*Memory, Mind & Media* (2022), **1**, e15, 1–13
doi:10.1017/mem.2022.8

MMM
Memory, Mind & Media

RESEARCH ARTICLE

# Identifying and minimising the impact of fake visual media: Current and future directions

Sophie J. Nightingale[1] and Kimberley A. Wade[2]

[1]Department of Psychology, Lancaster University, Lancaster, UK and [2]Department of Psychology, University of Warwick, Coventry, UK
Corresponding author: Sophie J. Nightingale, email: s.nightingale1@lancaster.ac.uk

(Received 2 September 2022; accepted 8 September 2022)

## Abstract

Over the past two decades, society has seen incredible advances in digital technology, resulting in the wide availability of cheap and easy-to-use software for creating highly sophisticated fake visual content. This democratisation of creating such content, paired with the ease of sharing it via social media, means that ill-intended fake images and videos pose a significant threat to society. To minimise this threat, it is necessary to be able to distinguish between real and fake content; to date, however, human perceptual research indicates that people have an extremely limited ability to do so. Generally, computational techniques fair better in these tasks, yet remain imperfect. What's more, this challenge is best considered as an arms race – as scientists improve detection techniques, fraudsters find novel ways to deceive. We believe that it is crucial to continue to raise awareness of the visual forgeries afforded by new technology and to examine both human and computational ability to sort the real from the fake. In this article, we outline three considerations for how society deals with future technological developments that aim to help secure the benefits of that technology while minimising its possible threats. We hope these considerations will encourage interdisciplinary discussion and collaboration that ultimately goes some way to limit the proliferation of harmful content and help to restore trust online.

**Keywords:** Image manipulation; Deep fake; Visual processing; Human perception; Digital image forensics; Misinformation; Social media; Ethics

Andrew Walz, according to his Twitter account, was a congressional candidate running for office in Rhode Island in the 2020 US Presidential election. At a glance, there was nothing unusual about Walz' account – it included details of why he was running, his political ambitions, his profile picture, and Twitter's coveted blue checkmark indicating that Walz had been verified as a genuine candidate. Walz, however, was not a real person; he was a fictional character created by a 17-year-old high school student who was suffering from boredom during the school holidays (O'Sullivan 2020). The student created the profile and inserted a photo of a middle-aged white man from the website thispersondoesnotexist.com – as its name suggests, this site generates faces of people who do not exist. Fortunately, the Andrew Walz incident did not lead to any harm; nonetheless, it raises concerns about the risks associated with the remarkable technological advances

that allow almost anyone to create visually compelling fake content. In line with the focus of *Memory, Mind, and Media*, here we discuss how technology has paved the way for the development of fake visual media, and we highlight directions for future, interdisciplinary, research on detecting fake visual media and for minimising its harmful effects on human memory and cognition.

## Technology has blurred the line between real and fake imagery

The premise of manipulating visual media is not new. Until relatively recently, though, the ability to do so remained in the hands of those with considerable digital editing expertise. Rapid and substantial technological advances mean that today almost anyone can manipulate digital images. The most recent technological leap involves the use of artificial intelligence, specifically generative adversarial networks (GANs) to synthesise images, like the one used by the prankster student to create Andrew Walz. The accessibility of such technologies – which allow users to manipulate and synthesise images both easily and cheaply – brings opportunities for education, art, and science, but also potential to use fabricated visual imagery for nefarious purposes (see Silbey and Hartzog 2019 for an insightful discussion on neutralising the destructive force of deep fakes).

The successful use of fake visual imagery for harm, of course, somewhat depends on whether people can sort the real from the fake. New studies have begun to answer this question, examining the extent to which observers can detect various types of digital manipulations (e.g., Farid and Bravo 2010; Robertson et al 2017; Shen et al 2019). In our own labs, we have explored people's ability to sort authentic from manipulated images of real-world scenes (Nightingale et al 2017, 2019, 2022). People's ability to determine if an image is real or fake depends on the type of manipulation applied to the image (e.g., airbrushing, shadow alterations, object insertion), but in general people perform only slightly above chance on this task (Nightingale et al 2019). Somewhat counterintuitively, in our research we have found no strong evidence that individual factors, such as having an interest in photography or experience of editing images, are associated with increased accuracy in distinguishing real from manipulated photos. In one study, we tested 15,873 individuals aged 15–75, and found that older adults categorised images less accurately than their middle-aged and younger counterparts, which is not overly surprising given that visual processing declines with age (Nightingale et al 2022). Perhaps more importantly, we found that despite their mediocre ability to sort the genuine photos from the fakes, participants of all ages tended to express confidence in their judgements (Nightingale et al 2022). These results suggest that people are not only poor at detecting manipulated photos, but also often unaware of just how poor they are.

