The Psychology of Deception: Why We Believe Deepfakes

Nejc Plohl,[1] Urška Smrke,[2]
Letizia Aquilino,[3, 4] Izidor Mlakar[2]

[1] University of Maribor, Faculty of Arts, Maribor, Slovenia
nejc.plohl1@um.si
[2] University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
urska.smrke@um.si, izidor.mlakar@um.si
[3] DEXAI – Artificial Ethics, Rome, Italy
[4] Università Cattolica Del Sacro Cuore, Milan, Italy
letizia.aquilino@dexai.eu

DOI
https://doi.org/10.18690/um.feri.2.2026.4

ISBN
978-961-299-109-8

Advances in artificial intelligence have enabled the creation of highly realistic deepfakes, yet their impact ultimately depends on how humans perceive and interpret them. This chapter examines the psychological processes underlying belief in deepfakes, focusing on perceptual mechanisms, individual differences, and downstream consequences. Despite widespread confidence in detection abilities, people generally struggle to distinguish authentic from manipulated videos, often performing at or near chance. To move beyond binary detection measures, we introduce the construct of perceived trustworthiness, defined as the extent to which a video is experienced as authentic. We describe the development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ), which captures two dimensions: trustworthiness of content (plausibility and source credibility) and trustworthiness of presentation (perceived realism of delivery, including technical quality, voice, and behaviour). This tool enables systematic examination of perceptual features that make deepfakes believable across contexts. We further show how sociodemographic, motivational, and cognitive factors shape susceptibility, and demonstrate that perceived trustworthiness predicts attitudes toward climate change and immigration as well as intentions to share content. Overall, the chapter highlights the need for psychological, not only technological, interventions.

Keywords:
deepfakes, psychology, trustworthiness, individual differences, attitudes

University of Maribor Press

1 Introduction

While chapter 1 introduced the technical layers of deepfakes, explaining how advances in artificial intelligence, machine learning, deep learning, and generative adversarial networks make it possible to create hyper-realistic synthetic media, technology is only half of the story. The other half lies in human perception, specifically in how we see, interpret, and, in the end, decide whether to believe what is placed before our eyes and ears. No matter how sophisticated a deepfake's creation process is, its final impact depends on the processes occurring within the person encountering it. However, these perceptual and cognitive processes depend on broader individual characteristics and are particularly complex in the context of multimodal media, making it difficult to fully grasp why people come to believe deepfakes.

Specifically, how we judge a video's authenticity is not shaped solely by the sensory information it provides, but also by who we are as individuals, our prior knowledge, worldviews, cognitive styles, and even habitual media use (Somoray et al., 2025). Two people can watch the same deepfake and come away with very different conclusions, depending on factors such as political orientation, trust in institutions, or media literacy. This highlights the importance of individual differences, which interact with perceptual processes to shape how a given deepfake is received and interpreted.

Second, the challenge is compounded by the fact that deepfakes are multimodal, targeting several channels of human perception simultaneously (Lee & Shin, 2022). They can look real, sound real, and convey a message we are already predisposed to accept. This convergence of visual, auditory, and semantic cues can create a powerful sense of authenticity, making it harder for viewers to engage in critical evaluation. Even when technical imperfections are present (e.g., slightly unnatural facial movements, subtle audio mismatches), a coherent and plausible message can override scepticism, fostering misplaced but compelling trust. Understanding how these different pathways interact, and how they are related to individual characteristics, is essential for building a comprehensive account of why people believe deepfakes and how they can be influenced by them. Crucially, such influence is not limited to the moment of exposure; perceptions of authenticity can shape downstream psychological outcomes (Rijo & Waldzus, 2023), including changes in attitudes toward the depicted topic and intentions to engage with or share the

content. These behavioural consequences, ranging from private opinion shifts to the viral spread of misinformation, make the study of deepfake perception a matter of detection accuracy and of understanding their broader persuasive power.

