiveness of the indicator-based approach in real-world usage is still needed to investigate if there is a long-term learning effect of applying indicators and if this approach prevents misinformation sharing. It is not the aim of our study to identify statistically significant effects, such as reducing the sharing of misinformation. This has been addressed in other studies [9, 20] and remains worthy of investigation in the context of video-based content for future research. It will be interesting to monitor whether the integration of the app via the share button is accepted by users in everyday life compared to a direct integration within TikTok, and how the approach is perceived in a practical usage scenario where the app's output is occasionally incorrect, and the app's consultation is useful for a limited number of videos. *Second*, evaluating the approach with a larger sample of videos might provide additional insights and could be combined with the suggested quantitative study design for future research. *Third,* our prototype is based on a simulation regarding the detection of misinformation and the detection of the individual indicators. Digital misinformation interventions often rely on a successful prior decision (automatic or manual) as to whether a piece of information contains misinformation. Prefiltering whether a social media post contains misinformation is a non-trivial task and a problem that has been the subject of research in recent years [82, 94], especially in the field of machine learning. Displaying our indicators by default on all videos would lead to an excessive number of false positives and is therefore not appropriate. Future work could evaluate the approach fully implemented with state-of-the-art detection techniques for pre-filtering. In our user studies, we observed a group of young people with different levels of learning abilities, ranging from very young adolescents (13 years old) to 16 year-old adolescents. By including an entire school class in combination with participants from a youth center, biases towards a special interest in technology could be well balanced and the sample size is also well in line with the usual measures of qualitative studies [21]. Nevertheless, it would be interesting for future studies to - *fourth* - include a larger sample, e.g., to expand the age range of the subgroups. The use of TikTok is restricted to individuals over the age of 13, but there is no defined upper limit. People in older age groups are discovering TikTok for themselves, wich makes them a particularly interesting group of participants for studies.

6 CONCLUSION

Through our mixed-methods user study with 39 adolescents, we qualitatively evaluated the potential and limitations of an indicator-based misinformation intervention for TikTok as a core VSP.

We advance misinformation research by applying existing knowledge about indicators and controversially discussed media literacy interventions to the very current modality of short-videos. Thereby, we adopt the user perspective of adolescents as a particularly relevant user group. We generate novel insights into the opportunities, challenges, and limitations of indicator-based misinformation interventions. Our core contributions and findings are (C1) evaluating the comprehensibility and perceived usefulness of video-based indicators to (F1) find how indicators on diverse levels (e.g., the video itself, profiles, and interactions) expand and confirm the perceptions of adolescents, and (F2) highlight how the identification of characteristics of manipulated videos is an intuitive approach pursued by adolescents. Further, we (C2) implemented and evaluated a smartphone app prototype as indicator-based misinformation and received (F3) an overall positive feedback on the approach and a substantive understanding of its features, and identified (F4) potentials of transferability to new content. Lastly, we (C3) assessed opportunities, challenges, and limitations of the indicator-based approach and found that (F5) it was particularly well received in terms of transparency, while it (F6) carries limitations, such as missing realistic concerns about the robustness of the intervention. On a social level, our findings underscore the balancing act of adolescents being regarded as digital experts to evaluate social media content, but also being vulnerable to the complex landscape of (video-based) misinformation.

ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry for Education and Research (BMBF) in the project NEBULA (13N16361) and in the project CYLENCE (13N16636), and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We sincerely thank the Haus der digitalen Medienbildung, the Institut für Medienpädagogik und Kommunikation Hessen e.V., and the Jugendzentrum Martinsviertel, along with the teachers and students whose guidance and participation played a crucial role in enhancing the quality of our study.

REFERENCES

[1] Malik Almaliki. 2019. Online Misinformation Spread: A Systematic Literature Map. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining (ICISDM 2019)*. Association for Computing Machinery, New York, NY, USA, 171–178. https://doi.org/10.1145/3325917.3325938

[2] Nadia Alonso-López. 2021. Beyond Challenges and Viral Dance Moves: TikTok as a Vehicle for Disinformation and Fact-Checking in Spain, Portugal, Brazil, and the USA. *Analisi: Quaderns de Comunicacio i Cultura* 64, 1 (2021), 65–84.

[3] Filipe Altoe and H. Sofia Pinto. 2023. Towards a Personalized Online Fake News Taxonomy. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Limassol Cyprus, 96–105. https://doi.org/10.1145/3565472.3592963

[4] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence. *AI & SOCIETY* 35, 3 (Sept. 2020), 611–623. https://doi.org/10.1007/s00146-019-00931-w

[5] A Ardevol-Abreu, P Delponti, and C Rodriguez-Wanguemert. 2020. Intentional or Inadvertent Fake News Sharing? Fact-checking Warnings and Users' Interaction with Social Media Content. *Profesional de la Informacion* 29, 5 (2020), 1–13. https://doi.org/10.3145/epi.2020.sep.07

[6] Patricia Aufderheide. 1992. *A Report of the National Leadership Conference on Media Literacy.* Technical Report. The Aspen Institute Wye Center, Queenstown. 9–16 pages.