A relatively new type of image manipulation – face morphing – has raised significant concerns about the opportunity afforded by advanced digital editing software for committing identity theft. Face morphs are created by digitally combining images of two or more individuals to create a new image that resembles each of the original identities. The wide availability of face morphing software means that this type of fake image now poses a serious threat to border controls and other face recognition systems used in security settings (see Pikoulis et al 2021), and worryingly, several studies show that people cannot reliably detect face morphs (Kramer et al 2019; Nightingale et al 2021; Robertson et al 2017, 2018). In one recent study, participants attempted to determine whether two face images depicted the same person or not (Robertson et al 2017). Over 49 trials, each participant viewed three types of face pairs: (1) two facial images of the same individual, (2) faces belonging to two different individuals, and (3) one individual's face, alongside a 50/50 morph combining that same individual with another person's face. Participants incorrectly endorsed the 50/50 morphs as being a 'match', on average, 68 per cent of the

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

time with error rates under 10 per cent for the same individual and different individual face pairs.

In the previous few years, researchers have turned their attention to an even more sophisticated technology for producing manipulated media: deep fakes (Agarwal et al 2019). The advent of AI-synthesised content has signalled a giant leap forward in the creation of photorealistic fake visual media. Deep fake images are typically synthesised using GANs, a type of machine learning that works by pitting two neural networks (a generator and a discriminator) against one another in an iterative back-and-forth process. This GAN structure can be used to synthesise any type of fake image, as well as video and audio. Highlighting the photo-realism of the outputs of synthesis machines, studies have begun to reveal that people are frequently fooled by deep fakes and cannot reliably distinguish between real and synthetic faces (Hulzebosch et al 2020; Lago et al 2022; Nightingale and Farid 2022) or videos (Groh et al 2021; Hughes et al 2021), yet show overconfidence in their ability to recognise deep fake content (Sütterlin et al 2022).

## Fake imagery influences thought, intent, and behaviour

The research evidence is clear: Distinguishing between doctored and authentic visual media is difficult, and research dating back to the early 2000s illustrates why we should be concerned. Fake media can have detrimental effects on human memory, cognition, and behaviour. Studies have shown that doctored photos of prominent public events can change not only what people recollect about those events but also people's attitudes and behavioural intentions (Frenda et al 2013; Nash 2018). Sacchi et al (2007), for example, created doctored photos of the 1989 Tiananmen Square protest in Beijing and the 2003 Iraq war protest in Rome. For the Rome event, aggressive-looking demonstrators and police officers wearing riot gear were inserted into an original image of the peaceful demonstration. During a memory test, people who viewed the doctored Rome photo were more likely to state that the protest involved physical confrontation, significant injuries, and damage to property, and indicated they would be less inclined to participate in future protests, than people who viewed the original photo. Doctored images can also distort how people recall self-involving, significant, childhood experiences (e.g., taking a hot air balloon ride, Hessen-Kayfitz and Scoboria 2012; Wade et al 2002) or recent everyday actions (e.g., shuffling cards or counting to 20, Nash et al 2009; Nash and Wade 2009; Wright et al 2013). Doctored videos have been used to induce people to falsely confess to committing an undesirable act (e.g., cheating in a gambling task) or to provide erroneous testimony about another person's actions (Nash and Wade 2009, Wade et al 2010 Wright et al 2013).

More recent research has demonstrated the potential for doctored imagery to change consumers' memories of what they have purchased and the brands they prefer. Under the guise of conducting consumer research, Hellenthal et al (2016) instructed participants to compile a basket of 12 food items made by their preferred brands. The researchers then took a photo of the participant with their 'personal brand lifestyle basket'. Approximately a week later, participants viewed a doctored version of the photo in which 4 of their 12 selected items were replaced with similar items made by different brands. Participants were asked whether they were comfortable with the photo being included in their brand profile. The next day, participants completed a surprise memory test in which they were asked to indicate which items they included in their original basket. Participants displayed the classic 'misinformation effect' – frequently remembering items that were merely suggested to them in the fake photo. Their brand preference ratings also changed after viewing the doctored photo, with many participants indicating a positive shift in attitude and behaviour towards brands that were suggested to them.

