Reducing the Impact of Deepfakes

Niklas Lübbeling (s2151448)
University of Twente

07.07.2022

Abstract

Deepfakes are a form of machine learning and artificial intelligence enabling the visual and auditory manipulation of video material which is mostly used to transpose individuals faces on those of other people. Deepfakes can cause reputational damage, distrust or even hatred towards the people being deepfaked, or can be used for political propaganda leading to polarization among those believing in these videos. Due to this, this study tested the effectiveness of an intervention including inoculation against deepfakes to reduce the spread of deepfakes. Students between 18-29 years of age (*N* = 152) were randomly allocated to one of two conditions receiving either neutral information about deepfakes or negative information about the consequences of deepfakes. It was expected that students in the Inoculation present condition showed lower sharing intentions and that this effect on intentions was mediated by students' attitudes and their perceived coping abilities. However, none of these expected relationships were supported. Apart from this, the influence of social media trust on the expected relationships was investigated. Yet, social media trust did not affect the other relationships. Moreover, different dimensions investigating why people share videos were tested. Findings show that students' sharing intentions could be predicted by their ratings of funniness of the individual deepfakes. Although this study was not able to reduce people's sharing intentions, it did reveal crucial relationships between the individual variables which are in line with prior literature. Furthermore, important implications for further studies investigating the potential of inoculation on reducing the impact of deepfakes were inferred.

Keywords: deepfakes, psychological inoculation, boosting, sharing intentions, social media

Reducing the Impact of Deepfakes

The Internet offers many possibilities for information being shared quickly. Such a fast distribution of messages can have many advantages such as improved communication (Krimsky, 2007). However, by making information easily accessible to a large number of people, also false information can be quickly spread (Burkhardt, 2017). This is because, especially on social media, people can distribute information without being controlled or fact-checked by experts (Pennycook & Rand, 2021).

More recently, a new form of false information became prevalent, the so-called deepfakes. Deepfakes are a form of machine learning and artificial intelligence which enables the visual and auditory manipulation of video material (Kietzmann et al., 2020). With this form of technology, for example, individuals' faces can be transposed on faces of other people and their voices can be imitated so that they ostensibly make statements that they would never make in reality (Kietzmann et al., 2020). Hence, people can be placed in a context in which they are made to act or behave in ways they would not behave by themselves.

There are many problems connected to this new form of false information. Firstly, deepfakes can cause significant damage to the person displayed on the manipulated video footage (Lyu, 2020). Deepfakes attracted public attention when in late 2017 faces of celebrities were transposed onto pornographic videos (Lyu, 2020). Such misusage of these technologies creates reputational damage to the person involved and could therefore be used as a form of bullying others. Moreover, by faking statements of politicians, deepfakes are also abused as a means for political sabotage and propaganda (Maras & Alexandrou, 2018) and hence, may lead to international conflicts (Caporusso, 2020). Additionally, the rapid advances in deepfake methods make deepfakes more realistic and less distinguishable from real video footage. This in turn can lead to distrust or even hatred towards the people being deepfaked and can cause polarization among those believing in these videos (Yadlin-Segal & Oppenheim, 2021). Due to this, it is important that the impact of deepfakes is reduced by raising people's awareness towards this form of false information and that the spread of deepfakes is limited.

In order to reduce the impact of deepfakes, an intervention would be needed which helps people to assess a video's believability, to reject deepfakes and lead them to not share deepfakes when encountering them on social media. Due to the rapidly evolving and developing deepfake methods, current digital interventions to detect manipulated video footage are not enough to counteract deepfakes (Pennycook & Rand, 2021; Yang et al., 2020).

This is because the quickly changing possibilities for creating deepfakes require deepfake detection programs to constantly be revised as well. If, however, people could be made aware of the negative consequences of deepfakes, they may resist deepfakes themselves and not spread them further among social media. This may have a longer lasting effect and reduce the impact of deepfakes more efficiently. Due to this, this study tries to investigate whether an intervention could successfully increase people's awareness about deepfakes and help them understand the potentially negative consequences of deepfakes.

However, not only awareness needs to be raised, but also people's intention to share and hence, spread deepfakes among social media needs to decrease. The problem with people's sharing intentions is that they are independent from people's accuracy judgements regarding the content of the respective message (Pennycook & Rand, 2021). In other words, people may still share false information even when being aware of it being false or misleading. Some people may do so deliberately to create chaos (Petersen et al., 2018) or because they would find such information interesting if it was true (Altay et al., 2021). Other people, however, may do so because they simply did not attend to the accuracy of the information (Pennycook & Rand, 2021). Linked with this, Pennycook et al. (2020) found that when simply reminding people to assess the information's accuracy that they are better able to discern real from false information and that this improves people's sharing decisions. Hence, a study which raises awareness about the negative consequences about the distribution of deepfakes may lead people to avoid spreading deepfakes among social media.

For this study, students were chosen as the target group. The reason for this is that deepfakes are mostly shared on social media (Kwok & Koh, 2021). This increased quantity of fake videos on social media is due to the mere accessibility of deepfake software which can be used by everyone without extensive technical skills (Kietzmann, 2020). Thereby, students from European universities were found to trust social media more than traditional media and use social media as their main news source (Bantimaroudis et al., 2020; Statista, 2021). Moreover, statistics show that people in the ages from 18-29 most frequently use social media platforms (Khoros, 2021). This means that students between the age of 18-29 may be more likely to be exposed to deepfakes among social media and may believe these deepfakes to reflect reality. Hence, an intervention aimed at raising awareness among students between the age of 18-29 will be tested. The research question of this paper is: "How can the awareness of students between 18 and 29 years of age be increased regarding the negative consequences of sharing deepfakes?"

Theoretical Framework

Belief in deepfakes. There are different factors influencing why people believe such manipulated footage. According to Reichelt (2021), the believability of deepfakes is dependent upon five categories. The first two categories are the technical and auditory qualities of the deepfake which help to create the illusion that the manipulated footage is real. However, as methods of making deepfakes improve rapidly, flaws in these categories are likely corrected over time (Langguth et al., 2021). Hence, regarding these categories, deepfake detection programs will be necessary to detect manipulations on video footages.

Apart from these more technological aspects, there are also factors on the side of the person consuming the deepfakes influencing whether deepfakes are perceived as believable. Generally, people trust the voices they know and are more likely to believe videos than photos or images (Kietzmann et al., 2020). Thereby, believability increases when the content of the deepfakes is in line with what the audience would be expecting from the person displayed on the deepfake (Reichelt, 2021). In other words, the source of the deepfake as well as prior experience or knowledge of the receiver are important for whether deepfakes are believed or not (Reichelt, 2021). Kietzmann (2020), however, proposes that people may still trust what they perceive with their eyes, and hence, they may trust deepfakes, even if the content may be perceived as unlikely. As mentioned above, people may even share false information despite knowing it to be false or even misleading (Pennycook & Rand, 2021). Therefore, behaviour change on the side of the consumer is necessary to reduce the impact of deepfakes.

Behaviour change. To reduce the impact of deepfakes, students' sharing intentions need to be decreased. They should not only learn about deepfakes generally, but also about their potential of misusage and negative consequences for the persons displayed. According to the theory of planned behaviour (TPB), for behaviour to change, people need to intend or be motivated to do so (Ajzen, 1985). Such intentions, in turn, depend on people's attitudes, perceived behaviour control and subjective norms. In other words, if people have positive attitudes towards a certain behaviour, think that the behaviour is socially desirable and feel capable to apply such behaviour, their motivation to change their behaviour increases (Ajzen, 1985). If students learn about the negative consequences of deepfakes and socially undesirable misusages, they will form more negative attitudes towards deepfakes, thereby resisting deepfakes in the future and avoid sharing them on social media.

