Cyberpsychology, Behavior, and Social Networking
Volume 24, Number 3, 2021

DOI: 10.1089/cyber.2020.0149

Deepfakes Unmasked: The Effects of Information Priming and Bullshit Receptivity on Deepfake Recognition and Sharing Intention

Serena Iacobucci, PhD,[1–3] Roberta De Cicco, PhD,[1,2,4] Francesca Michetti, MA,[1,2] Riccardo Palumbo, PhD,[1–3] and Stefano Pagliaro, PhD[5]

Abstract

The study aims to test whether simple priming of deepfake (DF) information significantly increases users' ability to recognize DF media. Although undoubtedly fascinating from a technological point of view, these highly realistic artificial intelligent (AI)-generated fake videos hold high deceptive potential. Both practitioners and institutions are thus joining forces to develop debunking strategies to counter the spread of such difficult-to-recognize and potentially misleading video content. On this premise, this study addresses the following research questions: does simple priming with the definition of DFs and information about their potentially harmful applications increase users' ability to recognize DFs? Does bullshit receptivity, as an individual tendency to be overly accepting of epistemically suspect beliefs, moderate the relationship between such priming and DF recognition? Results indicate that the development of strategies to counter the deceitfulness of DFs from an educational and cultural perspective might work well, but only for people with a lower susceptibility to believe willfully misleading claims. Finally, through a serial mediation analysis, we show that DF recognition does, in turn, negatively impact users' sharing intention, thus limiting the potential harm of DFs at the very root of one of their strengths: virality. We discuss the implications of our finding that society's defense against DFs could benefit from a simple reasoned digital literacy intervention.

Keywords: deepfake, deception, bullshit receptivity, sharing intention

Introduction

The spread of fake news in the digital environment is emerging as one of the preeminent technological threats of our ''post-truth'' era.[1] Fictitious content is purposely developed and spread with the aim of malevolently deceiving public opinion.[2–5] Facts become less influential and are overridden by fake manipulated content leveraging emotions, biases, and personal beliefs.[6] The recent uproar over this noxious spread of information is, for many, being framed as an information literacy problem rather than only a technological problem,[7,8] and an urgent call has been raised to equip citizens with the required knowledge and ability to discern facts from fake stories.[9] However, the digital threat has found renewed strength since the rise of fake video content, marking a turning point in information warfare.[10,11]

Within the AI-generated technologies that are making the manipulation of online content easier and more accurate—and thus increasingly deceitful—the recent research attention that has been paid to deepfakes (DFs) is surely justified. As fascinating as they are uncanny, DFs can first be defined, from a merely technical point of view, as synthetic media developed through machine-learning-based software that allows users to fabricate fake videos in which one person's face and voice are superimposed onto existing footage. This results in realistic-looking videos where anyone can appear to have said or done something they have not. A more comprehensive definition of DFs, however, needs to take into account their unprecedented, highly deceptive, and manipulative potential.[11]

Being ranked as the most serious AI crime threat in 2020,[12] DFs not only represent the ''newest way to commit the oldest crime[s],'' (p. 1)[13] such as political sabotage,

[1]Department of Neurosciences, Imaging and Clinical Sciences, University ''G. d'Annunzio'' of Chieti-Pescara, Chieti, Italy.
[2]Center for Advanced Studies and Technologies (CAST), University ''G. d'Annunzio'' of Chieti-Pescara, Chieti, Italy.
[3]European Capital Markets Cooperative Research Centre, Chieti, Italy.
[4]Católica Porto Business School, Universidade Católica Portuguesa, Porto, Portugal.
[5]Group Processes and Morality Lab, University ''G. d'Annunzio'' of Chieti-Pescara, Chieti, Italy.

identity theft, harassment, blackmail and fraud,[14] but alarmingly facilitate a great variety of newest cybercrimes as well, from cyberbullying to revenge porn. In this sense, DF technology has now become the latest frontier of AI-based cybercrime.

Scholars[15,16] have recently provided a motivational taxonomy of cybercrime based on perpetrators' primary benefits, as well as victims' primary losses, namely the Tripartite Cybercrime Framework, which identifies socioeconomic, psychosocial, and geopolitical motivations as the three main factors behind cybercrimes. Within such a taxonomy, DFs generally pose a threat not only to individuals but also to businesses, political systems, and national security.[17]

On the one hand, it is a daunting truth that, as of today, the majority of DFs tend to fully leverage their believability[11] for harmful purposes,[18] with pornographic examples accounting for 96 percent of all DFs online.[19,20] On the other hand, not all DFs are created for malicious purposes and many might be intended purely for entertainment or educational ends.[21]

Indeed, in this study, we focus on a DF intended for entertainment purposes that, although not explicitly psychologically motivated, might nonetheless cause some level of distress. Even DFs designed to entertain may, for instance, provoke psychological distress in users when they realize they have been deceived or if they perceive what they have seen as uncanny or potentially harmful. Moreover, even if we set any malicious intent aside, the ethical dilemma of DFs' potential deceitfulness still holds. Authorities, policy makers, and researchers are currently joining forces to develop suitable techniques to address this, focusing on technological, legal, and administrative solutions.[18]

From a technological perspective, one counter-DF strategy is the development of AI-based detectors. These tools, however, represent a double-edged sword: with the two systems basically working in parallel, DF makers promptly work to overcome each new DF detection strategy,[22] thus further perfecting their techniques and making it only a matter of time before the fakes become indiscernible from genuine content.[23] From a legal perspective, the scenario is similarly uncertain and discouraging: researchers have underlined regulators' lack of required technical expertise and resources, which seems to confirm that current regulatory responses alone will not suffice.[24]