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

It is clear, from over 20 years of empirical research, that doctored images can have powerful effects on cognition and behaviour. Interestingly, psychologists conducting doctored photo research in the early 2000s were of the belief that 'most of us will never be confronted with images of ourselves doing things we have never done, or in places we have never been' (Strange et al 2005, 240). Today we are not so sure, given the numerous technologies that can capture, archive, and visualise images of our personal experiences, and given the widespread availability of digital editing software (not to mention the emerging synthesis technologies, but see Murphy and Flynn 2021).

People's limited ability to detect fake visual media, paired with its powerful effects, makes it highly effective for nefarious purposes. Furthermore, the prevalence of social media platforms allows content – both real and fake – to be rapidly disseminated across the world. We have already witnessed the harm that synthesised and manipulated images can have, for example, in creating non-consensual sexual imagery, committing financial fraud and identity theft, and fuelling misinformation campaigns (e.g., Kalpokas & Kalpokiene 2022; Sleigh 2021; Wakefield 2022; Westerlund 2019). We now turn to an important question that has received relatively little attention from researchers, particularly social scientists: What can be done to improve the detection of fake visual content?

## Improving detection of fake imagery

### *Computational detection*

There is a substantial literature on digital image forensics; however, here we will mention a selection of the recent techniques proposed for detecting deep fakes (for an overview of digital forensic techniques, see Farid 2016). The forensic techniques tend to be categorised as low- or high-level. Low-level techniques detect pixel-level artefacts that are not visually perceivable by human observers, for example evidence of warping or blending (Li and Lyu 2018; Li et al 2020). These low-level approaches can often achieve high accuracy in detecting fakes; however, a limitation is that they are sensitive to counter attacks (e.g., the addition of noise or image resizing can destroy the artefact used for detection; Carlini and Wagner 2017). High-level techniques focus on artefacts within semantically meaningful information, such as eye blinks, head pose, and mannerisms. One interesting new finding is that deep fake videos can be reliably detected by identifying an inconsistency between a synthesised person's mouth shape and a spoken phoneme (Agarwal et al 2020). The authors point out that for words containing the letters 'B', 'M', or 'P', it is near impossible to make those sounds without closing your lips – for example, if you try to say 'mother', it likely is impossible to enunciate clearly without your lips touching. This realisation has led to the development of a forensic technique using phoneme-viseme mismatches to detect state-of-the-art deep fake videos (Agarwal et al 2020).

Another slightly different approach proposed by forensic image experts is to introduce watermarks into any new technologies for synthesising media content. The development of synthesis technology involves using machine learning to train a model that generates photorealistic content without any human input – a generative model. Drawing on the already widespread use of watermarking for copyrighting digital property, it is possible to take a similar approach with synthesis technology. In this instance, a watermark is inserted within the set of real images used for training the model, such that all of the training images will also contain the unique identification information. Crucially, research has shown that when training such a model using a watermarked dataset of images, the model takes on this identification information during its learning; the information then becomes embedded within any subsequently synthesised content (Yu et al 2021). As such, the watermarking approach allows fake images to be reliably detected downstream.

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

With the rapid development of AI-synthesised content, this watermarking approach offers a proactive solution to the threat of deep fakes.

### *Human detection*

Scientists have tested certain interventions for improving people's ability to detect fake imagery. One line of research has, to a large extent, been driven by the disturbing finding that unrealistic and unattainable beauty 'ideals' depicted in (manipulated) images of fashion models can cause psychological harm to observers (e.g., Grabe et al 2008). In an attempt to mitigate the damaging effects of viewing these impossible beauty standards, researchers have explored the possible benefits of adding disclaimer labels to images to indicate, for example, when the body of a model has been digitally manipulated. Several studies have found that the addition of such labels does not necessarily help people to discount these images – in fact, some results suggest that this approach might even prove counterproductive by encouraging viewers to direct more, rather than less, attention to the model's body (Selimbegović and Chatard 2015; Slater et al 2012; Tiggeman et al 2013). Related studies have revealed the limitations of warning labels by showing that they are only partially effective in reducing people's belief in fake news headlines (Ecker et al 2010; Pennycook et al 2020) and in fake photos of public events (Nash 2018).