In the present chapter, we hence focus on the human aspect of deepfakes, with a particular emphasis on the advancements made within the SOLARIS project (which are presented in detail in our research articles; Plohl et al., 2024, 2025a, 2025b, 2025c). We start by reviewing the key literature on human detection of deepfakes. Next, we move beyond detection and introduce the concept of perceived trustworthiness of deepfakes to provide some insight into the perceptual elements of deepfakes that make people more or less inclined to believe them. We then accompany these perceptual aspects with broader individual characteristics, which may contribute to individuals' susceptibility to deepfakes. Lastly, we finish the chapter with a brief section on why deepfake detection and perceived trustworthiness matter. Altogether, the chapter provides a brief but comprehensive insight into the psychological processes underlying how people perceive and respond to deepfakes, highlighting both perceptual and individual factors that shape susceptibility and resistance.

2 Do We Actually Believe Deepfakes?

People generally believe that they can reliably detect deepfakes and overestimate their performance in deepfake detection tasks (e.g., Köbis et al., 2021; Somoray & Miller, 2023), which is particularly true for those who actually perform the worst in such tasks (Plohl et al., 2025c), illustrating a phenomenon called the Dunning-Kruger effect (Kruger & Dunning, 1999). However, in reality, the existing studies suggest that we are generally bad at recognizing whether the video is real or manipulated. For example, Köbis and colleagues (2021) exposed participants to 16 videos lasting about 10 seconds and found the overall accuracy level to be 57.6%, just slightly above what would be achieved with coin-tossing (50.0%). Similarly, another recent study (Somoray & Miller, 2023) found the mean categorization accuracy of 20 videos lasting 10 seconds to be 60.7%, which, again, only slightly exceeded chance levels. Moreover, our recently conducted study revealed that detection accuracy varies based on deepfake quality, manipulated by (mis)aligning the content of the message with the depicted person's actual stance on the topic and changing the technical proficiency (e.g., voice quality, lip-syncing). In this study, 43.5-60.4% of individuals correctly identified lower-quality deepfakes (characterized

by misaligned content and low technical proficiency), whereas higher-quality deepfakes (characterized by aligned content and high technical proficiency) were correctly detected only by about a third of participants (30.9-36.6%; Plohl et al., 2025b).

The findings of individual studies have recently been summarized in a comprehensive systematic review investigating deepfake detection. Diel and colleagues (2024) synthesized the evidence on the human ability to detect deepfakes of different modalities, including audio, image, and video. They found 56 studies involving more than 86,000 participants that involved some kind of deepfake stimuli and detection performance measures (which varied between the studies). They found the total deepfake detection accuracy of 55.5% (audio: 62.1%, images: 53.2%, video: 57.3%), which is not significantly above the chance level. Similar results emerged for other metrics beyond analyses of proportions. Hence, the available evidence suggests that individuals' decisions regarding video authenticity are close to decisions one would make by blind guessing, with detection accuracy likely facing additional challenges once deepfakes become more and more sophisticated.

3 Moving Beyond Detection to Understand Why We Believe Deepfakes

Focusing solely on detection and employing simple dichotomous questions asking whether a video is real or a deepfake offers an interesting insight into the extent to which people may believe deepfakes. However, such research cannot convincingly answer how these judgments are formed, or, in other words, why people believe deepfakes. To address this gap, we proposed a new construct, "perceived trustworthiness of deepfakes', defined as the extent to which individuals perceive deepfakes as authentic (i.e., not fabricated). From the beginning, perceived trustworthiness was hypothesized to be multidimensional, consisting of various aspects that may contribute to deepfakes being perceived as more or less trustworthy. Due to specific aspects determining these perceptions not being well-understood and the lack of measures capable of capturing this newly-proposed construct, we set out to develop a new scale by employing a complex process combining various methodologies (i.e., qualitative and quantitative research), stakeholders (i.e., experts and general population), and cultural backgrounds (i.e., participants from the United Kingdom, Italy, and Slovenia).

Specifically, the development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ; Plohl et al., 2024) occurred in three phases to ensure the scale's validity and conceptual depth. The first phase was dedicated to the development of the initial pool of items. We reviewed the literature to collect items from existing relevant scales (e.g., Hameleers et al., 2024; Hwang et al., 2021; Lee & Shin, 2022) and generate new items based on aspects identified as important in previous, mostly qualitative, studies, such as blurriness on the eye region, abnormal mouth movements, and unnatural voice (e.g., Hameleers et al., 2023; Tahir et al., 2021; Thaw et al., 2021). Furthermore, we conducted face-to-face interviews with students and an online survey with citizens, journalists, and experts. In both interviews and the online survey (overall $N = 26$), participants were asked to watch multiple videos, some of which were deepfakes, decide whether they trust each of them, and share all the thoughts that popped into their heads while forming these decisions. The relevant statements collected qualitatively were transformed into questionnaire items. Lastly, we generated additional items using the Psychometric Item Generator (Götz et al., 2023), a machine-learning solution to developing items for psychometric scales. Altogether, the first phase resulted in 419 initial items.