As both legal and technological solutions are thus either lagging behind or facing strong slowdowns, there is a growing and compelling need to tackle the issue from an educational point of view by developing cultural awareness around DFs. Debunking strategies against this potential threat should include increased user awareness from a cultural and educational perspective, with the aim of educating users to start questioning even the most realistic types of content.[25,26]

This study

To assess the underexplored educational approach, this study develops a threefold perspective. First, literacy interventions should enhance users' inoculation against the potentially harmful effects of exposure to misleading information[7,27]; for that reason, and to enhance the audience's knowledge, interventions should also implement simple priming with the definition of DFs and information about their potentially harmful applications. Thus, we explore whether priming users with a DF definition and explanation of its potential threats can lead to increased DF recognition.

Second, we investigate whether and how this priming effect works in participants with differing degrees of individual relatively stable disposition to believe in fake content, defined as their bullshit (BS) receptivity.[28] Finally, we address the virality of DFs, investigating the impact that DF recognition has on attitudes toward the video and, in turn, on sharing intention.[29]

Priming and DF recognition. As reported by McNamara,[30] priming can be defined as ''an improvement in performance in a perceptual or cognitive task, relative to an appropriate baseline, produced by context or prior experience'' (p. 3). Within the priming paradigm, priming is applied to manipulate or increase knowledge activation, and applied to a target object in an effort to produce specific outcomes.[31] The recent intensification in online fake news has caught authorities' interest and stimulated the search for factors that could decrease individual susceptibility. So far, among the various approaches, some have been based on inoculation theory,[32] which posits that individuals can be inoculated against persuasive attempts in the same way that organisms can be immunized against viral diseases.

Accordingly, by making certain issues salient, previous alerting stimuli (e.g., primes) that individuals have encountered can determine how they evaluate subsequent stimuli.[33] Priming has been widely explored in several fields, but deception-related findings come from advertising. Priming participants with the notion of covert forms of advertising (e.g., native advertising) not only influences how they perceive and evaluate the advertising, but also leads to greater recognition.[33,34]

As research on priming offers substantial evidence that individuals are primed with messages seen in media, it is worth considering the potential strength of educational messages acting as a warning. In fact, by emphasizing certain issues in the media, individuals ''can be more cognizant of fake news, especially when it is at the top of their mind and easily accessible'' (p. 32).[35] Hence, bearing in mind the need to find mechanisms to address the DF phenomenon, we believe that priming users with the definition of DF and its potential threats may be fruitfully implemented to enhance audiences' knowledge and capacity to question and discern whether a video is real or fake. Drawing from research on priming and knowledge activation,[34] we hypothesize the following:

H1. Priming participants with the notion of DFs will lead to greater DF recognition compared with control groups (not primed with the notion of DFs).

The role of BS receptivity in fake content believability. Quoting George Carlin, ''Bullshit is everywhere,'' (p. 1) today more than ever.[36] Information spreads online at an ever-increasing pace and without quality control, thus increasingly exposing people to BS. In an initial investigation of the issue of BS—defined as an insidious way of speaking that is ''unconnected with the truth'' (p. 9)[37]—Pennycook et al.[28] focused on a specific form that they defined as pseudoprofound BS, which consists of seemingly impressive statements that—even though presented as profoundly meaningful and with the syntactic correctness necessary for linguistic cognition—are completely meaningless.[38]

Individuals' tendency to ascribe profundity to BS statements has been consistently associated with a wide range of psychological factors, including a less reflective cognitive style, lower cognitive ability, and increased epistemically suspect beliefs, such as proneness to ontological confusions and conspiratorial ideation.[39]

In later studies, Pennycook and Rand[40] assessed the role of pseudoprofound BS receptivity in the domain of fake news, as untrustworthy information is seemingly characterized by a lack of regard for the truth. Interestingly, they provided evidence that BS receptivity correlated positively with the perceived accuracy of fake news, and negatively with media-truth discernment. We expect the same process to lead individuals with higher levels of BS receptivity to ineffectively judge the untruthfulness of the DF video in our study. We thus predict that individuals with higher BS receptivity will apply less cognitive processing and mental resources compared with individuals with lower BS receptivity in assessing the veracity of the video.

As depicted in Figure 1, our predictions call for a two-way interaction where priming with DF information and its potential threats should have a lower positive effect on DF recognition for individuals with a higher BS receptivity than for those with a lower BS receptivity:

H2. BS receptivity moderates the relationship between priming and DF recognition, such that the less individuals demonstrate BS receptivity, the more they will be able to recognize DFs after being primed, and vice versa.