[7] Jackie Ayoub, X. Jessie Yang, and Feng Zhou. 2021. Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models. *Information Processing & Management* 58, 4 (July 2021), 102569. https://doi.org/10.1016/j.ipm.2021.102569

[8] Bence Bago, David G. Rand, and Gordon Pennycook. 2020. Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines. *Journal of Experimental Psychology: General* 149, 8 (Aug. 2020), 1608–1613. https://doi.org/10.1037/xge0000729

[9] Joseph B. Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Emma S. Spiro, Kate Starbird, and Jevin D. West. 2022. Combining Interventions to Reduce the Spread of Viral Misinformation. *Nature Human Behaviour* 6, 10 (Oct. 2022), 1372–+. https://doi.org/10.1038/s41562-022-01388-6

[10] Ava Bartolome and Shuo Niu. 2023. A Literature Review of Video-Sharing Platform Research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581107

[11] Ranojoy Barua, Rajdeep Maity, Dipankar Minj, Tarang Barua, and Ashish Kumar Layek. 2019. F-NAD: An Application for Fake News Article Detection Using Machine Learning Techniques. In *2019 IEEE Bombay Section Signature Conference (IBSSC)*. Institute of Electrical and Electronics Engineers, Mumbai, 1–6. https://doi.org/10.1109/IBSSC47189.2019.8973059

[12] Corey H. Basch, Grace C. Hillyer, and Christie Jaime. 2022. COVID-19 on TikTok: Harnessing an Emerging Social Media Platform to Convey Important Public Health Messages. *International Journal of Adolescent Medicine and Health* 34, 5 (Oct. 2022), 367–369. https://doi.org/10.1515/ijamh-2020-0111

[13] Federico Battaglia, Denis Gross, Hanna Herweg, and Eva Kappl. 2023. *Cats, War and Creators: TikTok as Place for (Dis)Information for Teenagers (Translated from German).* Technical Report. Amadeu Antonio Stiftung, Berlin.

[14] Puneet Bhargava, Katie MacDonald, Christie Newton, Hause Lin, and Gordon Pennycook. 2023. How Effective Are TikTok Misinformation Debunking Videos? *Harvard Kennedy School Misinformation Review* 4, 2 (March 2023), 1–17. https://doi.org/10.37016/mr-2020-114

[15] Prasanta Bhattacharya, Raj Kumar Gupta, and Yinping Yang. 2023. Exploring the Contextual Factors Affecting Multimodal Emotion Recognition in Videos. *IEEE Transactions on Affective Computing* 14, 2 (April 2023), 1547–1557. https://doi.org/10.1109/TAFFC.2021.3071503

[16] MD Momen Bhuiyan, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. Designing Transparency Cues in Online News Platforms to Promote Trust: Journalists' &amp; Consumers' Perspectives. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 395:1–395:31. https://doi.org/10.1145/3479539

[17] Danah Boyd. 2017. Did Media Literacy Backfire? *Journal of Applied Youth Studies* 1, 4 (2017), 83–89.

[18] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, Pranee Liamputtong (Ed.). Springer, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[19] Monica Bulger and Patrick Davison. 2018. The Promises, Challenges, and Futures of Media Literacy. *Journal of Media Literacy Education* 10, 1 (2018), 1–21. https://doi.org/10.23860/JMLE-2018-10-1-1

[20] Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. YouTube Recommendations and Effects on Sharing Across Online Social Platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 11:1–11:26. https://doi.org/10.1145/3449085

[21] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 981–992. https://doi.org/10.1145/2858036.2858498

[22] Valerio Capraro and Tatiana Celadin. 2022. "I Think This News Is Accurate": Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News. *Personality and Social Psychology Bulletin* 0, 0 (2022), 1–11. https://doi.org/10.1177/01461672221117691

[23] Sijing Chen, Lu Xiao, and Akit Kumar. 2022. Spread of Misinformation on Social Media: What Contributes to It and How to Combat It. *Computers in Human Behavior* 141 (Dec. 2022), 107643. https://doi.org/10.1016/j.chb.2022.107643

[24] Correctiv.org. 2023. Wie erkenne ich Falschmeldungen?