After reducing the number of items by only keeping those that were unique and general enough (i.e., suitable for different deepfake videos), 123 items were reviewed by 13 experts for content validity. Specifically, the experts were asked to assess the relevance and clarity through a classic content validity procedure. For each item, we then calculated the content validity ratio (a measure of relevance) and content validity index (a measure of clarity), with only items above the acceptable thresholds being retained further. This procedure resulted in a 31-item version covering key dimensions such as the content of the video, the behaviour of the person in the video, the video's source, and its technical features. The items were then translated into Italian and Slovene using the translation-back translation procedure.

In the last step, we conducted large-scale surveys across English, Italian, and Slovene samples ($N = 733$) to investigate the factorial structure of the questionnaire, measurement equivalence of the three language versions, internal reliability of the questionnaire, construct validity, and incremental validity. The results of exploratory and confirmatory factor analyses supported a two-factor structure of the final 22-item scale, consisting of perceived trustworthiness of content (i.e., evaluations of the presented information and its source; 11 items) and perceived trustworthiness of presentation (i.e., evaluations of how the information is presented, including the

speaker's behaviour and the video's technical sophistication; 11 items). For instance, a deepfake of a politician delivering factual information aligned with what they usually advocate for may score high on content trustworthiness but low on presentation trustworthiness if the lip-syncing is misaligned. In addition, we found support for configural and metric invariance across the three languages, suggesting that the factor structure and factor loadings are similar across different versions of the questionnaire.

The scale demonstrated strong psychometric properties, including high reliability (α = .83–.92). Moreover, construct and incremental validity analyses confirmed that PDTQ scores relate meaningfully to some of the established correlates of misinformation susceptibility (reviewed in section 4) and predict relevant behavioural outcomes beyond existing measures (reviewed in section 5). Taken together, these results position the PDTQ as a psychometrically robust, multilingual instrument for studying perceived trust in deepfakes across diverse contexts. The final English version of the scale can be seen in Table 1.

Table 1: English version of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ)

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
1.The presented information seemed convincing.	1	2	3	4	5	6	7
2.The mouth movements of the person in the video did not completely match the sound.	1	2	3	4	5	6	7
3.The background in the video contained irrelevant or out-of-place objects.	1	2	3	4	5	6	7
4.The presented information seemed plausible.	1	2	3	4	5	6	7
5.I found the voice of the person in the video unnatural.	1	2	3	4	5	6	7
6.I found the voice of the person in the video to be different from their usual voice.	1	2	3	4	5	6	7

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
7.The audio was low quality.	1	2	3	4	5	6	7
8.The presented information was something that I already know to be true.	1	2	3	4	5	6	7
9.The source of the video is verified in some way.	1	2	3	4	5	6	7
10.The facial features of the person in the video changed during the video.	1	2	3	4	5	6	7
11.The person's gestures in the video did not seem natural.	1	2	3	4	5	6	7
12.The video quality was inconsistent.	1	2	3	4	5	6	7
13.The source of the video is well-known.	1	2	3	4	5	6	7
14.The face of the person in the video (or parts of it) was distorted.	1	2	3	4	5	6	7
15.The presented information was consistent with my previous knowledge.	1	2	3	4	5	6	7
16.The source of the video seems credible.	1	2	3	4	5	6	7
17.The mouth of the person in the video was moving strangely.	1	2	3	4	5	6	7
18.The presented information seemed questionable.	1	2	3	4	5	6	7
19.The face of the person in the video (or parts of it) was blurry.	1	2	3	4	5	6	7
20.The content of the video is consistent with what this source has published previously.	1	2	3	4	5	6	7

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
21.The video was posted by a reputable source.	1	2	3	4	5	6	7
22.The presented information seemed credible.	1	2	3	4	5	6	7

Instructions: The following questionnaire contains items that aim to capture your perception of the video you just watched. Please read each item carefully and indicate your agreement using a 7-point scale ranging from »Strongly disagree« to »Strongly agree«. If you feel that you cannot answer a particular item, please choose »Neutral«.