Fake content recognition, attitudes, and behavioral intents. The persuasion knowledge model (PKM)[41] explains how priming with understanding of a topic helps to develop knowledge to cope with persuasion episodes. Similarly, according to Amazeen and Bucy's procedural news knowledge (PNK) approach,[42] understanding of professional news operations and procedures facilitates audiences' ability to parse news-related content from imposter content. Thus, according to both of these theories, developing consumers' knowledge of covert techniques becomes crucial in enhancing their awareness.[34]

Such knowledge, however, not only serves as an antecedent to the recognition of disinformation but also affects consequent coping responses.[42] PNK and PKM are two good examples of how priming with salient knowledge can inoculate audiences against disinformation and close the resistance gap. When an individual cognitively elaborates on whether a message contains persuasive intent, negative thoughts might arise and resistance occur. A number of studies have shown that coping mechanisms heighten suspicion or skepticism and negatively influence the consumer's attitude and behavioral intent.[41,43–53]

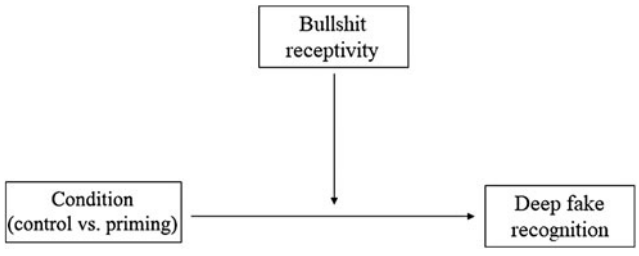


FIG. 1. The figure reports the hypothesized moderation model.

As already mentioned, in the case of covert advertising where recent studies examined the effects of coping mechanisms on attitude, individuals primed with a definition of native advertising were found more likely to recognize such content as an advertisement, identify its persuasive intent, and, accordingly, be more critical of it.[54] As a result of such a coping-response strategy in terms of resistance-induced counterargument against persuasive intent, consumers will have a greater negative attitude toward the content.[55]

Furthermore, according to the theory of planned behavior (TPB),[29] such an unfavorable attitudinal predisposition will result in more negative behavioral intentions. The TPB posits a strong attitude–behavior relationship, where a change in behavior occurs by means of attitude change. Following this rationale, we expect that familiarizing consumers with knowledge of DF practice will prompt them to recognize its duplicity and, therefore, develop a lower attitude toward the video in question. We further expect a strong attitude–behavior relationship between the attitude toward the video and the intention to share it:

H3. Priming participants with the definition of DFs reduces intention to share the video compared with control through DF recognition and attitude toward the video.

Methods

Design, participants, and procedure

We adopted a single factor between-participant design in which priming with the DF definition was manipulated at two levels (conditions: control vs. priming). Participants were recruited from a database of volunteers who had agreed to take part in experiments carried out by their university's Department of Behavioral Economics.

We performed data collection throughout the month of December 2019. Clicking on a link redirected participant to our questionnaire, which was hosted through Qualtrics. Participants were informed of the university's ethical approval process and data storage and processing procedures, which adhere to Regulation (EU) 2016/679 GDPR. Following the ethical standards of the 1964 Declaration of Helsinki, participants were informed of their right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal.

A total of 152 Italian participants explicitly consented to participate and started the online survey, completing all measures of interest (73 percent female; $M_{age} = 36.09$; age range = 20–69; standard deviation [*SD*] = 12.82). Some participants were randomly instructed to read a short explanation of DFs and their applications, including potentially harmful ones (i.e., priming condition); others were given no such text (i.e., control condition).

All participants then viewed the DF video developed by the YouTube channel ''Ctrl Shift Face'' through the DeepFaceLab tool, where the face of original actor Jack Nicholson was digitally replaced with that of Jim Carrey. The video was accompanied by a caption that praised Carrey's facial ability to perfectly re-enact the original scene, by mimicking Nicholson's facial acting features. As the false astounding mimicry was fundamental for the manipulative power of the DF, we showed participants the unaltered scene from the original movie from which the clip was taken to help them familiarize themselves with or recall this original version (Fig. 2).

FIG. 2. The figure reports a comparison between the two scenes in eight thumbnails from the original (*left*) versus the DF (*right*) video. Thumbnails from the original video were retrieved from a cut of the digital version of ''The Shining'' movie, Warner Bros. Entertainment, All rights Reserved. Thumbnails from the DF video were retrieved from a cut of the YouTube ''The Shining starring Jim Carrey: Episode 1—Concentration [DeepFake],'' from user Ctrl Alt Face.[a]

To maximize participants' feeling that they were being exposed to the video through their Facebook newsfeed, we framed a first snapshot of the DF video in a Facebook-style format (Fig. 3).

Measures

Participants completed a list of measures based on previous relevant research. BS receptivity was measured according to Pennycook et al.[28] Participants were asked to rate the profoundness of pseudoprofound sentences (10 items, e.g., ''Imagination is inside exponential space time events''; ''We are in the midst of a high-frequency blossoming of interconnectedness that will give us access to the quantum soup itself'') on a 5-point scale from 1 = ''not at all profound'' to 5 = ''very profound'' ($\alpha = 0.87$). Higher scores indicate an inability to spontaneously discern or detect pseudoprofound BS.