[25] Jonas De Keersmaecker and Arne Roets. 2017. 'Fake News': Incorrect, but Hard to Correct. The Role of Cognitive Ability on the Impact of False Information on Social Impressions. *Intelligence* 65 (Nov. 2017), 107–110. https://doi.org/10.1016/j.intell.2017.10.005

[26] Lindsay Levkoff Diamond, Hande Batan, Jennings Anderson, and Leysia Palen. 2022. The Polyvocality of Online COVID-19 Vaccine Narratives That Invoke Medical Racism. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–21. https://doi.org/10.1145/3491102.3501892

[27] Elena-Alexandra Dumitru. 2020. Testing Children and Adolescents' Ability to Identify Fake News: A Combined Design of Quasi-Experiment and Group Discussions. *Societies* 10, 3 (Sept. 2020), 71. https://doi.org/10.3390/soc10030071

[28] Jared Duval, Ferran Altarriba Bertran, Siying Chen, Melissa Chu, Divya Subramonian, Austin Wang, Geoffrey Xiang, Sri Kurniawan, and Katherine Isbister. 2021. Chasing Play on TikTok from Populations with Disabilities to Inspire Playful and Inclusive Technology Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445303

[29] Serge Egelman and Eyal Peer. 2015. The Myth of the Average User: Improving Privacy and Security Systems through Individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*. ACM, Twente, 16–28.

[30] K J Kevin Feng, Kevin Song, Kejing Li, Marshini Chetty, and Oishee Chakrabarti. 2022. Investigating How University Students in the United States Encounter and Deal With Misinformation in Private WhatsApp Chats During COVID-19. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, Vol. 18. USENIX Association, Boston Massachusetts USA, 427–446.

[31] Diana Freed, Natalie N. Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. 2023. Understanding Digital-Safety Experiences of Youth in the U.S.. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3581128

[32] Tomoya Furuta and Yu Suzuki. 2021. A Fact-checking Assistant System for Textual Documents. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, Tokyo, 243–246. https://doi.org/10.1109/MIPR51284.2021.00046

[33] Sukeshini Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. 2021. By the Crowd and for the Crowd: Perceived Utility and Willingness to Contribute to Trustworthiness Indicators on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (July 2021), 1–24. https://doi.org/10.1145/3463930

[34] Xinning Gui, Yubo Kou, Kathleen H. Pine, and Yunan Chen. 2017. Managing Uncertainty: Using Social Media for Risk Assessment during a Public Health Crisis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 4520–4533. https://doi.org/10.1145/3025453.3025891

[35] Lee Hadlington, Lydia J. Harkin, Daria Kuss, Kristina Newman, and Francesca C. Ryding. 2023. Perceptions of Fake News, Misinformation, and Disinformation amid the COVID-19 Pandemic: A Qualitative Exploration. *Psychology of Popular Media* 12, 1 (Jan. 2023), 40–49. https://doi.org/10.1037/ppm0000387

[36] Loni Hagen, Mary Falling, Oleksandr Lisnichenko, AbdelRahim A. Elmadany, Pankti Mehta, Muhammad Abdul-Mageed, Justin Costakis, and Thomas E. Keller. 2019. Emoji Use in Twitter White Nationalism Communication. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. ACM, Austin TX USA, 201–205. https://doi.org/10.1145/3311957.3359495

[37] Sven Ove Hansson. 2017. Science Denial as a Form of Pseudoscience. *Studies in History and Philosophy of Science Part A* 63 (June 2017), 39–47. https://doi.org/10.1016/j.shpsa.2017.05.002

[38] Katrin Hartwig, Frederic Doell, and Christian Reuter. 2023. The Landscape of User-centered Misinformation Interventions – A Systematic Literature Review. https://doi.org/10.48550/arXiv.2301.06517 arXiv:2301.06517 [cs]

[39] Katrin Hartwig and Christian Reuter. 2019. TrustyTweet: An Indicator-Based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter. In *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*, Vol. 14. AIS, Siegen, Germany, 1844–1855.

[40] Amelia Hassoun, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M. Russell. 2023. Practicing Information Sensibility: How Gen Z Engages with Online Information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3581328

[41] Paula Herrero-Diz, Jesús Conde-Jiménez, and Salvador Reyes de Cózar. 2020. Teens' Motivations to Spread Fake News on WhatsApp. *Social Media + Society* 6, 3 (July 2020), 2056305120942879. https://doi.org/10.1177/2056305120942879

[42] Paula Herrero-Diz, Jesús Conde-Jiménez, and Salvador Reyes-de-Cózar. 2021. Spanish Adolescents and Fake News: Level of Awareness and Credibility of Information (Los Adolescentes Españoles Frente a Las Fake News: Nivel de Conciencia y Credibilidad de La Información). *Culture and Education* 33, 1 (Jan. 2021), 1–27. https://doi.org/10.1080/11356405.2020.1859739

[43] https://www.facebook.com/cbckidsca. 2023. WATCH — How to Spot What's Not Real on TikTok CBC Kids News. https://www.cbc.ca/kidsnews/post/watch-how-to-spot-whats-not-real-on-tiktok/.