In line with this, boosting is a promising strategy which can be used to raise students' awareness about deepfakes and their negative impact. Boosting involves the fostering of specific competences and has already been successfully applied in a variety of domains

(Hertwig & Grüne-Yanoff, 2017). Such competences may be behavioural routines, motivational competences, or information about a certain topic. In this study, awareness and knowledge about deepfakes and their negative consequences would be the required competences that would have to be fostered. Hence, people's knowledge about deepfakes should be broadened or boosted.

In line with the TPB, boosts require motivational participation for behaviour change to occur. Boosts are transparent to the persons intended to be boosted and they can then decide themselves whether to apply the boost or not (Hertwig & Grüne-Yanoff, 2017). Another advantage of boosting is that the gained competences may also be applied across situations. In other words, boosting awareness about deepfakes may help students to also think more critically in general when assessing information on social media. Thus, boosting may trigger critical thinking and motivate students to resist deepfakes and lead them to avoid sharing deepfakes among social media.

As already mentioned, for behaviour change and boosting to be successful, people need to be motivated. According to protection motivation theory (PMT), whether people respond maladaptively to a threat or whether they become motivated to counteract a threat is dependent on two distinct appraisal processes (Norman et al., 2005). On the one hand, there is a threat appraisal consisting of two distinct constructs, perceived severity as well as perceived susceptibility towards the risk. On the other hand, there is a coping appraisal involving the coping strategies that the individual possesses (self-efficacy) and whether these strategies are perceived to be effective in counteracting the threat (response efficacy).

In line with this, the Extended Parallel Process Model (EPPM) predicts that people firstly need to perceive a threat as relevant for engaging in finding strategies to counteract the threat (Gore & Bracken, 2005). For risk messages to be effective then, people need to feel capable of counteracting the threat to accept the message and raise protection motivation. If people do not feel adequately capable of dealing with a risk, the EPPM states that people react maladaptively to the fear elicited by the threat. Typical maladaptive coping mechanisms in this case include denial, reactance, or avoidance, leading to a rejection of the risk message (Gore & Bracken, 2005). Hence, people need to feel capable to counteract the threat in order to react adaptively, i.e., to not sharing deepfakes.

However, not only people's cognitive appraisals of a threat decide about people's actions, but also their emotional appraisal. According to Loewenstein et al. (2001), students' risk perception is also influenced by affective responses in the form of negative emotions. In other words, if an intervention could raise people's perceived threat, fear towards the risk is elicited. This emotion should then lead people to decide about whether to approach or avoid the risk (Loewenstein et al., 2001). For this study, this means that students would have to be warned of the threat elicited by deepfakes and how easily they could be exposed or even targeted by deepfakes. However, at the same time they need to be convinced that by not sharing deepfakes, that the impact of deepfakes can be successfully reduced. By this, their cognitive and emotional threat perception should be increased while it would be ensured that students cope adaptively with the risk. This should increase their coping appraisal and hence, their protection motivation to accept the boost.

Apart from this, people’s prior knowledge can affect whether people intend to resist deepfakes and counteract them. In accordance with this, the persuasion knowledge model (PKM) predicts that people’s way of coping with a persuasive attempt is determined by three different types of knowledge (Friestad & Wright, 1994). These include individuals’ knowledge about the topic at hand, the source of the information and the persuasive intent of the source. If people know about the persuasive attempt of the source, they develop higher skepticism and more negative attitudes towards the source (Iacobucci et al., 2021). This means that when students are warned about the persuasive attempt of deepfakes to spread misinformation, they should develop more negative attitudes about the deepfakes.

Intervention models. To raise awareness about deepfakes and increase protection motivation, students may get inoculated against deepfakes. Inoculation theory involves that, similar to a medical vaccination, people are given a weaker version of false information as a means to trigger critical thinking (Banas & Miller, 2013; Van der Linden & Roozenbeek, 2020). Inoculation necessitates two essential message characteristics to be effective. Firstly, people need to be warned of an incoming persuasive attempt (Banas, & Miller, 2013). This warning message serves to boost people’s awareness about the persuasive intents of deepfakes and their severe consequences. According to the PKM, students’ attitudes towards deepfakes then should become more negative (Iacobucci et al., 2021). Following the TPB, such negative attitudes should then lead to a lower intention to share deepfakes (Ajzen, 1985). Furthermore, the warning message should increase students’ perceived threat of the severity of deepfakes. In line with the EPPM, this should lead students to attend to the risk message which in turn should heighten their protection motivation as predicted by the PMT (Norman et al., 2005). The second characteristic is the refutational preemption in which people have to create arguments to bolster their negative attitudes against the incoming threat (Banas, & Miller, 2013). This in turn should lead to lower sharing intentions among students. Thus, inoculation may be helpful in reducing the impact of deepfakes.

Inoculation theory has already been widely applied to counteract false information. For example, inoculation has been shown to be effective in combating conspiracy theories (Banas & Miller, 2013). Thus, it can be assumed that by inoculating students against deepfakes, they may become more resisting towards deepfakes and have a lower intention to share deepfakes among social media. Hence, inoculation may be able to reduce the impact of deepfakes being spread among social media. Nevertheless, research into the benefits of inoculation on reducing the impact of deepfakes is still necessary (Sankaranarayanan et al., 2021). Therefore, this study adds to prior literature by investigating to what extent inoculation theory can be applied to raise awareness about deepfakes and reduce their impact by reducing people's intention to share them among social media.

The current study

This study investigates whether inoculating students is effective in reducing the impact of deepfakes being shared on social media. Based on previous findings (Banas & Miller, 2013; Van der Linden & Roozenbeek, 2020), it was expected that inoculation could increase students' awareness about deepfakes and decrease their intent to share deepfakes among social media. Furthermore, with perceived threat being the main mechanism of the PMT and EPPM (Gore & Bracken, 2005; Norman et al., 2005), it was used as a manipulation check. An increase in both students' cognitive and emotional threat was assumed. Moreover, in line with prior inoculation studies (Banas & Miller, 2013) and the PKM (Iacobucci et al., 2021), students' attitudes towards deepfakes were expected to become more negative which in turn should reduce students' sharing intentions. Hence, attitudes towards deepfakes were investigated as a mediator for the effect of the intervention on sharing intentions. Moreover, by explaining students how to counteract deepfakes, their perceived coping abilities should increase which in turn should also reduce their sharing intentions (Gore & Bracken, 2005). Therefore, perceived coping ability was added as an additional mediator. In summary, the hypotheses for this study are the following:

H1: An intervention which inoculates students against deepfakes will reduce their intention to share deepfakes among social media.

H2: Students' attitudes towards deepfakes mediate the effect of inoculation on sharing intentions.

H3: Students' perceived coping abilities mediate the effect of inoculation on sharing intentions.