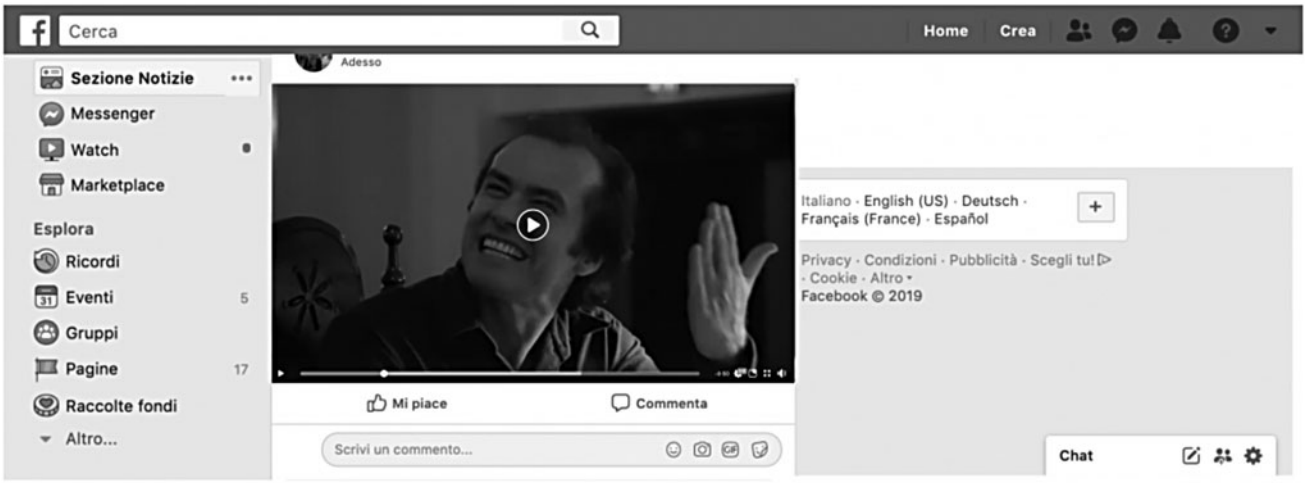
In line with prior research, pseudoprofound statements were randomly intermixed with prototypically profound (non-BS) sentences (10 items, e.g., ''Your teacher can open the door, but you must enter by yourself''; ''A river cuts through a rock not because of its power but because of its persistence'').[28]

DF recognition was measured with a single item: ''The similarity of the remake to the original scene is due to the actor's abilities and not to digital video editing technologies.'' Participants answered this question on a scale ranging from 1 = ''not at all'' to 5 = ''absolutely.'' Before the analysis, the item was reversed: the higher the score, the higher the DF recognition.

Attitude toward the video was measured with an established six-item, 5-point semantic differential scale (bad/good, foolish/clever, unpleasant/pleasant, useful/useless, boring/interesting, and negative/positive)[56] ($\alpha = 0.87$).

Intention to share the video was measured according to Hsieh et al.[57] through 3 items: ''I think this video is worth sharing with others; I will recommend this video to others; I will share this video to my friends through the Internet'' on a 5-point Likert scale ranging from 1 = ''strongly disagree'' to 5 = ''strongly agree'' ($\alpha = 0.91$).

The questionnaire was administered in Italian and all measures were translated, except for the prototypically profound non-BS sentences (which are traditionally held as truths or advice, such as proverbs and famous quotes; thus, the Italian versions were retrieved through a Google search) and DF recognition, which was specifically created for this



Note: The frame increases the salience of being exposed to a content appearing on one's news feed rather than just embedded in the Qualtrics environment.

FIG. 3. A screenshot of the video framed in a Facebook-style format. The frame increases the salience of being exposed to a content appearing on one's newsfeed rather than just being embedded in the Qualtrics scenario.

study and thus directly developed in Italian. Given the singularity of the BS-receptivity scale, a holder of an MSc in linguistics with an academic background and professional translation expertise forward translated our adaptation; a panel of three experts then reviewed the scale. The subsequent evaluations balanced the lack of any concern for meaning or truth with a correct syntactic structure, so that participants would ascribe meaning and credibility to the semantic information depending on their levels of sensitivity to BS. At the same time, suspicion arousal had to be avoided so that pseudoprofound statements were not easily recognized as nonsense.[28,58] The Italian version of the BS-receptivity scale is provided as Supplementary Table S1.

Results

A set of confound checks revealed that differences in the distribution of participants in terms of gender (χ^2 (1, 152) = 0.202, $p = 0.653$), education ($t(150) = 1.076$, $p = 0.284$), and age [$t(150) = 0.509$, $p = 0.612$] between the two conditions were not statistically significant. Means, *SD*s, and bivariate correlations for study variables are given in Table 1.

We performed an independent samples *t*-test to test whether priming participants with the notion of DF actually led to greater DF recognition. The *t*-test showed that this condition had a significant effect on DF recognition, $t(150) = -2.67$, $p = 0.008$, 95% confidence interval (CI) [–0.76 to –0.11], $d = -0.43$. In line with H1, participants in the priming condition showed greater DF recognition ($M = 3.25$, $SD = 1.29$) than did participants in the control condition ($M = 2.69$, $SD = 1.31$). To test for the hypothesized moderating effect of BS receptivity on the relationship between condition and DF recognition, we relied on PROCESS, the MACRO for SPSS developed by Hayes.[59] We tested model number 1 with 5,000 bootstrap resamples. BS receptivity was modeled as a dichotomous moderator by means of a median split, since the empirical median ($Mdn = 3.00$) coincided with the theoretical median. The overall equation was significant: $R^2 = 0.02$, $F(3, 148) = 4.40$, $p = 0.005$. Crucially for the present purpose, the condition of BS–receptivity interaction significantly increased the explained variance, $\Delta R^2 = 0.02$, $F(1, 148) = 3.78$, $p = 0.05$. The interaction effects are depicted in Figure 4.