[44] Institut für Medienpädagogik und Kommunikation and Wissenschaftsstadt Darmstadt. 2023. *Media Check: What Smartphone and Internet Can Mean for You (Tranlated from German)*. Technical Report. Haus der digitalen Medienbildung (HddM).

[45] Se-Hoon Jeong, Hyunyi Cho, and Yoori Hwang. 2012. Media Literacy Interventions: A Meta-Analytic Review. *Journal of Communication* 62, 3 (2012), 454–472. https://doi.org/10.1111/j.1460-2466.2012.01643.x

[46] Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Bias Misperceived:The Role of Partisanship and Misinformation in YouTube Comment Moderation. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 278–289. https://doi.org/10.1609/icwsm.v13i01.3229

[47] Hyerim Jo, Fan Yang, and Qing Yan. 2022. Spreaders vs Victims: The Nuanced Relationship between Age and Misinformation via FoMO and Digital Literacy in Different Cultures. *New Media & Society* 11, 1 (Nov. 2022), 146144482211304. https://doi.org/10.1177/14614448221130476

[48] Joseph Kahne and Benjamin Bowyer. 2017. Educating for Democracy in a Partisan Age: Confronting the Challenges of Motivated Reasoning and Misinformation. *American Educational Research Journal* 54, 1 (Feb. 2017), 3–34. https://doi.org/10.3102/0002831216679817

[49] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. A Systematic Literature Review on Disinformation: Toward a Unified Taxonomical Framework. *New Media & Society* 23, 5 (May 2021), 1301–1326. https://doi.org/10.1177/1461444820959296

[50] Kulvinder Kaur and Pawan Kumar. 2022. Social Media: A Blessing or a Curse? Voice of Owners in the Beauty and Wellness Industry. *The TQM Journal* 34, 5 (Nov. 2022), 1039–1056. https://doi.org/10.1108/TQM-03-2021-0074

[51] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27. https://doi.org/10.1145/3415211

[52] Markus Langer, Cornelius J. König, Caroline Back, and Victoria Hemsing. 2023. Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias. *Journal of Business and Psychology* 38, 3 (June 2023), 493–508. https://doi.org/10.1007/s10869-022-09829-9

[53] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Chapter 8 - Interviews and Focus Groups. In *Research Methods in Human Computer Interaction (Second Edition)*, Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser (Eds.). Morgan Kaufmann, Boston, 187–228. https://doi.org/10.1016/B978-0-12-805390-4.00008-X

[54] Clayton Lewis. 1982. *Using the Thinking-Aloud Method in Cognitive Interface Design*. IBM TJ Watson Research Center, NY.

[55] Chen Ling, Krishna P. Gummadi, and Savvas Zannettou. 2023. "Learn the Facts about COVID-19": Analyzing the Use of Warning Labels on TikTok Videos. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (June 2023), 554–565. https://doi.org/10.1609/icwsm.v17i1.22168

[56] Eugène Loos, Loredana Ivan, and Donald Leu. 2018. "Save the Pacific Northwest Tree Octopus": A Hoax Revisited. Or: How Vulnerable Are School Children to Fake News? *Information and Learning Science* 119, 9/10 (Oct. 2018), 514–528. https://doi.org/10.1108/ILS-04-2018-0031

[57] C Martel, M Mosleh, and DG Rand. 2021. You're Definitely Wrong, Maybe: Correction Style Has Minimal Effect on Corrections of Misinformation Online. *Media and Communication* 9, 1 (2021), 120–133. https://doi.org/10.17645/mac.v9i1.3519

[58] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A System to Support the Analysis of Propaganda Techniques in the News. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 287–293. arXiv:2005.05854

[59] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580695

[60] Ashlee Milton, Leah Ajmani, Michael Ann DeVito, and Stevie Chancellor. 2023. "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing*

Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3581489

[61] Muhammad Naeem and Ozuem Wilson. 2022. Understanding Misinformation and Rumors That Generated Panic Buying as a Social Practice during COVID-19 Pandemic: Evidence from Twitter, YouTube and Focus Group Interviews. *Information Technology & People* 35, 7 (2022), 2140–2166.