Method

Design

An experimental between-groups design was used with Inoculation (present or absent) as the independent variable. Participants were allocated to either condition. Data was collected from participants using an online survey. The main dependent variable was participants' intention to share the deepfakes among social media. Moreover, students' attitudes towards deepfakes and their perceived coping ability were measured as additional confounding variables. Perceived threat was considered as a manipulation check consisting of two separate tests. On the one hand, the cognitive side of perceived was investigated by perceived susceptibility and severity. On the other hand, the emotional side of perceived threat was measured in form of negative emotions elicited by deepfakes.

Apart from this, several variables were measured as exploratory variables to gain more insights into the underlying mechanisms for why people share deepfakes among social media. Firstly, social media trust was measured to gain more insights into the relation between social media trust and sharing intentions. Moreover, different dimensions according to which people may share videos on social media were measured to gain more information on the reasons for why videos are shared on social media more generally (see Appendix A).

Participants

The population for this study were students between the age of 18-29 years. A non-probability sampling design was chosen with the application of quota sampling. Quota sampling involves selecting a sample convenient to the researcher based upon a priori decided upon inclusion criteria (Kumar, 2018). The inclusion criteria were the age of the participants as well as the necessary occupation as a student. Otherwise, participants were referred to the end of the survey. In addition, participants were collected via the SONA system of the University of Twente. When having participated via SONA, students received credit points for the completion of the study. Otherwise, no incentives were given. In total, 204 participants took part in the study. However, 16 participants had to be removed due to not having met the inclusion criteria. Furthermore, 36 cases had to be removed due to missing data. After the removal of invalid cases, 75 participants remained in the Inoculation absent group (*N* = 75), while 77 participants (*N* = 77) remained in the Inoculation present group. The gender distribution was 28.3% male (*N* = 43), 70.4% female (*N* = 107) and 1.3% diverse (*N* = 2). The age of the participants ranged from 18 to 26 years (*M* = 20.44, *SD* = 1.73). Moreover, 54.6% of the participants were German (*N* = 83) followed by 27.6% Dutch (*N* = 42) and 17.8% other nationalities (*N* = 27).

Materials

The online survey used for this study consisted of several parts. These included an informed consent, demographic questions, the inoculation in the form of a message displayed at the beginning of the intervention, the deepfakes used to measure participants' sharing intentions, a questionnaire to measure the dependent variables, as well as a final question asking for further improvements. The study was designed and conducted with the online software tool Qualtrics.

Inoculation. Participants in both groups received a message about deepfakes at the start of the intervention. However, the context of the message differed between groups. Participants in the Inoculation absent condition merely received neutral information about deepfakes. Students in the Inoculation present group were made aware of the negative consequences of deepfakes and the different ways deepfakes can be misused. This should raise participants' perceived threat in the Inoculation present group. At the same time, students' perceived coping ability should be increased. This was achieved by raising students' self-efficacy by explaining how easily the impact of deepfakes can be reduced by limiting their spread. Additionally, response efficacy should be increased by making students aware that by not sharing deepfakes, their impact can be successfully reduced. Furthermore, the source credibility of the message, as a component of persuasion knowledge, was raised by stating that the messages were taken from a scientific article about recent findings about deepfakes. This should ensure that participants attend to the message and that their attitudes became more negative as a result. By this, students' sharing intentions should decrease. The messages for both groups can be found in Appendix B and C.

Deepfakes. The second part of the intervention consists of eight deepfakes which are spread among social media. The deepfakes were received from YouTube or other video platforms on the internet and merely included non-explicit content with relatively harmless applications of deepfakes. For example, one of the deepfakes used for this study involves an ostensible statement of Mark Zuckerberg, co-founder and chief executive of Facebook, in which he seems to declare that the data collection of his website is used to manipulate and control people (Brandalism Project, 2019). Above the videos, participants were informed that the videos involved deepfakes to ensure that students knew that the content of the videos was manipulated.

Measures

Sharing intentions. Participants' sharing intentions regarding the deepfakes displayed were captured using a 7-point Likert scale ranging from 1 (*unlikely*) to 7 (likely). Participants were asked to answer the question "Imagine that you would have encountered the following video on social media. Please indicate how likely you would have shared this video among social media". Higher scores indicated a higher sharing intentions. This question was asked for each of the eight deepfakes. The question regarding people's sharing intentions was newly formulated for this study.

Attitude towards deepfakes. Students' attitude towards deepfakes was measured with the question "Please indicate to what extent you think that deepfakes are...". This question was answered using six bipolar adjective pairs (*negative/positive, bad/good, unfavourable/favourable, unacceptable/acceptable, wrong/right, foolish/wise*) on a scale from 1 to 7. Higher scores indicated more positive attitudes towards deepfakes. This measure for attitudes has already been validated in prior inoculation studies (e.g., Banas & Miller, 2013; Pfau et al., 2009). In this study, this scale showed an excellent internal consistency with Cronbach's Alpha being $\alpha = .90$. According to Nunnally (1967, as cited in Anderson & Agarwal, 2010), an alpha of above .7 can be considered acceptable for confirmatory analyses. Furthermore, Lambda 2 was inspected as it is independent from the number of items in a scale and hence, a more robust and precise estimate of a scale's reliability (Statistics How To, 2016; Tavakol & Dennick, 2011). The Lambda 2 for this scale was $\lambda = .90$.

Perceived coping ability. Participants' perceived coping ability was measured through a combined measure of their self-efficacy and response-efficacy. Both were captured using three questions each on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher scores indicated higher self-efficacy or response-efficacy respectively. One example of a self-efficacy item was "I feel like I have the knowledge that is necessary to deal with deepfakes on my own" while an example of a response-efficacy item was "By not sharing deepfakes among social media, misusage of deepfakes can be prevented". The items for this scale were taken from De Kimpe et al. (2021) but rewritten to fit the context of deepfakes. This scale showed an acceptable internal consistency with Cronbach's Alpha being $\alpha = .75$ and Lambda 2 being $\lambda = .80$.

Perceived threat. Perceived threat was measured through a combined measure of its cognitive as well as affective components. The cognitive component included participants' perceived severity and perceived susceptibility towards deepfakes. Again, both constructs were captured using three questions each on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher scores indicated higher severity or susceptibility respectively. One example of a severity item was "Deepfakes are a serious threat." while an example of a susceptibility item was "It is likely that I will be exposed to deepfakes on social media". The items for this scale were also taken from De Kimpe et al. (2021) yet again rewritten to fit the context of deepfakes. This scale showed an internal consistency of Cronbach's Alpha being α = .81 and Lambda 2 being λ = .84.

The emotional component of perceived threat was investigated by measuring negative emotions using the question "When thinking about deepfakes, to what extent do you feel...". To answer this question students were asked to rate the applicability of seven different negative emotions (*distressed*, *angry*, *insecure*, *worried*, *irritated*, *powerless*, *anxious*) on a 5-point Likert scale ranging from 1 (*Not at all*) to 5 (*Extremely much*). These emotions were taken from the 7-Level Emotional Scale (Markalanfish, 2020) and were chosen by their fit regarding deepfakes. Although there are already established scales measuring negative affect such as the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), these items mostly did not relate to cybersecurity or deepfakes more specifically. Thus, the negative emotion scale was also created by the author of this paper. The scale demonstrated high internal consistency with Cronbach's Alpha being α = .89 and Lambda 2 being λ = .89. The complete questionnaire can be found in Appendix D.