Another area of research that aims to help people identify fake imagery has involved encouraging people to look for the visual artefacts left behind by any editing or synthesis process. When morphing faces, for example, the morph might contain tell-tale signs of digital editing such as a ghost-like outline of another person's face or hair over the forehead (Robertson et al 2017). Initially, this simple training approach seemed promising, as research showed that participants who were trained to detect morphing artefacts mistakenly accepted 50/50 morphs as being a 'match' 21 per cent of the time on average, compared to 68 per cent for untrained participants. Yet in two subsequent studies that used higher-quality face morphs, training led to no reliable improvement in performance (Kramer et al 2019; Nightingale et al 2021). With GANs now being used to streamline the generation of high-quality face morphs, it seems likely that training people to spot artefacts in images is, or soon will be, redundant (Venkatesh et al 2020). Similarly, our own research has shown that informing participants about the common artefacts associated with synthesising content only enhanced their accuracy slightly, compared to participants who were given no information (Nightingale and Farid 2022). As technology improves, any artefacts produced in the manipulation process will likely disappear again limiting the usefulness of any detection technique that rests on alerting people to such artefacts.

## Why are people limited in their ability to detect fake images?

An important question that remains largely unanswered is *why* people have such limited ability to determine when an image is real and when it is fake. One obvious, potential explanation is that fake images are now so sophisticated and realistic that there are no perceptible clues to alert people that the image has been manipulated or synthesised. Yet, in studies involving images that do contain detectable signs of manipulation, including changes that are physically impossible (such as a scene containing cast shadows that are inconsistent with the lighting source), people still frequently fail to notice that those images are manipulated (e.g., Nightingale et al 2019). Based on these findings, it is reasonable to think there may be value in training people to detect signs of manipulation. However, given the evidence outlined above, which suggests limited benefits of training interventions, more empirical work is needed to advance our understanding of exactly when and why people fail to successfully use such signs.

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

When thinking about why people fail to notice signs of digital manipulation, a good starting point is to consider the limits of human perception. Decades of cognitive science has shown that people's capacity to perceive the visual world is finite, with seminal studies demonstrating that people can fail to notice even highly conspicuous events unfolding right in front of them (e.g., Neisser 1979; Neisser and Becklen 1975). Perceptual failures have been shown in change blindness and inattentional blindness studies, where people are surprisingly unaware of significant changes to, or the presentation of, stimuli outside of their focus of attention (e.g., Rensink et al 1997; Simons and Chabris 1999). One of the most famous examples of inattentional blindness is the 'invisible gorilla' study (Simons and Chabris 1999) in which participants observed a video of a ball game while counting the number of balls passes made between the players in the game. When engaged in this task, approximately half of participants failed to see a person dressed as a gorilla walk through the middle of the ball game. Furthermore, these perceptual failures are affected by the observer's perceptual load; when people are tasked with processing a lot of information, they are less likely to detect changes in scenes (e.g., Carmel et al 2011). Thus, it remains possible that attention is another crucial factor that impacts whether or not people notice when an image has been manipulated.

It is also important to think about the challenge of distinguishing between authentic and manipulated media in a digital world where the internet, and particularly social media, offers endless content (more than 3.2 billion images are shared online each day; Thomson et al 2020). Research drawing on cognitive and evolutionary theory, along with behavioural economics, shows us that when people have access to vast amounts of information, the way they search that information shapes the decisions they make (Hills and Hertwig 2010). Technological developments afford an ever-increasing ability to store and share information, yet the psychological limits on people's capacity to process information remain unchanged, resulting in a state of information overload (Henkel et al 2021; Hilbert and López 2011; Hills 2019; van den Bosch et al 2016). With such overload, people must select what to attend to, what to believe, and what to share. However, not all information is equal: through evolution, humans have developed cognitive heuristics that make certain types of information more attention-worthy, such as negative information and information that is consistent with existing beliefs (Hills 2019). As such, it might be that some manipulations are detectable in principle, yet in a world overloaded with information, human perceptual limits lead people to overlook types of evidence that would indicate foul play.

]Alternatively, it might be that people simply do not know what to look for, and rely on unhelpful strategies when trying to verify the authenticity of an image. In a recent study, participants were asked to distinguish between manipulated and genuine photos of real-world events, and to report the strategies they used to determine whether an image had been manipulated or not (Nightingale et al 2022). Overall, people's success was similar regardless of whether or not they reported using a specific strategy, yet there were some interesting differences when looking at the specific types of strategy used. For example, those who reported paying careful attention and systematically 'zooming in' to look at different parts of the image were more accurate than those who did not report using this strategy. Although this notion of paying attention might seem obvious, only 2 per cent of participants (263/15,873) mentioned it. This finding suggests that people might be able to improve their detection of manipulated images simply by changing the way they approach the task, echoing Hills and Hertwig's (2010) finding that search strategy can play a crucial role in decision accuracy.