[62] Shuo Niu, Zhicong Lu, Amy X. Zhang, Jie Cai, Carla F. Griggio, and Hendrik Heuer. 2023. Building Credibility, Trust, and Safety on Video-Sharing Platforms. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3573809

[63] Thomas Nygren and Mona Guath. 2019. Swedish Teenagers' Difficulties and Abilities to Determine Digital News Credibility. *Nordicom Review* 40, 1 (Feb. 2019), 23–42. https://doi.org/10.2478/nor-2019-0002

[64] Brendan Nyhan and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32, 2 (June 2010), 303–330. https://doi.org/10.1007/s11109-010-9112-2

[65] Niall J. O'Sullivan, Greg Nason, Rustom P. Manecksha, and Fardod O'Kelly. 2022. The Unintentional Spread of Misinformation on 'TikTok'; A Paediatric Urological Perspective. *Journal of Pediatric Urology* 18, 3 (June 2022), 371–375. https://doi.org/10.1016/j.jpurol.2022.03.001

[66] Katherine O'Toole. 2023. Collaborative Creativity in TikTok Music Duets. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581380

[67] Marinella Paciello, Giuseppe Corbelli, and Francesca D'Errico. 2023. The Role of Self-Efficacy Beliefs in Dealing with Misinformation among Adolescents. *Frontiers in Psychology* 14 (2023), 1–7.

[68] Concetta Papapicco, Isabella Lamanna, and Francesca D'Errico. 2022. Adolescents' Vulnerability to Fake News and to Racial Hoaxes: A Qualitative Analysis on Italian Sample. *Multimodal Technologies and Interaction* 6, 3 (March 2022), 20. https://doi.org/10.3390/mti6030020

[69] Ana Pérez, Anna Potocki, Marc Stadtler, Mônica Macedo-Rouet, Johanna Paul, Ladislao Salmerón, and Jean-François Rouet. 2018. Fostering Teenagers' Assessment of Information Reliability: Effects of a Classroom Intervention Focused on Critical Source Dimensions. *Learning and Instruction* 58 (Dec. 2018), 53–64. https://doi.org/10.1016/j.learninstruc.2018.04.006

[70] Erez Porat, Ina Blau, and Azy Barak. 2018. Measuring Digital Literacies: Junior High-School Students' Perceived Competencies versus Actual Performance. *Computers & Education* 126 (Nov. 2018), 23–36. https://doi.org/10.1016/j.compedu.2018.06.030

[71] Cynthia Putnam, Melisa Puthenmadom, Marjorie Ann Cuerdo, Wanshu Wang, and Nathaniel Paul. 2020. Adaptation of the System Usability Scale for User Testing with Children. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–7. https://doi.org/10.1145/3334480.3382840

[72] Thomas Reinhold and Christian Reuter. 2023. Zur Debatte über die Einhegung eines Cyberwars: Analyse militärischer Cyberaktivitäten im Krieg Russlands gegen die Ukraine. *Zeitschrift für Friedens- und Konfliktforschung* 12, 1 (April 2023), 135–149. https://doi.org/10.1007/s42597-023-00094-y

[73] Hilda Ruokolainen and Gunilla Widén. 2020. Conceptualising Misinformation in the Context of Asylum Seekers. *Information Processing & Management* 57, 3 (May 2020), 102127. https://doi.org/10.1016/j.ipm.2019.102127

[74] Joni Salminen, Soon-Gyo Jung, Shammur Chowdhury, Sercan Sengün, and Bernard J. Jansen. 2020. Personas and Analytics: A Comparative User Study of Efficiency and Effectiveness for a User Identification Task. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376770

[75] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. Misinformation Interventions Are Common, Divisive, and Poorly Understood. *Harvard Kennedy School Misinformation Review* 2, 5 (Oct. 2021), 1–25. https://doi.org/10.37016/mr-2020-81

[76] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3411763.3451807

[77] Carmen Sanchez and David Dunning. 2018. Overconfidence among Beginners: Is a Little Learning a Dangerous Thing? *Journal of Personality and Social Psychology* 114, 1 (Jan. 2018), 10–28. https://doi.org/10.1037/pspa0000102

[78] Anastasia Schaadhardt, Yue Fu, Cory Gennari Pratt, and Wanda Pratt. 2023. "Laughing so I Don't Cry": How TikTok Users Employ Humor and Compassion to Connect around Psychiatric Hospitalization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3544548.3581559

[79] Stefka Schmid, Katrin Hartwig, Robert Cieslinski, and Christian Reuter. 2022. Digital Resilience in Dealing with Misinformation on Social Media during COVID-19 A Web Application to Assist Users in Crises. *Information Systems Frontiers* 0, 0 (2022), 1–23. https://doi.org/10.1007/s10796-022-10347-5

[80] Wolfgang Schweiger, Sarah Heinisch, Sophie Kitzmann, Jennifer Madelmond, Marco Herbst, and Constantin Schell. 2020. *Informationskompetenz: Erkennen, Was Wahr Und Richtig Ist.* Technical Report. Landesmedienzentrum Baden-Württemberg, Stuttgart.