Exploratory variables. Students' social media trust and the ratings of the different dimensions rated for each deepfake were measured as exploratory variables. Social media trust was measured using two items on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher scores indicated higher social media trust. An example of a social media trust item was "I have every confidence that trusting information being shared among social media is safe". Again, these items were already used by De Kimpe et al. (2021) and were rewritten to match the context of this study. The social media trust items had an unacceptable internal consistency of Cronbach's Alpha being α = .58 and Lambda 2 being λ = .58. Thus, it can be assumed that the two items measured different underlying concepts. For this reason, only the above stated item was kept for the analysis as this item more clearly asked participants for their trust in information shared on social media.

Apart from this, to receive more information on why people would or would not share the videos, students were asked to rate the videos on several dimensions using the question "Please rate the video among the following dimensions". This question was answered using six bipolar adjective pairs (*not funny/funny, not informative/informative, not polarizing/ polarizing, not harmful/harmful, unintimidating/intimidating, unacceptable/acceptable*) on a scale from 1 to 7. This question regarding the dimensions was newly conceptualized for this study.

Procedure

Firstly, the students filled out an informed consent in which they were given more general information that the study is about how people evaluate deepfakes shared among social media. After having consented to participate, students in both groups were asked to answer demographic questions to make sure that they met the inclusion criteria for this study. If participants did not meet the inclusion criteria, they were transferred to the end of the survey. This study was approved by the BMS Ethics Committee of the University of Twente.

Next, students were randomly assigned to one of two conditions, either the Inoculation present or Inoculation absent condition. Students in both conditions firstly received the manipulation of this study, an introduction message about deepfakes. The message for the Inoculation present condition informed students that they will be shown deepfakes and explained the problems and potentially negative consequences of deepfakes. This warning message was meant to increase their protection motivation to resist sharing deepfakes among social media. Students in the Inoculation absent condition were only given neutral information about deepfakes.

After that, both groups were presented eight different deepfakes. Each deepfake was followed by one question asking for the likelihood of students sharing each deepfake if they would have encountered the respective video on social media. Additionally, participants were asked to rate the respective videos among the different dimensions. After the intervention, students were shown a questionnaire which measured the additional dependent variables. More specifically, students' attitudes towards deepfakes were measured first. After that, the emotional component of perceived threat was assessed. Next, students' cognitive component of perceived threat and their perceived coping ability were assessed. Moreover, their social media trust was determined.

After having completed the intervention and the questionnaire, students were debriefed about the real intention of the study. It was explained that the real aim was to reduce students' sharing intentions and, in turn, limit the impact of deepfakes. They were then enabled to choose again whether their data should be used for the analysis or not. If they chose to withdraw from the study, their data was deleted. Lastly, participants in both groups were asked one open question to evaluate the study for further improvements in the future. Overall, the completion of the study took participants around 30 minutes.

Data Analysis Plan

For the data analysis, the Statistical Package for the Social Sciences (SPSS: Version 26) was used. Two separate independent samples t-tests were conducted for perceived threat, which was used as a manipulation check, one for the cognitive component and one for the emotional component of threat. To test the hypotheses, a parallel mediation model was applied with PROCESS to investigate the main effect of inoculation of sharing intentions, but also the separate mediation effects of attitudes towards deepfakes and perceived coping ability. Moreover, social media trust was included as an exploratory variable. In other words, attitudes towards deepfakes, perceived coping ability and social media trust were included in the assumed model. Before the main analyses, however, a multiple linear regression analysis was conducted to see which of the evaluation dimensions could predict the likelihood of sharing the respective deepfakes. For all tests, the significance threshold was set to $\alpha \leq .05$, with $\alpha < .1$ being considered marginally significant.

Results

Descriptives

A summary of the means and standard deviations of each scale as well as the confirmatory inter-scale correlations are displayed in Table 1. These included students' sharing intentions of deepfakes, their perceived threat, their perceived coping ability, their attitudes towards deepfakes as well as their social media trust. The means of cognitive perceived threat (*M* = 5.29, *SD* = 0.97) and perceived coping ability (*M* = 4.59, *SD* = 0.99) were above the midpoints of their scales. Emotional perceived threat (*M* = 2.55, *SD* = 0.92) was scored around the midpoint while sharing intentions (*M* = 2.84, *SD* = 1.03) and attitudes towards deepfakes (*M* = 2.56, *SD* = 0.93) were scored below the midpoints of their respective scales. This means that although participants rated the severity and susceptibility of deepfakes to be high, they also thought of themselves as being capable to deal with the threat. Moreover, they seemed to be affected only moderately by deepfakes emotionally. Apart from this, the general attitudes were rather negative towards deepfakes and sharing intentions were also rather low.

Notably, sharing intentions were weakly positively correlated to emotional perceived threat $r(150) = .27$, $p < .01$, yet nonsignificantly to participants' cognitive perceived threat $r(150) = .11$, $p > .05$. According to Schober et al. (2018), correlations below $r = .40$ can be considered weak. Thus, only if negative emotions of participants increased, they were also more likely to intend to share the deepfakes. This suggests that people's emotions towards deepfakes are more decisive about whether people eventually share deepfakes than their cognitive perceptions of threat. The other scales were not significantly related to sharing intentions. Moreover, emotional perceived threat was further significantly related to all other

Table 1

Means (M), Standard Deviations (SD) and Confirmatory Inter-scale Correlations [a]

Scales	M	SD	1	2	3	4	5
1. Sharing Intentions	2.84	1.03					
2. Perceived Threat Cognitive	5.29	0.97	.11				
3. Perceived Threat Emotional	2.55	0.92	.27**	.48**			
4. Perceived Coping Ability	4.59	0.99	-.15	-.07	-.37**		
5. Attitudes Towards Deepfakes	2.56	0.93	-.04	-.31**	-.50**	.15	
6. Social Media Trust	1.89	0.90	.08	-.18*	-.07	-.01	.16*

Note. [a] $N = 152$; *$p < .05$, ** $p < .01$, two-tailed.

scales, thereby moderately positive with cognitive perceived threat $r(150) = .48, p < .01$, weakly negative with perceived coping ability $r(150) = -.37, p < .01$, and moderately negative with attitudes towards deepfakes $r(150) = -.50, p < .01$ (Schober et al., 2018). In other words, the higher participants felt threatened emotionally by deepfakes, the higher they also perceived the threat cognitively and vice versa. Therefore, emotional and cognitive threat perceptions indeed seem to be connected rather than being independent in the case of deepfakes. Moreover, with higher emotional threat, they showed lower perceived coping abilities and more negative attitudes towards deepfakes. Hence, the more emotionally threatened participants felt, the less likely they felt able to adequately cope with deepfakes. Attitudes were also weakly negative related to cognitive perceived threat $r(150) = -.31, p < .01$ (Schober et al., 2018). Thus, both when students perceived deepfakes to be more threatening, whether cognitive or emotional, their attitudes towards deepfakes became more negative and vice versa. Apart from this, social media trust was weakly negative related to perceived cognitive threat and weakly positive related to students' attitudes. This means that the more students trust social media, the less threatening they evaluated deepfakes and more positive attitudes they showed towards deepfakes.

Manipulation checks

Two manipulation checks were conducted to assess whether the intervention was able to increase students' perceived threat as an essential component of inoculation theory. Thus, cognitive perceived threat and emotional perceived threat were each compared between groups. For this, two separate independent samples t-test were conducted.