Congruent with H2, the relationship between the condition and DF recognition was positive and significant among people with low BS receptivity ($b = 0.48$, CI: [0.20 to 0.76]). By contrast, the relationship between the condition and DF recognition was not reliable among people with high BS receptivity ($b = 0.07$, CI: [–0.23 to 0.38]). This partially confirmed H2, showing that priming users with DF knowledge influences their ability to recognize a DF, but only when they are not strongly inclined to the reception of BS. Of note, our findings hold in terms of both associations and significance levels controlling for the profundity rating of mundane statements, indicating that our results were not influenced by participants' tendency to see profoundness in everything.[60]

Finally, we tested whether priming DF knowledge reduced the intention to share the video through DF recognition and attitude toward the video itself. We thus

Table 1. Means, Standard Deviations, and Correlations of the Variables

	Mean	SD	*(1)*	*(2)*	*(3)*	*(4)*
(1) DF recognition	2.96	1.33	—			
(2) Bullshit receptivity	2.95	0.80	–0.112	—		
(3) Attitude toward the video	3.03	0.82	–0.205*	–0.146	—	
(4) Intention to share the video	2.05	1.16	–0.318**	0.086	0.455*	—

*$p < 0.05$, **$p < 0.01$.
DF, deepfake; *SD*, standard deviation.

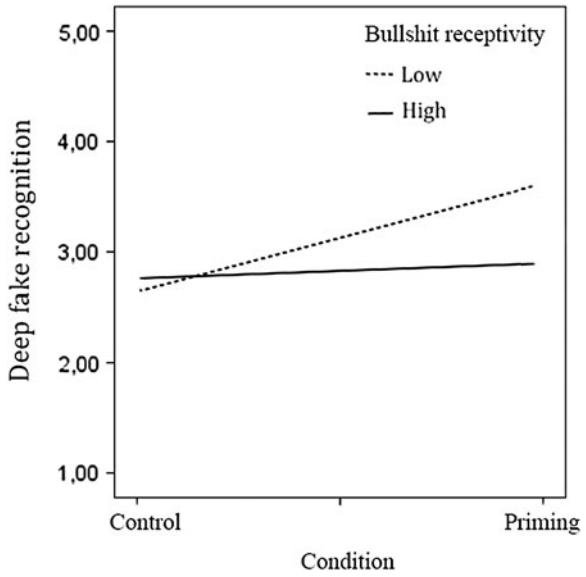


FIG. 4. Moderation effect of bullshit receptivity on DF recognition at the different levels of condition.

relied on PROCESS model n. 6. Condition (0 = control vs. 1 = priming) was modeled as an independent variable, and DF recognition and attitude toward the video were modeled as sequential mediators, assuming that the presence of DF information priming would have elicited higher DF recognition and this, in turn, would have induced more negative attitude toward the video. This reflects the order in which the questionnaire assessed these two constructs. The overall equation was significant ($R^2 = 0.27$, $F(3, 148) = 18.56$, $p < 0.001$).

As shown in Figure 5, the indirect effect of DF information priming on the intention to share the video through the hypothesized causal chain was significant (5,000 resamples; $b = -0.02$, boot $SE = 0.01$, 95% bias-corrected 5,000 bootstrap confidence intervals [BCBCI]: −0.0519 to −0.0013). H3 was thus confirmed: higher levels of DF recognition (induced through priming) reduce intention to share the video through attitudes toward the video. Direct and indirect effects are displayed in Table 2.

Discussions, Limitations, and Future Research

Given the lack of DF-related research in social sciences and, more specifically, from a media psychology perspective, this study provides a first broader look into this issue.

First, our results contribute to the idea that debunking strategies should consider users' education and awareness of the existence of manipulated online content, which, as of 2020, represents the primary AI-based cyberthreat.[12] On this matter, we confirmed the positive effect of DF information priming on users' ability to successfully identify DF content. We show that, as happens with fake news in PNK, simple information priming serves as a soft form of inoculation and creates a form of ''umbrella protection'' against the risk of being deceived by DFs.[7,61] This provides users with the required knowledge and skills to correctly identify and interpret the real nature of the content they encounter, which is consistent with Amazeen and Bucy's PNK theory.[42]

Second, we report that prior knowledge generated by DF information priming can represent a necessary but not sufficient condition for triggering DF recognition. There are, indeed, many individual characteristics that might hamper DF detection. Specifically, we investigated the individual tendency to believe in pseudoprofound online content that prior literature has confirmed as negatively correlated to the ability to discern media truth[40]—that is, BS receptivity.

More importantly, we record an interaction between simple priming with DF definition and information about its potentially harmful applications and BS receptivity: priming users with prior information about DF does enhance their ability to identify the videos' deception, but only for individuals with low levels of BS receptivity.

Although encouraging, this study is not exempt from limitations, which may provide avenues for further research. First, participants were exposed to a brief broad overview on DFs rather than a specific refutational-same inoculation message. Further studies might assess whether different in-depth inoculation strategies rather than simple information priming can more effectively immunize participants with a higher tendency to trust faked online content.[61] Second, in line with the idea that fake-content recognition could be triggered by a top-down recognition process due to disclosure characteristics,[44,55] future research should build on prior disclosure studies to support practitioners in developing disclosure labels to signal such misinformative content.

Finally, as recognition of disinformation is fundamental for users to activate the coping strategies that allow them to react to DFs, we confirm that information priming elicited higher DF recognition, and this recognition triggered resistance to persuasion in line with the PKM. In fact, our results highlight how priming with an understanding of DF helps in developing knowledge to contend with such misleading content. Moreover, consistent with the strong attitude–behavior relationship posited by the TPB, having recognized the manipulative attempt, users develop a more negative attitude toward the video, followed by a lower sharing intention.