Scoring key (R denotes that the item needs to be reverse-coded): Trustworthiness of content = (I1+I4+I8+I9+ I13+I15+I16+I18R+I20+I21+I22)/11. Trustworthiness of presentation = (I2R+I3R+I5R+I6R+I7R+I10R+ I11R+I12R+I14R+I17R+I19R)/11.

4 Beyond the Video: How Individual Differences Shape Deepfake Perception

While individuals' perception of deepfakes is a good starting point, any answers to why people believe deepfakes are incomplete without taking into account individual differences. In other words, perceived trustworthiness of deepfakes does not exist in a vacuum; instead, as demonstrated by the fact that the same videos can be perceived vastly differently by different individuals, our perception of videos is heavily influenced by our past experiences (i.e., sociodemographic variables), worldviews (i.e., motivational variables), and knowledge (i.e., cognitive variables). These factors have previously been extensively investigated in the broader misinformation context, whereas research on how they operate in the context of deepfakes and how they are specifically associated with each of the two dimensions of perceived deepfake trustworthiness is only beginning to emerge.

Starting with sociodemographic variables, previous literature has revealed that age and social media use may be important in the context of misinformation (van der Linden, 2022). In our studies, age was significantly positively associated with individuals' judgments regarding the trustworthiness of deepfakes, their content, and presentation. In other words, older individuals were more inclined to trust manipulated videos (Plohl et al., 2024). On the other hand, the frequency of using social media as a source of news was positively associated with the perceived trustworthiness of content but not the perceived trustworthiness of presentation (Plohl et al., 2024), meaning that repeated social media use may make individuals more vulnerable to questionable arguments, but may not be related to their ability to discern authentic video presentations from the manipulated ones.

Based on various theories, such as the theory of motivated reasoning, which explains that decisions are often based on pre-determined goals and desirability rather than an accurate reflection of the evidence (Kunda, 1990), researchers have identified a few individual variables that may motivate the person to believe misinformation they are exposed to. These include political orientation (Chen et al., 2023; van der Linden, 2022), belief in conspiracy theories, and trust in institutions such as media, when the media at hand is not reliable (Chen et al., 2023). Our study (Plohl et al., 2024) suggests that the importance of these factors translates to the deepfake context to some degree, but that there is an additional complexity to judging deepfakes due to their multimodal nature. Specifically, conservatism was positively associated with the perceived trustworthiness of deepfake content but was not associated with the perceived trustworthiness of presentation at all, demonstrating informational bias but no difference in deepfake recognition skills pertaining to their presentation and technical aspects.

Additionally, our unpublished results, obtained during the validation study, showed no association between conspiracy beliefs and the two dimensions measuring the perceived trustworthiness of deepfakes. As such, the role of conspiracy mentality in the perception of deepfakes remains relatively unclear. It is likely that this variable is highly context-specific; in general, it may increase distrust in the presented information, however, when deepfakes advocate for conspiracy theories, it may increase perceived trustworthiness. Lastly, in our study, trust in media was significantly positively associated with the perceived trustworthiness of deepfake content but not the perceived trustworthiness of deepfake presentation. It hence seems likely that trust in media represents a double-edged sword; trust is a necessary ingredient in communication, facilitating the spread of credible information, but, when unwarranted, it may make individuals more vulnerable to deception – a phenomenon known as misplaced trust (O"Brien et al., 2021).

In addition to demographic and motivational variables, previous research has also explored the role of cognitive abilities and other related variables. The so-called inattention account posits that being bombarded with information, coupled with limited time and resources, interferes with individuals' ability to accurately reflect on the content (van der Linden, 2022). In line with this, previous research has found that education, media literacy, reflective thinking (i.e., ability to suppress intuition and cognitively reflect when making decisions; Frederick, 2005), and so-called "bullshit receptivity" (i.e., ascribing profundity to randomly generated sentences;

Pennycook & Rand, 2019) are relatively consistently associated with the processing of misinformation, even when the content is congruent with individuals' pre-existing beliefs (Roozenbeek et al., 2020; van der Linden, 2022). In our study, we found that education was not significantly associated with the perceived trustworthiness of deepfake content or presentation. In contrast, we found significant associations between media literacy, reflectiveness, and "bullshit receptivity" on one side and the trustworthiness of content on the other side, with "bullshit receptivity" emerging as a particularly strong contributing factor. However, none of these cognitive variables were significantly associated with the trustworthiness of the presentation. The only cognitive variable significantly (albeit weakly) related to the perceived trustworthiness of presentation, not just content, was specific deepfake knowledge (Plohl et al., 2024). This suggests that while general cognitive tendencies shape how individuals evaluate the credibility of content, knowledge specific to deepfakes plays a uniquely important role in shaping perceptions of their presentation.