Firstly, an independent samples t-test was conducted to investigate the differences of cognitive perceived threat between groups. An assumption check of the independent samples t-test revealed that the scores of both groups were sufficiently normally distributed. Furthermore, there was one outlier which, however, was due to natural variation and thus, was kept in order to ensure the representativeness of the data collected. Moreover, Levene's test for equality of variances was found to be non-significant, *F*(1, 150) = 0.50, *p* = .48, meaning that equality of variances could be assumed. The t-test showed that cognitive perceived threat did not differ between groups, *t*(150) = -0.42, *p* = .67. Thus, the intervention did not raise student's perceived severity and susceptibility of deepfakes.

Secondly, participants' emotional perceived threat in form of negative emotions towards deepfakes was compared between groups. Again, the assumptions of normality and no distorting outliers was met. Furthermore, Levene's test of equal variances indicated that equal variances between groups could be assumed, *F*(1, 150) = 0.02, *p* = .90. The independent samples t-test revealed that also student's emotional perceived threat did not differ between groups, *t*(150) = -0.07, *p* = .95. In other words, the intervention did also not raise students' negative emotions towards deepfakes significantly. This means that the manipulation was unsuccessful as it was not able to increase participants' perceived threat which was considered a prerequisite for inoculating students against deepfakes.

Sharing Dimensions

Before the main analyses, a multiple linear regression was conducted to investigate whether the dimensions of the separate deepfakes could predict people's sharing intentions. By this, insights should be gained about the reasons for why people would share the displayed videos at all. Again, the assumption of normality was met. However, an inspection of the correlations of the dimensions showed that harm and intimidation were highly correlated, suggesting multicollinearity. Nevertheless, the according collinearity diagnostics showed that the variance inflation factors of both dimensions were far below the threshold of 10 and hence, they were both kept for the analysis. Lastly, there was one outlier regarding sharing intentions in the Inoculation present group. Again, however, this outlier was due to natural variation and hence, kept for the analysis. Lastly, the assumption of homoscedasticity was met as well.

The regression showed that the combined effect of the dimensions could significantly predict students' sharing intentions, $R^2 = .30$, $R^2_{Adjusted} = .27$, F (6, 145) = 10.50, $p < .001$. An investigation of the individual predictors further revealed that the funniness dimension could significantly predict students' sharing intentions, $B = 0.39$, $t = 4.97$, $p < .001$. Thus, to the extent that students evaluated the videos to be funny, the more likely they intended to share the videos. Furthermore, polarization tended to predict sharing intentions, yet only marginally significant, $B = 0.17$, $t = 1.90$, $p = .06$. The other dimensions did not significantly predict sharing intentions. A summary of the effects of each predictor dimension can be found in Table 2.

Table 2

Effects of Each Dimension on Sharing Intentions

			95% CI	
Coefficient	B	t	LL	UL
Dimension Funny	0.39^{**}	4.97	0.23	0.54
Dimension Informative	0.11	0.92	-0.13	0.36
Dimension Polarizing	0.17	1.90	-0.01	0.34
Dimension Harmful	-0.15	-1.19	-0.41	0.10
Dimension Intimidating	0.17	1.58	-0.04	0.37
Dimension Acceptable	-0.01	-0.11	-0.21	0.19

Note. $^{a}N = 152$; $^{*}p < .05$, $^{**}p < .01$

In order to avoid potential misunderstandings, it was explored whether the eight videos differed in their ratings of funniness which predicted people's sharing intentions. This is because if some of the videos were not considered funny by participants, they would also not be shared by participants. If this was the case, this may potentially undermine the potential effect of the intervention on sharing intentions and some deepfakes would have to be excluded from the analyses. However, an inspection of the funniness ratings of the individual deepfakes showed that the ratings of the videos were all around 3.0. Thus, it could not be expected that there was an undermining of the effect and thus, all eight videos were kept.

Testing the hypotheses

In order to test the hypotheses, a parallel mediation model was conducted to assess the effect of inoculation on sharing intentions while also investigating the mediation effects of students' attitudes towards deepfakes and their perceived coping ability. The first hypothesis stated that the intervention should reduce students' intention to share deepfakes among social media. However, the model revealed that there was no significant direct effect of the inoculation intervention and their sharing intentions of deepfakes, $B = 0.27$, $t = 1.61$, $p = .11$. This means that the intervention was not able to decrease students sharing intentions and thus, the first hypothesis needs to be rejected.

Moreover, it was expected that the effect of the inoculation on sharing intention was mediated by students' attitudes towards deepfakes and their perceived coping abilities. However, the Inoculation did not significantly change students' attitudes towards deepfakes, *B* = -0.09, *t* = -0.61, *p* = .54. Furthermore, the effect of attitudes on sharing intentions was not significant in the model including the condition and students' perceived coping ability, *B* = -0.01, *t* = -0.09, *p* = .93. Hence, the second hypothesis was rejected.

Apart from this, the condition could not significantly predict students' perceived coping ability, *B* = 0.13, *t* = 0.84, *p* = .11. Additionally, perceived coping ability could not predict sharing intentions in the overall model, *B* = -0.17, *t* = 1.97, *p* = .051. Due to this, the third hypothesis was rejected as well. Thus, overall, the intervention could not reduce students' sharing intentions of deepfakes and, therefore, can be considered unsuccessful.

Social Media Trust

An additional parallel mediation model was conducted including the effect of social media trust and its potential influence on the inoculation and people's sharing intentions. Due to this, social media trust was included as an additional variable in the mediation model used for the hypothesis testing. It was found that the intervention did not affect social media trust significantly, *B* = 0.17, *t* = 1.20, *p* = .23. Moreover, there was no effect of social media trust on sharing intentions, *B* = 0.07, *t* = 0.78, *p* = .44. This means that in this study, social media trust did not have an influence on the relation between inoculation and students' sharing intentions. Apart from this, the effects of the other variables considered in the model to test the hypotheses remained nonsignificant. In other words, social media trust was not a significant predictor of sharing intentions and did not change the predictive power of the model.

Discussion

Due to the immense impact of conspiracy theories and their various negative consequences, the current study aimed at investigating whether an intervention was able to reduce students' sharing intentions of deepfakes. This was tried to be accomplished by boosting students with the necessary knowledge about deepfakes and explain how their impact can be reduced. An online survey was designed on the basis of inoculation theory.

Why do people share videos among social media?

In this study, the funniness of a videos significantly predicted whether people would intend to share deepfakes or not. In other words, the funnier people considered the deepfakes to be, the more likely they would share them among social media. Furthermore, polarization was a marginal predictor for sharing intentions. This is in line with findings of Lagger et al.

(2017) who also found that one of the main reasons for sharing videos is their entertainment or fun aspect. People share these videos with other people to make them laugh about or be amused with the video as well. Furthermore, sharing videos is used to inform other people of interesting videos or important news (Lagger et al., 2017). This could be a reason for why polarization factor of a video tended to predict sharing intentions. When sharing information among social media, people do not base their intention on the veracity of information (Pennycook & Rand, 2021) but rather disseminate them out of curiosity if the information was true or to spread chaos (Altay et al., 2021; Petersen et al., 2018). Thus, the more polarizing or attention-grabbing videos (i.e., deepfakes) are, the more likely people share them even if knowing that they are fake. Hence, this study confirmed basic sharing mechanisms for videos among social media which were found in prior research and give potential links for how these may apply to the increasing spread of deepfakes over the last years.