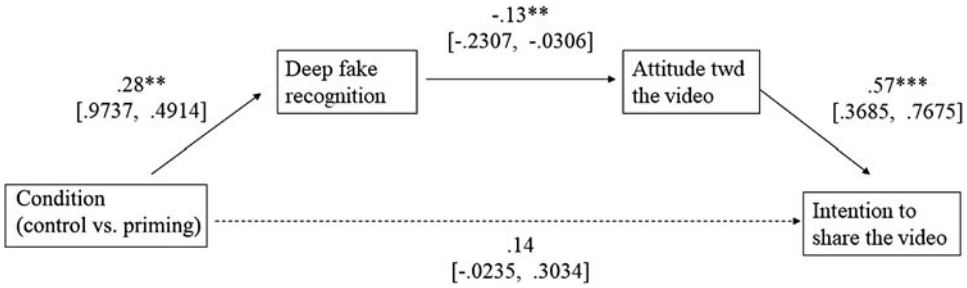


FIG. 5. Mediation model with values indicating the unstandardized path coefficients. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 2. Direct and Indirect Effects

	b (SE)	*Lower 95% BCBCI*	*Upper 95% BCBCI*
Direct effects			
Condition → DF recognition	0.28*** (0.11)	0.0737	0.4914
Condition → Attitude toward the video	0.02 (0.07)	–0.1084	0.1568
Condition → Intention to share the video	0.14 (0.08)	–0.0235	0.3034
DF recognition → Attitude toward the video	–0.13* (0.05)	–0.2307	–0.0306
DF recognition → Intention to share the video	–0.23*** (0.06)	–0.3529	–0.1009
Attitude toward the video → Intention to share the video	0.57*** (0.10)	0.3685	0.7675
Indirect effects			
Condition → DF recognition → Attitude toward the video	–0.04* (0.02)	–0.0849	–0.0015
Condition → DF recognition → Intention to share the video	–0.06* (0.03)	–0.1291	–0.0137
Condition → Attitude toward the video → Intention to share the video	0.01 (0.04)	–0.0634	0.0906
DF recognition → Attitude toward the video → Intention to share the video	–0.07** (0.03)	–0.1413	–0.0111

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.
Condition: priming versus control.
$N = 152$.
Unstandardized *b* coefficients (with boot *SE* between parentheses).
BCBCI, bias-corrected 5,000 bootstrap confidence intervals; DF, deepfake.

In the online environment, incentives to share emotionally arousing and triggering content tend to outweigh the incentives to perform tiresome fact-checking work[62]—hence, sharing intention is one of the drivers of DF virality.[63] Virality represents one of the main reasons behind, in turn, DFs' success and the turmoil they cause; this final result is very interesting, as we confirm that recognition through prior knowledge can finally lead to lower dissemination potential for DF content. In this vein, again, focus on users' education and awareness building is confirmed to be an effective strategy against DF deception. However, we did not specifically disambiguate between the positive or negative nature of this sharing intention: interestingly, an intention to share the DF content accompanied with a negative or educational message from an aware user could work as electronic word of mouth at the service of other users' knowledge.

A further limitation might be that we somewhat disregarded the impact of source credibility on sharing intention. We provided all users with a general indication that the content has been shared by a friend, thus appearing on the participants' Facebook newsfeeds; however, no other indication was given about the credibility of this source. Further studies should investigate source credibility as a determinant of fake-content recognition, as well as how users' experience of having been victim of a hoax impacts the perceived credibility of the source.

Finally, although a set of confound checks reassured us about the potential confounding effect of gender, we report sample-related limitations that we hope might be of interest for further research on this topic. Female participants represented 70 percent of our sample; therefore, further investigation—not only with more equally balanced samples, but also with the direct aim of examining the effects of gender on BS receptivity and DF recognition—is highly desirable, particularly in light of prior findings that women are more likely to rate BS statements as both are more truthful and more profound than are men.[64]

Overall, we believe we have contributed to the literature by highlighting that DFs represent an undoubted threat to online users, although we encourage scholars to further investigate users' education on the issue, as priming with knowledge about the DF threat seems a fruitful strategy to debunk and counter its negative effects.

Note

a. https://www.youtube.com/watch?v=HG_NZpkttXE (accessed Sept. 2, 2020).

Authors' Contributions

S.I. conceived the idea of this article; S.I. and R.D.C. collected and analyzed the data; S.I., R.D.C., and F.M. drafted the article; and F.M., R.P., and S.P. contributed by interpreting the results and finalizing the article.

Acknowledgments

The authors express their gratitude to the two anonymous reviewers who contributed in improving the first version of this article with their valuable comments, and to Miloš Borozan for his writing assistance and for carefully proofreading the first version of the article.

Authors Disclosure Statement

No competing financial interests exist.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Supplementary Material

Supplementary Table S1

References

1. Albright J. Welcome to the era of fake news. Media and Communication 2017; 5:87–89.
2. Brundage M, Avin S, Clark J. et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation 2018; arXiv preprint arXiv:1802.07228.