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

## The need for an interdisciplinary theoretical framework

An important next step in improving visual media authentication is to develop a theoretical framework for understanding how various factors – including individual, cognitive, environmental, and cultural – influence people's ability to detect manipulated images. As mentioned above, a small but rapidly growing body of empirical research spanning multiple disciplines speaks to this issue; much of this work could inform theory development.

Within cognitive psychology – our own discipline – one framework in particular could guide theoretical thinking: the source monitoring framework (*SMF*; Johnson et al 1993). Briefly, the SMF aims to explain how people distinguish between mental experiences that result from perception (i.e., memories of real events) versus mental experiences that result from internal processes (i.e., memories of dreams or thoughts). The SMF posits that people can determine the source of their mental experiences by evaluating the characteristics of those experiences. For example, when a memory or image comes to mind, one might consider how familiar, detailed, or coherent it is. If the mental experience has the characteristics typically associated with a memory of genuine experiences (i.e., it is sufficiently familiar, detailed, coherent), then the individual is likely to conclude that it is indeed a memory of something that really happened, rather than something that was merely imagined or thought about. Moreover, according to the SMF, people typically rely on two types of judgement processes to evaluate and classify their mental experiences – a slow systematic reflection and reasoning process, or a rapid, automatic heuristic process (Hasher and Zacks 1979; Johnson et al 1993). As you might expect, source misattributions (i.e., mistaking an imagined or internally generated event for a genuine memory) are more likely to occur when people rely on a rapid, heuristic, decision process.

We can apply the SMF judgement process to the task of distinguishing between genuine and fake visual imagery: If real and fake images differ in systematic and detectable ways, then people may engage in either a careful, systematic search of an image to detect clues that are indicative of a fake image, or they might rely on a more rapid and automatic judgement process to determine the image's authenticity. From a SMF perspective, we might predict that various extraneous factors could influence a person's ability to accurately evaluate an image and to determine whether it has been manipulated or not. One such factor is a person's political perspective, yet the evidence is mixed. Some research shows that people are more likely to buy into fake news, and mistake fictitious for genuine stories, if the false information aligns with their political beliefs or worldview (Frenda et al 2013; Greene et al 2021; Walter and Tukachinsky 2020; Zhou and Shen 2022). Other research suggests that susceptibility to fake news is less about how closely information aligns with an individual's political ideology and more about the extent to which an individual engages in analytical thinking (Pennycook and Rand 2019). Adding further complexity, in another study, partisan-motivated reasoning affected participants' susceptibility to believing political-based misinformation, however, more so for authentic video content than deep fake video content (Barari et al 2021). According to the SMF, when false information aligns with an individual's own views, beliefs, and stereotypes, it is likely that they will either automatically feel that information to be true, or through motivated reasoning conclude that the information is likely to be true (Mazzoni and Kirsch 2002). In a similar way, it seems reasonable to expect that a person's personal views, beliefs, and stereotypes might affect their ability (or effort) to detect image manipulations. Indeed, research has already shown that people's expectations and preferences can influence how they perceive visual information (Balcetis and Dunning 2006; Bruner and Potter 1964). To date, few studies have explored what makes people better or worse at detecting manipulated images, and the majority so far have involved images that depict unfamiliar people partaking in fairly mundane events. The images are not manipulated to serve a

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

particular political goal, or to comment on culture or society, or to evoke an emotional reaction in the observer. Therefore, it remains possible that in real-world scenarios, where visual media are often manipulated to serve a specific goal, observers' own goals might decide whether they perform better, or worse, when distinguishing authentic from manipulated images.

Another important factor that warrants greater attention from researchers is the context in which the image is viewed, and its apparent source. Research has shown that media platforms vary in terms of their perceived credibility, and the extent to which people trust any particular source might influence their credulity toward images appearing on that platform (Metzger et al 2010). Computer science and communications experts have started to address this question, and the data from one study suggest that the reported source of an image, and other contextual factors such as how many 'likes' it has received, in fact does not significantly affect observers' perceptions of image credibility (Shen et al 2019). The data did, however, reveal that observers' attitudes and individual factors, such as their photo-editing experience, affected their perceptions of image credibility.[1]

## Ethical challenges when seeing is not believing

Finally, the issue of sophisticated fake visual media raises a number of ethical challenges. Consider the so-called liar's dividend: perhaps one of the most concerning consequences of how easily people can manipulate and synthesise visual digital content. In a world where practically any image, video, or audio can be manipulated, it is easy to dismiss *anything* as fake. Soldiers pictured committing human rights violations, a CEO captured in an embarrassing photo, or a politician at a party they had claimed they did not attend: All of these people could, with enough plausibility to satisfy at least their most willing audiences, argue that those images are fraudulent. We have seen this strategy used in recent years, with former US President Donald Trump denying the authenticity of the 2005 recording of him bragging about sexually assaulting women (Fahrenthold 2016). Below we highlight the need to consider how society deals with future technological developments, to help us to secure the benefits of that technology while minimising its possible threats.