Table 2: A summary of factors associated with perceived trustworthiness

Category	Potential factor	Perceived trustworthiness of content	Perceived trustworthiness of presentation
Demographic variables	Higher age	✓ (↑ Risk)	✓ (↑ Risk)
	Higher social media use	✓ (↑ Risk)	X
Motivational variables	Higher political conservatism	✓ (↑ Risk)	X
	Higher belief in conspiracy theories	X	X
	Higher trust in media	✓ (↑ Risk)	X
Cognitive variables	Higher education	X	X
	Higher media literacy	✓ (↓ Risk)	X
	Higher reflective thinking	✓ (↓ Risk)	X
	Higher "bullshit receptivity"	✓ (↑ Risk)	X
	Higher deepfake knowledge	✓ (↑ Risk)	✓ (↓ Risk)

Source: Plohl et al. (2024).

As shown in Table 2, our results suggest that many known correlates of misinformation susceptibility are also relevant in the context of deepfakes. In line with this, deepfakes may disproportionally affect older individuals who use social media to a greater extent, are more conservative, trust (media) to a higher degree, have lower media literacy, are less reflective, and are more receptive to finding meaning in pseudo-profound information. The use of our scale offers additional

insights. While more studies are needed, most of these factors are consistently associated with individuals' perception of the messages conveyed in deepfakes but not so much with their perception of deepfakes' presentation, which includes paying attention to the person in the video and technical aspects. In fact, only age (risk factor) and deepfake knowledge (protective factor) were associated with the perceived trustworthiness of deepfakes' presentation.

5 When Trust Turns into Influence: The Role of Perceived Trustworthiness in Shaping Attitudes and Intentions

In the previous sections, we established that people are generally bad at detecting deepfakes and provided some insight into why this is so (i.e., due to their perceptions of content and presentation, as well as demographic, motivational, and cognitive individual differences). As we approach the end of the chapter, it is worth noting why low detection and, specifically, perceived trustworthiness of deepfakes matter beyond just providing a better understanding of individuals' perception of deepfakes. We will specifically focus on associations with attitudes (i.e., psychological tendencies expressed by evaluating a particular entity with some degree of favour or disfavour; Eagly & Chaiken, 1993) and behavioural intentions (i.e., individuals' intention to perform a given act; Ajzen & Fishbein, 1972) - two outcomes related to behaviour (Ajzen, 1991).

One of our studies showed that low detection, across various deepfake videos, led to more favourable affective responses to videos (i.e., higher liking), which, in turn, led to increased intentions to share the manipulated videos on social media (Plohl et al., 2025b). Similar associations were found between sharing intentions and perceived trustworthiness of deepfakes, with these results offering additional insight into the complex relationship between variables. Specifically, in the original PDTQ validation study (Plohl et al., 2024), we investigated whether perceived trustworthiness of content and presentation explain variance in viral behavioural intentions (i.e., the intentions to like, share, and recommend the video) beyond basic demographic variables (i.e., age, education, political conservatism, social media use), individual differences (i.e., "bullshit receptivity", reflectiveness, trust in media, media literacy, deepfake knowledge), and a previous scale measuring participants' perception of the manipulated video (i.e., Message Believability Scale; Hameleers et al., 2023). We found that the newly developed scale explained a significant part of the variance (an additional 5.0%) in viral behavioural intentions over and above

other included variables. In the final model, which was able to explain 36.0% of the variance, age, "bullshit receptivity", reflectiveness, trust in media, deepfake knowledge, message believability, and trustworthiness of content, which was the strongest predictor, significantly predicted the outcome. Other variables, including the trustworthiness of the presentation, did not significantly predict viral behavioural intentions. These results suggest that individuals' intention to spread the videos may be particularly driven by the trustworthiness of the content. Nonetheless, the questionnaire explained a significant additional share of variance, highlighting the added value of a more comprehensive measurement of deepfake perception.