3. Qayyum A, Qadir J, Janjua MU, et al. Using blockchain to rein in the new post-truth world and check the spread of fake news. IT Professional 2019; 21:16–24.
4. Zannettou S, Sirivianos M, Blackburn J, et al. The web of false information: rumors, fake news, hoaxes, clickbait, and various other shenanigans. Journal of Data and Information Quality 2019; 11:1–37.
5. Oxford Languages, Word of the Year 2016. [online] languages.oup.com. https://languages.oup.com/word-of-the-year/2016 (accessed Feb. 17, 2020).
6. Zhang X, Ghorbani AA. An overview of online fake news: characterization, detection, and discussion. Information Processing & Management 2020; 57:1–26.
7. Jones-Jang SM, Mortensen T, Liu J. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. American Behavioral Scientist 2021; 65:371–388.
8. Leetaru K. A Reminder That 'Fake News' Is Information Literacy Problem—Not A Technology Problem. [online] Forbes.com. https://www.forbes.com/sites/kalevleetaru/2019/07/07/a-reminder-that-fake-news-is-an-information-literacy-problem-not-a-technology-problem/#737681166a6f (accessed Feb. 17, 2020).
9. Flynn DJ, Nyhan B, Reifler J. The nature and origins of misperceptions: understanding false and unsupported beliefs about politics. Political Psychology 2017; 38:127–150.
10. Dack S. Deep Fakes, Fake News, and What Comes Next 2019. [online] Isis.washington.edu. https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/ (accessed Feb. 17, 2020).
11. Kietzmann J, Lee LW, McCarthy IP, et al. Deepfakes: trick or treat? Business Horizons 2020; 63:1–12.
12. University College London. 'Deepfakes' ranked as most serious AI crime threat. [online] ScienceDaily. https://www.sciencedaily.com/releases/2020/08/200804085908.htm (accessed Aug. 31, 2020).
13. Spivak R. ''Deepfakes'': the newest way to commit one of the oldest crime. Georgetown Law Technology Review 2019; 3:339–400.
14. Albahar M, Almalki J. Deepfakes: threats and countermeasures systematic review. Journal of Theoretical and Applied Information Technology 2019; 97:3242–3250.
15. Ibrahim S. Social and contextual taxonomy of cybercrime: socioeconomic theory of Nigerian cybercriminals. International Journal of Law, Crime and Justice 2016; 47:44–57.
16. Lazarus S. Just married: the synergy between feminist criminology and the Tripartite Cybercrime Framework. International Social Science Journal 2019; 69:15–33.
17. Westerlund M. The emergence of deepfake technology: a review. Technology Innovation Management Review 2019; 9:39–52.
18. Chesney R, Citron D. Deepfakes and the new disinformation war: the coming age of post-truth geopolitics. Foreign Affairs 2019; 98:147–155.
19. Ajder H, Patrini G, Cavalli F, et al. The State Of Deepfakes: landscape, Threats and Impact. [online] Deeptracelabs.com. https://deeptracelabs.com/resources/ (accessed Feb. 17, 2020).
20. Thomas E. In the battle against deepfakes, AI is being pitted against AI. [online] Wired.co.uk. https://www.wired.co.uk/article/deepfakes-ai (accessed Feb. 17, 2020).
21. Silbey J, Hartzog W. The upside of deep fakes. Maryland Law Review 2019; 78:960–966.
22. Vincent J. Deepfake detection algorithms will never be enough. [online] Theverge.com. https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work (accessed Feb. 17, 2020).
23. Hanbury M. A deepfake pioneer says 'perfectly real' manipulated videos are just 6 months away. [online] Businessinsider.com. https://www.businessinsider.com/perfectly-real-deepfake-videos-6-months-away-deepfake-pioneer-says-2019-9?IR=T (accessed Feb. 17, 2020).
24. Meskys E, Liaudanskas A, Kalpokiene J, et al. Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice 2020; 15:24–31.
25. The antivirus for deepfakes [online] Deeptracelabs.com. https://deeptracelabs.com. (accessed Feb. 17, 2020).
26. Longo A. Raddoppia il numero di video deepfake. Nel mirino le aziende. [online] Ilsole24ore.com. https://www.ilsole24ore.com/art/raddoppia-numero-video-deepfake-mirino-aziende-ACDBInp (accessed Feb. 17, 2020).
27. McGuire WJ. (1964) Inducing resistance to persuasion: some contemporary approaches. In Berkowitz L, ed. *Advances in experimental social psychology*. 1st ed. New York, NY: Academic Press, pp. 191–229.
28. Pennycook G, Cheyne JA, Barr N, et al. On the reception and detection of pseudo-profound bullshit. Judgment and Decision Making 2015; 10:549–563.
29. Ajzen I. (1985) From intentions to actions: a theory of planned behavior. In Kuhl J, Beckmann J, eds. *Action control: from cognition to behavior.* Heidelberg: Springer, pp. 11–39.
30. McNamara TP. (2005) *Semantic priming: perspectives from memory and word recognition.* New York: Psychology Press.
31. Higgins ET, Bargh JA, Lombardi WJ. Nature of priming effects on categorization. Journal of experimental psychology: Learning, Memory, and Cognition 1985; 11:59–69.
32. McGuire WJ, Papageorgis D. The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. The Journal of Abnormal and Social Psychology 1961; 62:327–337.
33. Yi, Y. The effects of contextual priming in print advertisements. Journal of Consumer Research 1990; 17:215–222.
34. Wu M, Huang Y, Li R, et al. A tale of two sources in native advertising: examining the effects of source credibility and priming on content, organizations, and media evaluations. American Behavioral Scientist 2016; 60:1492–1509.
35. Van Duyn E, Collier J. Priming and fake news: the effects of elite discourse on evaluations of news media. Mass Communication and Society 2019; 22:29–48.
36. Walker AC, Turpin MH, Stolz JA, et al. Finding meaning in the clouds: illusory pattern perception predicts receptivity to pseudo-profound bullshit. Judgment and Decision Making 2019; 14:109–119.
37. Frankfurt H. (2005) *On bullshit*. Princeton: Princeton University Press.
38. Dalton C. Bullshit for you; transcendence for me. A commentary on ''On the reception and detection of pseudo-profound bullshit.'' Judgment and Decision Making 2016; 11:121–122.
39. Erlandsson A, Nilsson A, Tinghög G, et al. Bullshit-sensitivity predicts prosocial behavior. PLoS One 2018; 13: 1–12.
40. Pennycook G, Rand DG. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of Personality 2020; 88:185–200.
41. Friestad M, Wright P. The persuasion knowledge model: how people cope with persuasion attempts. Journal of Consumer Research 1994; 21:1–31.