One consideration is how to balance the practice of open code and software distribution with the ethical sharing of image manipulation and synthesising technology. Open science initiatives encourage scientists to make their methods, data, and analytical and computer code openly available, which serves to enhance scientific rigour and researcher integrity as well as encourage the collaborative development of technology. The scientific community should, however, more carefully consider when this sharing is ethical and when there might be good reason to keep certain resources out of the public domain. New technologies, including GANs, quickly become widely and freely available on sites like GitHub, often with walkthroughs for implementation. On the one hand, and for the most part, access to such technology is non-problematic and allows for further advances to be made and the potential to develop use for good. For example, through the use of deep fakes in the documentary 'Welcome to Chechnya', LGBT individuals were able to testify anonymously about their suffering and persecution in Russia (RD 2020). On the other hand, the open access also extends to malicious actors who wish to deploy the technology

[1] These findings seem difficult to reconcile with those of Nightingale et al (2017) who found that people's photo-editing experience did not predict their ability to determine whether a photo was genuine or not. It is possible that specific methodological differences across the studies, such as the phrasing of the image authenticity question, could play a role.

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

for harm – for example, to generate images that can be used to scam a victim or to create videos to support false claims posted on social media. The balance between open and ethical sharing is a complex issue and one that requires interdisciplinary discussion to ensure the development of sensible and useful guidelines.

Another, much broader, consideration is how the research community might develop appropriate guidelines for the ethical development and use of new technologies. The market has exploded with new applications using GANs to create deep fakes – either for free or at a relatively low cost (Cole 2018). One application – FakeApp – introduced in 2018, which allows users to create deep fake videos at the press of a button, gathered great interest with hundreds of thousands of downloads in the first month of its release (Marino 2018). Although the complexity of training a GAN still prevents many from creating their own models, the development of applications like FakeApp opens up the market to everyone. As such, the potential for misuse is wide; one of the most common abuses so far being the creation of non-consensual sexual imagery. In 2019, research conducted by a cybersecurity company, Deeptrace, revealed that 96 per cent of the deep fake videos online at that time were of a pornographic nature, and the victims overwhelmingly women (Ajder et al 2019; Wang 2019). The ethical and moral concerns surrounding these new technologies are highlighted in the steadily growing number of publications on this topic from fields such as law, information technology, and political science (e.g., de Ruiter 2021). We believe that there is much that researchers from a range of disciplines can contribute to this discussion.

A final consideration is for the giants of the technology sector to understand how their platforms are used for sharing and weaponising content, and to put substantial effort into preventing such misuses. Business media experts have posited that social media companies are doing a substandard job of keeping harmful content, such as COVID-19 vaccine misinformation, off their platforms (O'Sullivan et al 2021). In a recent study examining 30 anti-vaccine Facebook groups, researchers discovered that just 12 individuals accounted for sharing 70 per cent of anti-vax disinformation within these groups (Center for Countering Digital Hate 2021). Of course, this study considered only a subset of Facebook groups, but it does pose an interesting question: if researchers can find those responsible for posting this vaccine disinformation, why can't Facebook? The better question is perhaps why *won't* they, as opposed to why *can't* they. Meta (previously Facebook) reported that 97 per cent of its total revenue from October to December 2021 came from advertising (Johnston and Cheng 2022). The business model underpinning such success involves gleaning as much data as possible from site users, to build detailed profiles ripe for ad targeting. One way to keep users returning to social media sites is to show controversial and evocative content that captivates interest (Kim 2015) – fake content can achieve this goal extremely effectively, given that it is free from factual constraints (Lewandowsky and Pomerantsev 2022). Deep fakes might be particularly powerful when it comes to captivating users' attention, especially given humans' ability to quickly recognise and understand visual content (e.g., Greene and Oliva 2009; Isola et al 2013). Therefore, legislators should consider reasonable policy and regulation for ensuring that social media companies are accountable for real-world harms that might result from their services. Modest regulatory changes should incentivise companies to introduce safeguards, and as a result, help toward restoring trust in our digital world.