Sharing intentions

The first hypothesis of this study examined whether the manipulation was able to reduce students' sharing intentions of the videos displayed. However, it was found that students' sharing intentions in the Inoculation present group were not significantly lower than in the Inoculation absent group. Even after having controlled for people's funniness ratings, inoculation did not have an effect on sharing intentions. In line with inoculation theory (Banas & Miller, 2013), it was assumed that when students learn about the potential negative consequences of deepfakes for the persons involved, they may become motivated to resist deepfakes and do not share them among social media in the process. This means that in this study there was no support for the assumption that inoculation can influence people's sharing intentions. One reason for these findings is that students perceived threat could not be increased. As already mentioned, perceived threat is an essential component of inoculation (Banas & Miller, 2013), but also a necessary prerequisite for people to become motivated to counteract a threat (Gore & Bracken, 2005; Norman et al., 2005). As there was no significant difference between both groups in perceived threat, it may be that both groups were equally motivated to resist deepfakes. This is also underlined by the below midpoint scores on sharing intentions of both groups. This shows that students in both groups did rather not intend to share the displayed deepfakes.

Nevertheless, it was found that sharing intentions were positively correlated with perceived emotional threat which in turn was negatively related to students' perceived coping ability. These relationships are in line with the EPPM, suggesting that if people feel less able to cope with deepfakes, that they react more maladaptively. Because of this, they then cope inadequately, i.e., by sharing the videos more among social media (Gore & Bracken, 2005). However, according to Loewenstein et al. (2001), emotions can also be determinants of behaviours. This would, however, imply that when people react more emotionally and feel less able to cope with deepfakes that they would rather show avoidance behaviours (Loewenstein et al., 2001). In the case of deepfakes, however, it seems that people with higher emotions would tend to share more deepfakes. In other words, certain emotions (e.g., funniness) seem to increase the likelihood that people distribute deepfakes on social media. Thus, the more emotionally people react to deepfakes, the quicker they may share deepfakes out of curiosity or because they think such information to be interesting if it was true (Altay et al., 2021). This finding is particularly interesting because it suggests that there may be certain emotions which may overrule the fear elicited by a threat. Because of this, people do not refrain from the threat posed by deepfakes but rather seem to approach the threat by sharing it with other people. This underlines the importance of the sharing dimensions used for this study. They indicate which exact emotions may lead people to share deepfakes or could help in reducing the spread of deepfakes. Future studies may try to target particularly those emotions which help people to reduce their spread of deepfakes and elaborate on these emotions to reduce the impact of deepfakes.

Attitudes

The second hypothesis stated that people's attitudes towards deepfakes would mediate the effect of the inoculation on student's sharing intentions. However, this study did not find support for this hypothesis. Neither did the inoculation decrease people's attitudes towards deepfakes, nor did students' attitudes influence their sharing intentions. In this study, students learned about the persuasive attempt of deepfakes of displaying a false reality and they received insights into their negative consequences. According to the PKM, this increased persuasion knowledge should lead to higher skepticism and more negative attitudes towards deepfakes (Iacobucci et al., 2021). This in turn should lead to lower sharing intentions as predicted by the TPB (Ajzen, 1985). Yet, none of these relationships were found.

One explanation for this may be that students already knew about deepfakes and thus, already knew about their negative consequences. In other words, it may have been the case that students already possessed a high persuasion knowledge as described by the PKM. This would explain why their attitudes were rather low. With the rapidly increasing numbers of deepfakes being shared online over the last years (Kugler & Pace 2021), and people in the ages from 18-29 being most frequently use social media platforms (Khoros, 2021), it can be assumed that most students already encountered deepfakes. Hence, they already formed rather negative attitudes about them. This would explain why the groups did not differ significantly in their attitudes.

Perceived coping ability

The final hypothesis expected another mediation effect of perceived coping abilities on the effect of inoculation on sharing intentions. It was expected that students' perceived coping abilities would increase by raising awareness about deepfakes and showing that not sharing deepfakes can help to reduce their impact. In line with the EPPM, this increased perceived coping ability then should lead to less emotional reactions among students' towards deepfakes and in turn lower sharing intentions (Gore & Bracken, 2005). Again, however, no indications for such a mediation were found in this study. Neither seemed students' perceived coping abilities to be influenced by the inoculation, nor did they have an effect on students' sharing intentions. Generally, students showed relatively high coping abilities. This suggests that students' generally cope relatively well with deepfakes. This high perceived coping ability of students may have also been the reason for why students' cognitive and emotional perceived threat could not be raised. In other words, as students felt capable of dealing with deepfakes, they did not cope emotionally with deepfakes and did not show particularly negative emotions towards them.

One reason for this may be younger people's high exposure to and usage of social media (Khoros, 2021). It may be that students already knew deepfakes and/or may already have been exposed to some. Another explanation for the high perceived coping abilities is the so-called above-average effect. This effect suggests that people tend to describe themselves as being above average, thereby contradicting the logic of statistics (Kruger & Dunning, 1999). This effect was also found in people's expectations in their abilities to detect deepfakes (Köbis et al., 2021). However, deception detection literature suggests that people are generally not able to detect deception and are easily influenced by it (Hancock & Bailenson, 2021). In the near future, it can be expected that deepfakes cannot be distinguished anymore from real footage with the naked eye (Maras & Alexandrou, 2018). This will make it more difficult, if not impossible, to actually assess deepfakes as fakes and deal with them adequately in the process. Thus, it may be that students overestimate their abilities to detect deepfakes which would explain their confidence in dealing with deepfakes.

Social media trust

Social media trust was investigated as an exploratory variable on its potential relationships with the other variables investigated in this research. Prior research already found that deepfakes reduce trust in social media (Vaccari & Chadwick, 2020). Furthermore, it was found that the more people trust information on social media, the more likely people shared such information with other people (Stefanone et al., 2019). Hence, social media trust was investigated as an additional mediator of the effect of inoculation on sharing intentions. Again, however, no mediation of social media trust on the effect of inoculation on sharing intention could be found.

Once more it can be assumed that due to students' high exposure to social media including different forms of false information, that their social media trust was very low already prior to the present study. In line with Stefanone et al. (2019), this would explain why students' sharing intentions in both groups were low and did not show different evaluations of deepfakes. Nevertheless, prior social media trust should be considered in the future.

Limitations

There are some limitations which have to be considered to be able to put the current findings into perspective. Firstly, this study did not investigate real sharing behaviours, but only students' intentions to share the displayed videos. According to the TPB, intentions are direct antecedents for behaviour (Ajzen, 1985). However, meta-analyses showed that they cannot fully account for people's behaviours (Sutton, 1998). Nevertheless, sharing intentions were already found by prior studies to account for the virality of deepfakes (Iacobucci et al., 2021). Thus, despite not being equal to sharing behaviours, sharing intentions are sufficient to be used for investigating the extent to which people share videos among social media. Still, generalizations need to be treated with caution.

Furthermore, the finding that the funniness dimension was found to be a significant predictor of sharing intentions may have been caused by the content of the deepfakes used for this study. In reality, most deepfakes include pornographic material (Lyu, 2020) or they may target politicians as a means for political sabotage (Maras & Alexandrou, 2018). However, using explicit material or extremely polarizing content was not possible from an ethical standpoint in this study as such forms of deepfakes may have been disturbing for participants. This is why rather funny or harmless videos were used. Yet, also relatively harmless videos may cause harm to the persons involved and may have been created without the displayed person's consent. Nevertheless, this may explain why perceived threat could not be raised and why funniness predicted sharing intentions. If more extreme videos would be used, different emotions may be more important to be considered. Furthermore, it may be that the elicited fear of the intervention would not be overruled by other emotions. Therefore, the content of the videos used should be considered when interpreting the findings of this study.