42. Amazeen MA, Bucy EP. Conferring resistance to digital disinformation: the inoculating influence of procedural news knowledge. Journal of Broadcasting & Electronic Media 2019; 63:415–432.
43. Nelson MR, Wood ML, Paek HJ. Increased persuasion knowledge of video news releases: audience beliefs about news and support for source disclosure. Journal of Mass Media Ethics 2009; 24:220–237.
44. Evans NJ, Park D. Rethinking the persuasion knowledge model: schematic antecedents and associative outcomes of persuasion knowledge activation for covert advertising. Journal of Current Issues & Research in Advertising 2015; 36:157–176.
45. Campbell MC, Kirmani A. (2008) I know what you're doing and why you're doing it: the use of the persuasion knowledge model in consumer research. In Haugtvedt CP, Herr P, Kardes FR, eds. *The handbook of consumer psychology*. New York, NY: Lawrence Erlbaum Associates, pp. 549–573.
46. Friestad M, Wright P. Everyday persuasion knowledge. Psychology & Marketing 1999; 16:185–194.
47. Youn S, Kim S. Newsfeed native advertising on Facebook: young millennials' knowledge, pet peeves, reactance and ad avoidance. International Journal of Advertising 2019; 5: 651–683.
48. Ortmann A, Hertwig R. The costs of deception: evidence from psychology. Experimental Economics 2002; 5:111–131.
49. Campbell MC, Mohr GS, Verlegh PWJ. Can disclosures lead consumers to resist covert persuasion? The important roles of disclosure timing and type of responses. Journal of Consumer Psychology 2013; 23:483–495.
50. Shrum LJ, Liu M, Nespoli M, et al. (2012) Persuasion in the marketplace: how theories of persuasion apply to marketing and advertising. In Dillard J, Shen L, eds. *The persuasion handbook*. Thousand Oaks, CA: Sage, pp. 314–330.
51. Tutaj K, van Reijmersdal EA. Effects of online advertising format and persuasion knowledge on audience reactions. Journal of Marketing Communications 2012; 18:5–18.
52. Wentzel D, Tomczak T, Herrman A. The moderating effect of manipulative intent and cognitive resources on the evaluation of narrative ads. Psychology & Marketing 2010; 27:510–530.
53. Nyilasy G. Fake news: when the dark side of persuasion takes over. International Journal of Advertising, 2019; 38: 336–342.
54. Iacobucci S, De Cicco R. Users awareness of native advertising from Instagram media publishers: the effects of Instagram's branded content tool on attitudes and behavioural intent. International Journal of Internet Marketing and Advertising 2020; 14:71–90.
55. Amazeen MA, Wojdynski BW. Reducing native advertising deception: revisiting the antecedents and consequences of persuasion knowledge in digital news contexts. Mass Communication and Society 2018; 22:222–247.
56. Bruner GC, Hensel PJ, James KE. *Marketing scales handbook: a compilation of multi-item measures*. Chicago: American Marketing Association, 2001.
57. Hsieh JK, Hsieh YC, Tang YC. Exploring the disseminating behaviors of eWOM marketing: persuasion in Online Video. Electronic Commerce Research 2012; 12:201–224.
58. Bainbridge TF, Quinlan JA, Mar RA, et al. Openness/intellect and susceptibility to pseudo-profound bullshit: a replication and extension. European Journal of Personality 2019; 33:72–88.
59. Hayes AF. (2017) *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*. New York: The Guilford Press.
60. Pfattheicher S, Schindler S. Misperceiving bullshit as profound is associated with favorable views of cruz, rubio, trump and conservatism. PLoS One 2016; 11:1–7.
61. Roozenbeek J, Van Der Linden S. The fake news game: actively inoculating against the risk of misinformation. Journal of Risk Research 2019; 22:570–580.
62. Parkin S. Politicians fear this like fire: The rise of the deepfake and the threat to democracy. [online] Theguardian .com. https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy (accessed Feb. 17, 2020).
63. Kim JW. Rumor has it: the effects of virality metrics on rumor believability and transmission on Twitter. New Media & Society 2018; 20:4807–4825.
64. Čavojová V, Brezina I, Jurkovič M. Expanding the bullshit research out of pseudo-transcendental domain. Current Psychology 2020; 1–10.

Address correspondence to:
Dr. Serena Iacobucci
Department of Neurosciences, Imaging and Clinical Sciences
University of Chieti-Pescara
Via Luigi Polacchi 11
Chieti 66100
Italy

E-mail: serena.iacobucci@unich.it