Ultimately, the potential consequences of fake imagery mean that it is worthwhile examining new ways of improving people's ability to sort the fake from the genuine. These attempts stand to be useful, even if they were only to equip people to weed out the poorer attempts at manipulation. With the pace at which technology is improving, it is perhaps overly optimistic to think that people could learn to reliably detect the most sophisticated fakes that are now readily disseminated across the internet. Instead,

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

within the research community, we should continue to raise awareness of the current and emerging threats, with the aim to encourage more research in this area, including the development of improved computational methods of detection – or face the possibility that people will be fooled by scams far worse than that of the made-up congressional candidate, Andrew Walz.

**Acknowledgement.** Thanks to Rob Nash for his insightful comments on a draft of the manuscript.

**Funding statement.** This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Conflict of interest.** The authors declare that they have no competing interests.

## References

**Agarwal S, Farid H, Gu Y, He M, Nagano K and Li H** (2019) Protecting world leaders against deep fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 38–45.

**Agarwal S, Farid H, Fried O and Agrawala M** (2020) Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 660–661.

**Ajder H, Patrini G, Cavalli F and Cullen L** (2019) The state of deepfakes: Landscape, threats, and impact. September 2019. Available at https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

**Balcetis E and Dunning D** (2006) See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology* **91**, 612–625. https://doi.org/10.1037/0022-3514.91.4.612

**Barari S, Lucas C and Munger K** (2021) Political deepfake videos misinform the public, but no more than other fake media. OSF Preprints.

**Bruner JS and Potter MC** (1964) Interference in visual recognition. *Science* **144**, 424–425. https://doi.org/10.1126/science.144.3617.424

**Carlini N and Wagner D** (2017) Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, 39–57. https://doi.org/10.1109/SP.2017.49.

**Carmel D, Thorne JD, Rees G and Lavie N** (2011) Perceptual load alters visual excitability. *Journal of Experimental Psychology: Human Perception and Performance* **37**, 1350–1360. https://doi.org/10.1037/a0024320

**Center for Countering Digital Hate LTD** (2021) *Pandemic profiteers: The business of anti-vax*. Available at https://counterhate.com/wp-content/uploads/2022/05/210601-Pandemic-Profiteers-Report.pdf (accessed 1 May 2022).

**Cole S** (2018) *We are truly fucked: Everyone is making AI-generated fake porn now*. Available at: https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley (accessed 1 May 2022).

**de Ruiter A** (2021) The distinct wrong of deepfakes. *Philosophy and Technology* **34**, 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

**Ecker UKH, Lewandowsky S and Tang DTW** (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* **38**, 1087–1100. https://doi.org/10.3758/MC.38.8.1087

**Fahrenthold D** (2016) Trump recorded having extremely lewd conversation about women in 2005. *The Washington Post*, 8 October. Available at https://www.washingtonpost.com/politics/trump-recorded-having-extremely-lewd-conversation-about-women-in-2005/2016/10/07/3b9ce776-8cb4-11e6-bf8a-3d26847eeed4_story.html (accessed 1 May 2022).

**Farid H** (2016) *Photo Forensics*. Cambridge, MA: The MIT Press.

**Farid H and Bravo MJ** (2010) Image forensic analyses that elude the human visual system. In *Proceedings of SPIE*, vol. **7541**, 1–10. https://doi.org/10.1117/12.837788

**Frenda SJ, Knowles ED, Saletan W and Loftus EF** (2013) False memories of fabricated political events. *Journal of Experimental Social Psychology* **49**, 280–286. https://doi.org/10.1016/j.jesp.2012.10.013

**Grabe S, Ward LM and Hyde JS** (2008) The role of the media in body image concerns among women: A meta-analysis of experimental and correlational studies. *Psychological Bulletin* **134**, 460–476. https://doi.org/10.1037/0033-2909.134.3.460

**Greene MR and Oliva A** (2009) The briefest of glances: The time course of natural scene understanding. *Psychological Science* **20**, 464–472. https://doi.org/10.1111/j.1467-9280.2009.02316.x

**Greene CM, Nash RA and Murphy G** (2021) Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory* **29**, 587–604. https://doi.org/10.1080/09658211.2021.1923754

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

**Groh M, Epstein Z, Firestone C and Picard R** (2021) Deepfake detection by human crowds, machines, and machine-informed crowds. *PNAS* **119**, e2110013119. https://doi.org/10.1073/pnas.2110013119