The importance of these perceptions was further demonstrated in our experimental study (Plohl et al., 2025a), which examined the potential positive or negative effects of a single exposure to deepfake or authentic videos on individuals' attitudes toward climate change and immigration, two highly polarized, politically sensitive issues (Doss et al., 2022; Hameleers et al., 2022; Westerlund, 2019). Specifically, the study explored boundary conditions under which attitude change might occur, with a focus on video quality, perceived trustworthiness, and political alignment.

A total of 1,124 participants from the United Kingdom, Italy, and Slovenia watched real videos, high-quality deepfakes, or low-quality deepfakes advocating for or against climate action and immigration (Figure 1). The quality of videos was manipulated in terms of the content and presentation. For example, manipulations of content included changing the supposed source of the video and making the presented information more or less aligned with the target person's actual stance on the topic. In contrast, manipulations of presentation included alterations of mouth movements, voice, and video quality. All videos lasted approximately one minute and featured well-known proponents or opponents of climate change and immigration. Participants provided their demographic data and filled out the PDTQ (Plohl et al., 2024) directly after watching each of the two videos, whereas the Scepticism scale (a measure of attitudes towards climate change; Whitmarsh, 2011) and the Positive and Negative Perception of Immigrants Scale (a measure of attitudes towards immigration; Panno et al., 2023) were filled out before and after video exposure.

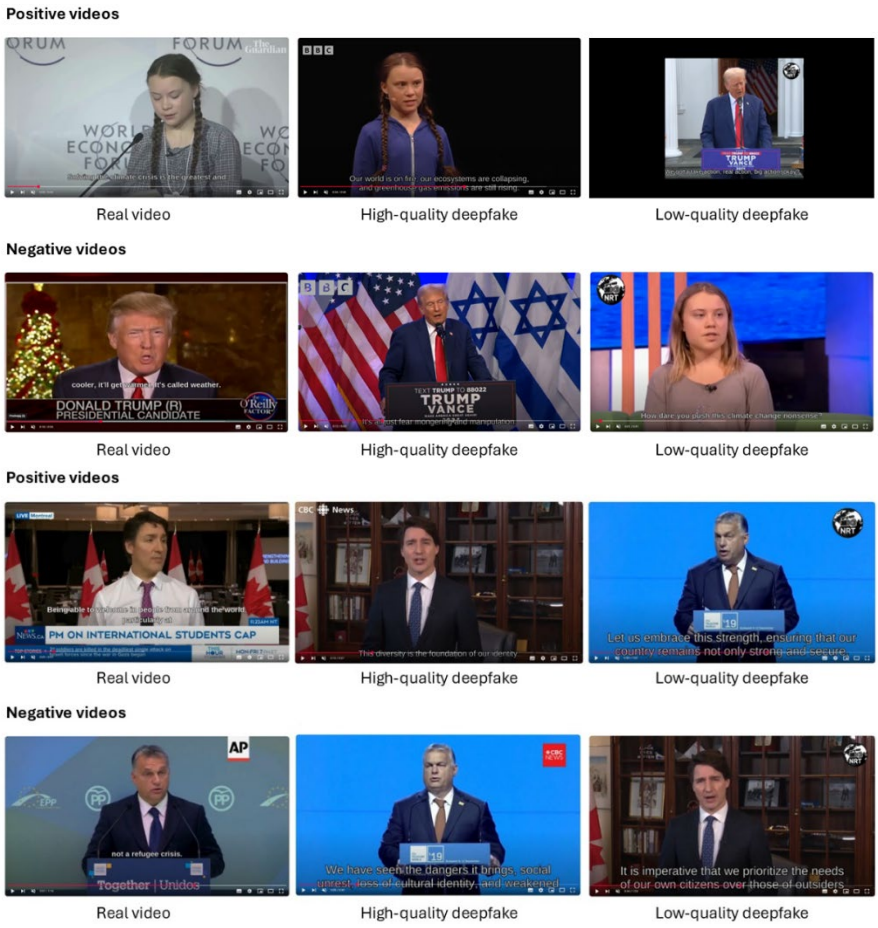

Figure 1: Stimuli related to climate action (first two rows) and immigration (last two rows)
Source: own.