Recommendations for future research

Regarding future research, there are different aspects which may be targeted. Firstly, the dimensions on which deepfakes were assessed may be tested further in future studies. These dimensions can give insights into why certain deepfakes may be shared. Furthermore, they help to put findings into perspective and can be used as a means to avoid potential misrepresentations of effects. Therefore, these dimensions should be added into or even elaborated on in future research. As already mentioned, this would also have the benefit to be able to target specific emotions leading people to limit the spread of deepfakes.

Apart from this, people's prior attitudes or prior knowledge about deepfakes may be added as additional variables. In this study, only attitudes towards deepfakes were measured after having been exposed to the intervention. This means that attitudes were only measured as a between-groups variable. By including prior attitudes towards deepfakes and compare them to participants' post-attitudes, potential changes in attitudes could be investigated as a within-groups factor. Furthermore, people's prior knowledge about deepfakes may be included. In this study, the investigated variables may have been influenced by students' existing knowledge about deepfakes as well as awareness about the negative consequences. Hence, prior knowledge may be used as an additional variable in future research.

In line with this, a different sample than students may be chosen. In the current study, students were chosen as the target group due to their increased usage of social media (Bantimaroudis et al., 2020; Khoros, 2021; Statista, 2021). This increased usage makes them susceptible to be exposed to deepfakes which made them suitable to be targeted by this intervention. Nevertheless, students were found to perceive themselves as capable to deal with deepfakes. Furthermore, their relatively high usage of social media may have led them to already having been exposed to deepfakes, thereby learning about their usage and consequences. Thus, choosing a target group with less exposure to social media and potentially lower coping ability and knowledge about deepfakes may benefit even more from an intervention raising awareness and teaching for how to deal with deepfakes.

Another interesting option for future studies would be to include both real videos and deepfakes and investigate people´s accuracy judgements. A previous study of Iacobucci et al. (2021) already found an effect of deepfake recognition on people's attitudes and sharing intention. In reality, deepfakes are often used to deceive people and are not labelled as such. As people are typically not able to distinguish real from fake footage (Hancock & Bailenson, 2021), their attitudes may be more positive and their sharing intentions higher when not knowing to be exposed to deepfakes. Thus, deepfake recognition should be controlled for when investigating sharing intentions. Another advantage of letting people distinguish real from fake footage is that it may not only have theoretical implications to examine whether sharing intentions may be reduced by an intervention. It also may have practical implications as people learn to detect deepfakes. Therefore, including deepfake recognition in future studies should be considered.

References

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action control* (pp. 11-39). Springer, Berlin, Heidelberg.

Altay, S., de Araujo, E., & Mercier, H. (2021). “If this account is true, it is most enormously wonderful”: Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, 1-22. doi:10.1080/21670811.2021.1941163

Anderson, C. L., & Agarwal, R. (2010). Practicing safe computing: A multimethod empirical examination of home computer user security behavioral intentions. *MIS quarterly*, 613-643. doi:10.2307/25750694

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research, 39*(2), 184-207. doi:10.1111/hcre.12000

Bantimaroudis, P., Sideri, M., Ballas, D., Panagiotidis, T., & Ziogas, T. (2020). Conspiracism on social media: An agenda melding of group-mediated deceptions. *International Journal of media & cultural politics*, *16*(2), 115-138. doi:10.1386/macp_00020_1

Brandalism Project (2019, June 14). *'I wish I could...' (2019)* [Video File]. YouTube. https://www.youtube.com/watch?v=3f66kBwfMto&t=19s

Burkhardt, J. M. (2017). History of fake news. *Library Technology Reports*, *53*(8), 5-9.

Caporusso, N. (2020). Deepfakes for the good: A beneficial application of contentious artificial intelligence technology. In *International Conference on Applied Human Factors and Ergonomics* (pp. 235-241). Springer, Cham. doi:10.1007/978-3-030-51328-3_33

De Kimpe, L., Walrave, M., Verdegem, P., & Ponnet, K. (2021). What we think we know about cybersecurity: an investigation of the relationship between perceived knowledge, internet trust, and protection motivation in a cybercrime context. *Behaviour & Information Technology*, 1-13. doi:10.1080/0144929X.2021.1905066

Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of consumer research*, *21*(1), 1-31. doi:10.1086/209380

Gore, T. D., & Bracken, C. C. (2005). Testing the theoretical design of a health risk message: Reexamining the major tenets of the extended parallel process model. *Health Education & Behavior*, *32*(1), 27-41. doi:10.1177/1090198104266901

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, behavior, and social networking*, *24*(3), 149-152. doi:10.1089/cyber.2021.29208.jth

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973-986. doi:10.1177/1745691617702496

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes Unmasked: The Effects of Information Priming and Bullshit Receptivity on Deepfake Recognition and Sharing Intention. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 194-202. doi:10.1089/cyber.2020.0149

Khoros. (2021). *The 2021 social media demographics guide.* https://khoros.com/resources/social-media-demographics-guide

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146. doi:10.1016/j.bushor.2019.11.006

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *Iscience*, *24*(11), 103364. doi:10.1016/j.isci.2021.103364

Krimsky, S. (2007). Risk communication in the internet age: The rise of disorganized skepticism. *Environmental hazards*, *7*(2), 157-164. doi:10.1016/j.envhaz.2007.05.006

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121. doi:10.1037//0022-3514.77.6.1121

Kugler, M. B., & Pace, C. (2021). Deepfake privacy: Attitudes and regulation. *Nw. UL Rev.*, *116*, 611. doi:10.2139/ssrn.3781968

Kumar, R. (2018). *Research methodology: A step-by-step guide for beginners*. Sage. doi:10.7748/nr.19.3.45.s5

Kwok, A. O., & Koh, S. G. (2021). Deepfake: a social construction of technology perspective. *Current Issues in Tourism*, *24*(13), 1798-1802. doi:10.1080/13683500.2020.1738357

Lagger, C., Lux, M., & Marques, O. (2017). What makes people watch online videos: An exploratory study. *Computers in Entertainment (CIE)*, *15*(2), 1-31. doi:10.1145/3034706

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*, *6*, 26. doi:10.3389/fcomm.2021.632317

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological bulletin*, *127*(2), 267. doi:10.1037/0033-2909.127.2.267

Lyu, S. (2020). Deepfake detection: Current challenges and next steps. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 1-6). IEEE. doi:10.1109/ICMEW46912.2020.9105991

Maras, M. H., & Alexandrou, A. (2018). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof*, *23*(3), 255-262. doi:10.1177/1365712718807226

Markalanfish (2020). *7-Level Emotional Scale*. https://markalanfish.com/2020/06/24/7-level-emotional-scale/

Norman, P., Boer, H., & Seydel, E. R. (2005). Protection motivation theory. *Predicting health behaviour*, *81*, 98-143. doi:10.1016/S0925-7535(97)81483-X

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770-780. doi:10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*. doi:10.1016/j.tics.2021.02.007

Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2018). A “need for chaos” and the sharing of hostile political rumors in advanced democracies. *American Political Science Association, Boston, MA*, *30*. doi:10.31234/osf.io/6m4ts