[81] Imani N. Sherman, Jack W. Stokes, and Elissa M. Redmiles. 2021. Designing Media Provenance Indicators to Combat Fake Media. In *24th International Symposium on Research in Attacks, Intrusions and Defenses*. ACM, San Sebastian Spain, 324–339. https://doi.org/10.1145/3471621.3471860

[82] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[83] H. Colleen Sinclair. 2020. 10 Ways to Spot Online Misinformation. http://theconversation.com/10-ways-to-spot-online-misinformation-132246.

[84] Francesca Spezzano. 2021. Using Service-Learning in Graduate Curriculum to Address Teenagers' Vulnerability to Web Misinformation. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 2*. ACM, Virtual Event Germany, 637–638. https://doi.org/10.1145/3456565.3460039

[85] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 1–26. https://doi.org/10.1145/3359229

[86] Kate Starbird, Dharma Dailey, Owla Mohamed, Gina Lee, and Emma S. Spiro. 2018. Engage Early, Correct More: How Journalists Participate in False Rumors Online during Crisis Events. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173679

[87] Stefan Strauß. 2021. Deep Automation Bias: How to Tackle a Wicked Problem of AI? *Big Data and Cognitive Computing* 5, 2 (June 2021), 18. https://doi.org/10.3390/bdcc5020018

[88] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's Common and a Part of Being a Content Creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501879

[89] Jan-Willem van Prooijen and Karen M. Douglas. 2018. Belief in Conspiracy Theories: Basic Principles of an Emerging Research Domain. *European Journal of Social Psychology* 48, 7 (Dec. 2018), 897–908. https://doi.org/10.1002/ejsp.2530

[90] Ingrid Volkmer. 2022. *Social Media and COVID-19: A Global Study of Digital Crisis Interaction among Gen Z and Millennials.* Technical Report. University of Melbourne, Melbourne, Australia.

[91] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine* 240 (Nov. 2019), 112552. https://doi.org/10.1016/j.socscimed.2019.112552

[92] Theresa Webb and Kathryn Martin. 2012. Evaluation of a Us School-Based Media Literacy Violence Prevention Curriculum on Changes in Knowledge and Critical Thinking Among Adolescents. *Journal of Children and Media* 6, 4 (Nov. 2012), 430–449. https://doi.org/10.1080/17482798.2012.724591

[93] Don Winiecki, Francesca Spezzano, and Chandler Underwood. 2023. Understanding Teenagers' Real and Fake News Sharing on Social Media. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. ACM, Chicago IL USA, 598–602. https://doi.org/10.1145/3585088.3593864

[94] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, Marina Del Rey CA USA, 637–645. https://doi.org/10.1145/3159652.3159677

[95] Alex J. Xu, Jacob Taylor, Tian Gao, Rada Mihalcea, Veronica Perez-Rosas, and Stacy Loeb. 2021. TikTok and Prostate Cancer: Misinformation and Quality of Information Using Validated Questionnaires. *BJU International* 128, 4 (Oct. 2021), 435–437. https://doi.org/10.1111/bju.15403

[96] Luisa Dolores Zozaya-Durazo, Charo Sádaba-Chalezquer, and Beatriz Feijoo-Fernández. 2023. "Fake or Not, I'm Sharing It": Teen Perception about Disinformation in Social Networks. *Young Consumers* 0, 0 (Jan. 2023), 1–18. https://doi.org/10.1108/YC-06-2022-1552

A INDICATORS FOR MISINFORMATION

Table 3: Indicators for Misinformation as derived from literature and represented in Step 1 and 2.