Contrary to expectations, neither video authenticity/quality nor political orientation moderated the impact of the videos on attitudes. On the other hand, perceived trustworthiness of deepfake content consistently predicted attitude change across both topics, while perceived presentation trustworthiness was associated with attitude shifts on immigration. Specifically, when individuals watched a video emphasizing that climate change is real and promoting positive attitudes towards immigrants and perceived it as highly trustworthy in terms of the content, this perception had larger positive effects on attitudes (and vice versa for videos opposing climate change and communicating negative attitudes towards immigrants). Similarly, when individuals perceived the immigration video as highly

trustworthy in terms of the presentation, the videos emphasizing positive attitudes towards immigrants exhibited larger positive effects on attitudes (and vice versa for videos communicating negative attitudes towards immigrants). These findings indicate that subjective perceptions of trustworthiness, rather than objective video features or ideological congruence, are central to understanding how deepfakes shape public opinion. Interestingly, our results also suggest that the perceived trustworthiness of a video's content exerts a more consistent and stronger effect than its presentation. Although visual and technical elements can enhance a video's sense of realism, it is the plausibility and coherence of the message that seem to play the more decisive role in shaping attitudes, at least in the political sphere, where audiences often possess prior knowledge about public figures; messages that align with these expectations may be perceived as more credible, even when their presentation is less polished.

6 Concluding Remarks

In conclusion, the evidence reviewed in this chapter paints a comprehensive picture of why people believe deepfakes and how such beliefs can shape attitudes and behavioural intentions. We began by highlighting that, despite public confidence in detection abilities, people are generally poor at distinguishing deepfakes from authentic videos, often performing only slightly above chance.

We then introduced the concept of perceived trustworthiness as a way to move beyond binary detection measures and capture the perceptual factors that drive belief in deepfakes. Our work distinguishes between the trustworthiness of a video's content (i.e., how plausible and credible the message appears) and its presentation (i.e., how authentic the visual, auditory, and behavioural cues seem). This distinction reveals that, due to their multimodal nature, judgments of deepfake videos go far beyond evaluations related to the factual accuracy of the content. While both dimensions matter, trustworthiness of content emerges as more strongly linked to individual differences such as political orientation, trust in media, and cognitive reflection, and more predictive of attitudinal outcomes, perhaps because audiences are not (yet) adept at scrutinizing subtle visual or behavioural inconsistencies.

We further examined how individual characteristics spanning demographic, motivational, and cognitive factors interact with perceptual processes to shape susceptibility. Factors such as age, social media use, media literacy, "bullshit

receptivity", and deepfake-specific knowledge influence whether viewers are more or less likely to accept deepfakes as genuine. Importantly, these variables are often more strongly associated with content-related trustworthiness than presentation-related trustworthiness.

Finally, we showed that perceptions of trustworthiness do not remain at the level of passive judgments; they can translate into measurable attitude change and behavioural intentions such as sharing content on social media. In our studies, the perceived trustworthiness of content consistently predicted shifts in views on polarized issues like climate change and immigration, regardless of objective video quality or political alignment. This highlights the broader persuasive potential of deepfakes; even imperfect manipulations can influence public opinion when their message resonates.

Taken together, these findings demonstrate that it is not the objective properties of a video, but the perceived credibility of its message and presentation, that drive its psychological impact. If deepfakes are a technological challenge, belief in deepfakes is a psychological one. Protecting the public will therefore require both technological detection tools and psychological interventions that address the perceptual, cognitive, and motivational factors underlying belief. In an era where seeing is no longer believing, this dual approach is essential for preserving informed decision-making, public trust, and democratic stability.

Building on this, the construct of perceived trustworthiness, along with the developed questionnaire, which represent the chapter's most significant contributions, may also guide policy and platform responses, explored in more detail in Chapter 8. Because the PDTQ quantifies how believable a deepfake appears to ordinary viewers, it can be used as an input for automated moderation pipelines or risk assessment systems, for example, by assigning each video a "harm score". Content that scores high on trustworthiness but is identified as synthetic could be prioritized for rapid review or removal, while lower-scoring deepfakes might be flagged for further verification without immediate action. Similarly, PDTQ items may be used to develop specific interventions prior to exposure and deliberation prompts at the point of exposure, helping users critically evaluate manipulative content before it shapes their beliefs or behaviour. In this way, psychological insights into why people believe deepfakes can directly inform scalable, evidence-based, and,

perhaps most importantly, citizen-empowering policy responses, bridging the gap between individual-level perception and systemic prevention strategies.