Pfau, M., Semmler, S. M., Deatrick, L., Mason, A., Nisbett, G., Lane, L., ... & Banas, J. (2009). Nuances about the role and impact of affect in inoculation. *Communication Monographs*, *76*(1), 73-98. doi:10.1080/03637750802378807

Reichelt, F. (2021). *What makes deepfakes believable?* doi:10.13140/RG.2.2.24357.63204

Sankaranarayanan, A., Groh, M., Picard, R., & Lippman, A. (2021). *The Presidential Deepfakes Dataset.* http://ceur-ws.org/Vol-2942/paper3.pdf

Schober, P., Boer, C., & Schwarte, L. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768. doi:10.1213/ANE.0000000000002864

Statista (2021). *Share of Europeans that tended to trust media in 2019, by education and medium*. https://www-statista-com.ezproxy2.utwente.nl/statistics/453791/europe-trust-in-media-by-education-and-medium/

Statistics How To (2016). *Guttman’s lambda-2: Definition, Examples.* https://www.statisticshowto.com/guttmans-lambda-2/

Stefanone, M. A., Vollmer, M., & Covert, J. M. (2019). In news we trust? Examining credibility and sharing behaviors of fake news. In *Proceedings of the 10th international conference on social media and society* (pp. 136-147). doi:10.1145/3328529.3328554

Sutton, S. (1998). Predicting and explaining intentions and behavior: How well are we doing?. *Journal of applied social psychology*, *28*(15), 1317-1338. doi:10.1111/j.1559-1816.1998.tb01679.x

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53. doi:10.5116/ijme.4dfb.8dfd

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, *6*(1), 2056305120903408. doi:10.1177/2056305120903408

Van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation.* doi:10.4324/9780429295379-11.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6), 1063. doi:10.1037/0022-3514.54.6.1063

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence*, *27*(1), 36-51. doi:10.1177/1354856520923963

Yang, C. Z., Ma, J., Wang, S., & Liew, A. W. C. (2020). Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, *16*, 1841-1854. doi:10.1109/TIFS.2020.3045937

Appendices

Appendix A

Question sharing dimensions

Please rate the video among the following dimensions.

	1	2	3	4	5	6	7	
Not funny	○	○	○	○	○	○	○	Funny
Not informative	○	○	○	○	○	○	○	Informative
Not polarizing	○	○	○	○	○	○	○	Polarizing
Not harmful	○	○	○	○	○	○	○	Harmful
Unintimidating	○	○	○	○	○	○	○	Intimidating
Unacceptable	○	○	○	○	○	○	○	Acceptable

Appendix B

Warning message Inoculation present group

In the following you will be confronted with different deepfakes which are shared among social media. The information below was taken from a scientific paper about recent insights about deepfakes. Please read the following information carefully.
Deepfakes are a form of machine learning and artificial intelligence which enable the visual and auditory manipulation of video material (Kietzmann et al., 2020). This form of video manipulation is used to transpose individuals' faces onto those of other persons and imitate their voices. This means that persons can be made to behave in a way in which they would not behave in reality.

There are, however, many problems connected to deepfakes. Most deepfakes are used for transposing people's faces onto pornographic content and hence, often create reputational damage for the persons involved (Lyu, 2020). Although deepfakes mostly target celebrities, everyone may potentially become victim of deepfakes. Moreover, by faking statements of politicians, deepfakes are also abused as a means for political sabotage and propaganda (Maras & Alexandrou, 2018) and hence, may lead to international conflicts (Caporusso, 2020). In other words, deepfakes can lead to distrust or even hatred towards the people being deepfaked and can lead to polarization among those believing in these videos (Yadlin-Segal & Oppenheim, 2021).

Due to these risks of deepfakes, their spread needs to be limited. Therefore, it is important that people do not share deepfakes among social media when encountering them (Pennycook & Rand, 2021). By placing them in a context which may damage their reputation or create polarization, hatred and distrust is incited. Due to this, the spread of deepfakes can be dangerous for the persons involved.

Please continue this study on the following page.

Appendix C

Warning message Inoculation absent group

In the following you will be confronted with different deepfakes which are shared among social media. The information below was taken from a scientific paper about recent insights about deepfakes. Please read the following information carefully.

Deepfakes are a form of machine learning and artificial intelligence which enable the visual and auditory manipulation of video material (Kietzmann et al., 2020). This form of video manipulation is used to transpose individuals' faces onto those of other persons and imitate their voices. This means that persons can be made to behave in a way in which they would not behave in reality.

Deepfakes became publicly known in late 2017 and were developed rapidly since then. Nowadays, deepfakes are easily accessible and can be created without much technical skills by everyone (Kietzmann et al., 2020). Modern deepfake techniques only require some photos or video recordings of the person to be deepfaked. With these images or videos the program learns itself how to use these images to create realistic fakes (Caporusso, 2020). Due to the quickly developing algorithms for the creation of deepfakes as well as algorithms to imitate the targeted person's voice, it can be assumed that within the next years, it will be much more difficult, if not impossible to distinguish real from fake video footages with the naked eye (Maras & Alexandrou, 2018).

Please continue this study on the following page.

Appendix D

Questionnaire measuring sharing intentions, perceived threat, attitudes perceived coping abilities and social media trust

Please indicate to what extent you think that deepfakes are...

	1	2	3	4	5	6	7	
Negative	○	○	○	○	○	○	○	Positive
Bad	○	○	○	○	○	○	○	Good
Unfavourable	○	○	○	○	○	○	○	Favourable
Unacceptable	○	○	○	○	○	○	○	Acceptable
Wrong	○	○	○	○	○	○	○	Right
Foolish	○	○	○	○	○	○	○	Wise

When thinking about deepfakes, to what extent do you feel...

	Not at all	A little	A moderate amount	A lot	Extremely much
Distressed	○	○	○	○	○
Angry	○	○	○	○	○
Insecure	○	○	○	○	○
Worried	○	○	○	○	○
Irritated	○	○	○	○	○
Powerless	○	○	○	○	○
Anxious	○	○	○	○	○

Please indicate to what extent you agree with the following statements.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I have every confidence that trusting information being shared among social media is safe	○	○	○	○	○	○	○
I am satisfied with the veracity of information being shared on social media	○	○	○	○	○	○	○
It is possible that I will be exposed to deepfakes on social media	○	○	○	○	○	○	○
It is likely that I will be exposed to deepfakes on social media	○	○	○	○	○	○	○
There is a great risk that I will be exposed to deepfakes on social media	○	○	○	○	○	○	○
The impact of deepfakes is significant	○	○	○	○	○	○	○
Deepfakes are a serious threat	○	○	○	○	○	○	○
Deepfakes are a severe threat	○	○	○	○	○	○	○
I feel capable of dealing with deepfakes on my own	○	○	○	○	○	○	○
I feel like I have the knowledge that is necessary to deal with deepfakes on my own	○	○	○	○	○	○	○

I feel like I have the skills that are necessary to deal with deepfakes on my own	○	○	○	○	○	○	○
If people do not share deepfakes among social media, their impact is reduced	○	○	○	○	○	○	○
By not sharing deepfakes among social media, misusage of deepfakes can be prevented	○	○	○	○	○	○	○
If people do not share deepfakes among social media, it is less likely that people will become victim of deepfakes	○	○	○	○	○	○	○