Category	**Indicator/Clue**		**Examples**	**Sources**
	Representation in Step 1	*Representation in Step 2*		
Spoken words	Tone of voice of the speaker evokes strong feelings like anger or fear	Strong emotions: Through tone of voice, facial expression, or body language, the person shows strong feelings or evokes strong feelings in you (e.g., anger or fear).	e.g., shouting or crying	[15, 24, 83]
video (= any motion or static pictures displayed in front or background)	Facial expression / Body language evokes strong feelings		e.g., angry or sad face	[15, 24, 83]
	Filter is misleading	Edited images: Details in the video look strange (e.g., lighting). Or, a filter in the video (such as a distorted face or voice) makes the video look different than it is.	e.g., appearance or voice sounds like child but is adult	[13, 31, 43, 50]
	Manipulated Images		e.g., something seems off, inaccurate representation of fingers	[13, 31, 43, 50]
	Origin / Time stamp of video is outdated or in wrong context	Wrong context: Outdated or mismatched sound.	e.g., video of military parade is displayed with description regarding ongoing war	[13, 43, 80]
Sound / Music	Origin of sound is outdated or in wrong context		e.g., gun shots are played but do not belong to the video	[13, 43]
	Sound / Music evokes strong feelings	Music or sound: The sound or music in the background evokes strong feelings (e.g., anger, fear).	e.g., emotional balades	[40]
Layout	Attention grabbing through color / capitalization / emoticons etc.	Attention grabbing layout: Colors, capitalization or emoticons are used to attract attention. It looks unserious.	e.g., red bold letters	[24, 36, 89]
	Difficult to take seriously due to spelling / grammatical errors	Hard to take layout seriously. E.g. spelling mistakes or grammatical errors (e.g. in the video description) look unserious.	e.g., wrong punctuation or spelling	[24, 40, 83]
Content / Message	In the video / picture / text / audio conspiracy theories are named	Conspiracy Theories: This is a common conspiracy theory that has been disproved by many fact checkers.	e.g., #simulationtheory	[37]
	Opinion only without a serious source	Missing sources: Only an opinion is named without giving serious sources.	/	[40, 44, 80, 83]
	(not yet included)	Humorous or ironic presentation: The video is not serious or making fun of things.	/	[83]
Profile	Strange or suspicious interests / beliefs in profile description / name / picture	Suspicious profile: The profile description, name, or profile picture seem strange or suspicious (e.g., lots of hateful or conspiratorial videos).	e.g., lots of hateful or conspiratorial videos	[13, 40]
Reactions / Interactions	There is a lot of questioning or refuting in comments	Reactions to the video: The video is challenged or refuted in the comments. Or: Comments on the video are disabled. This prevents other people from disagreeing with the video.	/	[13, 40]
	The comments on the video are disabled		/	[46]
Overall impression	Strong negative or positive feelings are evoked	Gut feeling: Is the story overall hard to believe? Do the arguments make sense? Does it fit with what you know about the subject? Do you know and trust the person?	/	[24, 40, 83]
	Gut feeling says the story is hard to believe overall		/	[83]

B STUDY PROCEDURES AND ITEMS APPLIED IN STEP 1 AND 2.

B.1 Detailed Study Procedure of Step 1

- Introduction, informed consent and collection of demographic information. Start of audio recording.
- (Repeated for all five videos:) Showing the TikTok video. Participants were always given the opportunity to watch the video again or to view additional information, such as the creator's profile, the video description, or comments from the video.
  - Individual paper survey on general questions about the video:
    - Do you know this video or a similar video on this topic? (yes, no, other)
    - Do you believe the content of the video is true or false?" (not at all true / not very true / somewhat true / very true)
    - How exactly do you come to that conclusion? What did you pay attention to? (free-text format)
  - If the video contained misinformation:
    - Participants were then given the following statement by the researcher *"Imagine that an algorithm (i.e., a computer program) has determined that the information is likely to be misleading or false. Also, an algorithm has automatically found the following cues. These are meant to help you understand that the content of the video is not true. We will now go through them one by one."*
    - Group discussions about how they understood each cue and whether they found it useful. This clarified whether the participants had actually understood the cue correctly.
  - If the video contained only true information (one of the five videos):
    - Participants were then given the following statement by the researcher *"Imagine that an algorithm (i.e., a computer program) has determined that the information is likely to be true."*
  - The researcher gives a final clarification about the truthfulness of the video.
- After following the procedure for all five videos, we presented three additional indicators that were not included in the selected videos but still seemed promising to evaluate. Again, the participants rated their comprehensibility and usefulness in group discussions.
- Final questions were asked orally to the group of participants:
  - What did you think of the cues overall?
  - Was something missing?
  - Do you have any additional comments or questions?
- Audio recording is stopped.

B.2 Detailed Study Procedure of Step 2

- Introduction and informed consent. Start of audio and screen recording.
- Demonstration of the think-aloud procedure
- The participant was handed an Android smartphone with the TikTok simulation open and was verbally given the following cue: *"Please imagine you are on TikTok and this is your ForYou page. You can click or swipe anywhere, you can't break anything."*
- Watching the first video for as long as they wanted and opening the comments, video description, or profile if they wanted to.
  - The participant was asked to share the video with the Misinfo-App and have it checked to see if it was true or false. The participant then had as much time as they wanted to look at the result of the Misinfo-App.
  - The participant then returned to the simulated ForYou page and could swipe to the next video.
- Repeating for all four videos. Then leaving the Misinfo-App and TikTok simulation and changing to an online survey:
  - System Usability Scale (SUS)
  - Demographic questions
- Verbally asking additional questions for in-depth insights:
  - Do you think the Misinfo-App can help you or others (e.g. younger siblings or an older neighbor) to spot misinformation or not? Why (not)?
  - Would you use the Misinfo-App? Why (why not)?
  - How would you change the Misinfo-App so that you or others would want to use it?
  - The Misinfo-App aims to automatically recognizes when information is false. It also gives you tips on how you could spot it yourself. Do you need these tips at all, or is it enough just to say: This video is fake?
  - The detection if a video is misinformation has been done by an algorithm (i.e., a computer program) within the Misinfo-App. Do you trust the Misinfo-App? Why (not)?
  - Can the Misinfo-App make mistakes?
  - What would happen if the Misinfo-App made mistakes?
- Recording is stopped.