End notes

All authors helped conceptualize the chapter and actively contributed to psychological studies carried out within the SOLARIS project, which are presented in the chapter. Nejc Plohl prepared the original draft, while Urška Smrke, Letizia Aquilino, and Izidor Mlakar contributed to reviewing and editing the chapter. Izidor Mlakar led this part of the project and supervised the writing process.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I., & Fishbein, M. (1972). Attitudes and normative beliefs as factors influencing behavioral intentions. *Journal of Personality and Social Psychology, 21*(1), 1–9. https://doi.org/10.1037/h0031930

Chen, S., Xiao, L., & Kumar, A. (2023). Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior, 141*, 107643. https://doi.org/10.1016/j.chb.2022.107643

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports, 16*, 100538. https://doi.org/10.1016/j.chbr.2024.100538

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports, 13*(1), 13429. https://doi.org/10.1038/s41598-023-39944-3

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods, 29*(3), 494–518. https://doi.org/10.1037/met0000540

Hameleers, M., van der Meer, T. G., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society, 8*(3), 20563051221116346. https://doi.org/10.1177/20563051221116346

Hameleers, M., van der Meer, T. G., & Dobber, T. (2023). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication, 39*(1), 56–70. https://doi.org/10.1177/0267323123118470

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior, 152*, 108096. https://doi.org/10.1016/j.chb.2023.108096

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

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

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), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lee, J., & Shin, S. Y. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology, 25*(4), 531–546. https://doi.org/10.1080/15213269.2021.2007489

O"Brien, T. C., Palmer, R., & Albarracin, D. (2021). Misplaced trust: When trust in science fosters belief in pseudoscience and the benefits of critical evaluation. *Journal of Experimental Social Psychology, 96*, 104184. https://doi.org/10.1016/j.jesp.2021.104184

Panno, A., Pellegrini, V., De Cristofaro, V., & Donati, M. A. (2023). A measure of positive and negative perception of migration: Development and psychometric properties of the Positive and Negative Perception of Immigrants Scale (PANPIS). *Analyses of Social Issues and Public Policy, 23*(1), 73–105. https://doi.org/10.1111/asap.12338

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Plohl, N., Mlakar, I., Aquilino, L., Bisconti, P., & Smrke, U. (2024). Development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ) in three languages. *International Journal of Human–Computer Interaction, 41*(11), 6786–6803. https://doi.org/10.1080/10447318.2024.2384821

Plohl, N., Mlakar, I., Aquilino, L., Brienza, M., Bisconti, P., & Smrke, U. (2025a). The moderating role of perceived trustworthiness in explaining the attitudinal effects of political deepfakes. *SSRN Preprint*.https://doi.org/10.2139/ssrn.5351533

Plohl, N., Mlakar, I., Aquilino, L., Brienza, M., Bisconti, P., & Smrke, U. (2025b). How deepfake quality, media literacy, and personal attitudes shape detection, liking, and social media sharing of political deepfakes. *PsyArXiv Preprint*.https://doi.org/10.31234/osf.io/knvby

Plohl, N., Mlakar, I., & Kecelj, Ž. (2025c). *Investigating the effects of an educational infographic and cognitive load on deepfake detection and metacognitive judgments.* Manuscript in preparation.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science, 7*(10), 201199. https://doi.org/10.1098/rsos.201199

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior, 149*, 107917. https://doi.org/10.1016/j.chb.2023.107917

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies, 2025*(1), Article 1833228. https://doi.org/10.1155/hbe2/1833228

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445699

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2021). How are deepfake videos detected? An initial user study. In *Proceedings of the 23rd HCI International Conference, HCII 2021, Part 1* (pp. 631–636). Springer. https://doi.org/10.1007/978-3-030-78635-9_80

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine, 28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Whitmarsh, L. (2011). Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global Environmental Change, 21*(2), 690–700. https://doi.org/10.1016/j.gloenvcha.2011.01.016