B.3 Adapted System Usability Scale (SUS)

The SUS was adapted to young people aged 13 to 16 following Putnam et al. [71].

- I think that I would like to use the Misinfo-App frequently.
- I was confused may times when using the Misinfo-App.
- I thought the Misinfo-App was easy to use.
- I think I need help from my parents or siblings to use the Misinfo-App.
- I always felt like I knew what to do next when using the Misinfo-App.
- Some of the things I had to do in the Misinfo-App did not make sense.
- I would imagine that most people my age would learn to use the Misinfo-App very quickly
- I felt the Misinfo-App was cumbersome to use.
- I felt very confident using the Misinfo-App.
- I needed to learn a lot of things before I could get going with the Misinfo-App.

B.4 Coding Scheme and Stimuli for Steps 1 and 2

Table 4: Coding scheme of Step 1 and 2 including codes and themes at different levels.

Themes Level 1	Themes Level 2	Codes Level 1	Codes Level 2
Indicators / Strategies	Literature-based indicators	Perceived usefulness of Comprehensibility of Limitations of	Video
			Sound & music
			Reactions & interactions
			Profile
			Content & message
			Layout
			Spoken words
			Overall impression
	Autonomous assessment strategies	Indicators for misinformation	
		Indicators for credible videos	
Applicability of the indicator-based approach	Usability of the misinformation intervention	Perceived usefulness	
		Barriers of usage	Effort
			Complexity
			Not relevant
		Personalization as mitigation	
	Proposed user groups	Adolescents	
		Other users	
		Other platforms	
		Responsibility and confidence	
Chances and Limitations of the indicator-based approach	Expected learning and applicability	Applying indicators	
		Changes in credibility assessment	
	Transparency	(Blindly) trusting in the algorithm	
		Imagining consequences of mistakes	
		Indicators versus binary feedback	

Table 5: Detailed description of videos used as stimuli in Steps 1 and 2.

	Video Pets	Video Trump	Video Weather	Video Clouds	Video Trains
Description	Misinformation: A TikToker expresses anger and frustration at the alleged call by the climate movement Fridays for Future to ban and euthanize pets that weigh over 10kg. The face of the Tiktoker is very close to the camera and superimposed on the video, allowing distinctive facial expressions to be easily recognized.	Deep fake: In the background there are images created by means of deep fake that supposedly show ex-US President Donald Trump in prison. In the foreground, a TikToker is shaking their head to ironically express their disbelief while rap music is playing.	Misinformation: A TikToker comments on the comparison of two weather maps, with one map in shades of green and the other in shades of red for identical temperatures. He sees this as an alleged manipulative attempt by the media to overstate global warming. Dramatic music is playing.	Satire: The TikToker makes fun of common conspiracy theories by pointing in the video to clouds that block out the sun and have an allegedly unusual shape. In this he recognizes an alleged conspiracy of the pharmaceutical and solarium industry. Bright graphics and inscriptions are also added to the video.	True news: A presenter reports on the sabotage of railway cables on the TikTok channel 'Funk', which is operated by two German public broadcasters. An offscreen voiceover reads out a statement by the Minister of the Interior on the subject.
Link	Has been deleted	Has been deleted	Link Weather [3]	Link Clouds [4]	Link Train [5]
Characteristics					
Evokes strong feelings (voice, face, body)	●			●	
Manipulated images		●			
Wrong origin or context			●		
Evokes strong feelings (music, sound)		●	●		
Attention-grabbing layout	●			●	
Dubious layout	●				
Conspiracy theory				●	
Opinion, missing source	●		●	●	
Humorous/ ironic/ sarcastic	●	●	●	●	
Suspicious profile	●			●	
Comments/ reactions			●	●	
Gut feeling: strong emotions	●	●	●	●	
Gut feeling: hard to believe	●	●	●	●	

[3] anonymized due to privacy of creator

[4] https://www.tiktok.com/@inspecta.wack/video/7216618291282840838?is_from_webapp=1&sender_device=pc&web_id=7244433874452071962

[5] https://www.tiktok.com/@funk/video/7161386786265124101?is_from_webapp=1&sender_device=pc&web_id=7244433874452071962