**Hasher L and Zacks RT** (1979) Automatic and effortful processes in memory. *Journal of Experimental Psychology: General* **108**, 356–388. https://doi.org/10.1037/0096-3445.108.3.356

**Hellenthal MV, Howe ML and Knott LM** (2016) It must be my favourite brand: Using retroactive brand replacements in doctored photographs to influence brand preferences. *Applied Cognitive Psychology* **30**, 863–870. https://doi.org/10.1002/acp.3271

**Henkel LA, Nash RA and Paton JA** (2021) 'Say Cheese!' How taking and viewing photos can shape memory and cognition. In Lane S and Atchley B (eds), *Human Capacity in the Attention Economy*. Washington, DC: American Psychological Association, 103–133.

**Hessen-Kayfitz JK and Scoboria A** (2012) False memory is in the details: Photographic details differentially predict memory formation. *Applied Cognitive Psychology* **26**, 333–341. https://doi.org/10.1002/acp.1839

**Hilbert M and López P** (2011) The world's technological capacity to store, communicate, and compute information. *Science* **332**, 60–65. https://doi.org/10.1126/science.1200970

**Hills TT** (2019) The dark side of information proliferation. *Perspectives on Psychological Science* **14**, 323–330. https://doi.org/10.1177/1745691618803647

**Hills TT and Hertwig R** (2010) Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science* **21**, 1787–1792. https://doi.org/10.1177/0956797610387443

**Hughes S, Ferguson M, Hughes C, Hughes R, Fried O, Yao X and Hussey I** (2021) Deepfaked online content is highly effective in manipulating people's attitudes and intentions. OSF Preprints. https://doi.org/10.31234/osf.io/4ms5a (accessed 16 August 2022).

**Hulzebosch N, Ibrahimi S and Worring M** (2020) Detecting CNN-generated facial images in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 642–643.

**Isola P, Xiao J, Parikh D, Torralba A and Oliva A** (2013) What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**, 1469–1482.

**Johnson MK, Hashtroudi S and Lindsay SD** (1993) Source monitoring. *Psychological Bulletin* **114**, 3–28. https://doi.org/10.1037/0033-2909.114.1.3

**Johnston M and Cheng M** (2022) How Facebook (Meta) makes money. *Investopedia*, 4 February. Available at https://www.investopedia.com/ask/answers/120114/how-does-facebook-fb-make-money.asp#:~:text=(FB)%2C%20the%20company%20that,communicate%20with%20family%20and%20friends (accessed 7 June 2022).

**Kalpokas I and Kalpokiene J** (2022) *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation*. Singapore: Springer Nature.

**Kim HS** (2015) Attracting views and going viral: How message features and news-sharing channels affect health news diffusion. *Journal of Communication* **65**, 512–534. https://doi.org/10.1111/jcom.12160

**Kramer RS, Mireku MO, Flack TR and Ritchie KL** (2019) Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications* **4**, 1–15. https://doi.org/10.1186/s41235-019-0181-4

**Lago F, Pasquini C, Böhme R, Dumont H, Goffaux V and Boato G** (2022) More real than real: A study on human visual perception of synthetic faces. *IEEE Signal Processing Magazine* **39**, 109–116. https://doi.org/10.1109/MSP.2021.3120982

**Lewandowsky S and Pomerantsev P** (2022) Technology and democracy: A paradox wrapped in a contradiction inside an irony. *Memory, Mind, & Media* **1**, e5. https://doi.org/10.1017/mem.2021.7

**Li Y and Lyu S** (2018) Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint* arXiv:1811.00656.

**Li L, Bao J, Zhang T, Yang H, Chen D, Wen F and Guo B** (2020) Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001–5010.

**Marino D** (2018) FakeApp: Groundbreaking or dangerous? Available at: https://www.artefactmagazine.com/2018/02/13/fakeapp-groundbreaking-or-dangerous/ (accessed 1 May 2022).

**Mazzoni G and Kirsch I** (2002) Autobiographical memories and beliefs: A preliminary metacognitive model. In Perfect TJ and Schwartz BL (eds), *Applied Metacognition*: Cambridge University Press, 121–145. https://doi.org/10.1017/CBO9780511489976.007

**Metzger MJ, Flanagin AJ and Medders RB** (2010) Social and heuristic approaches to credibility evaluation online. *Journal of Communication* **60**, 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

**Murphy G and Flynn E** (2021) Deepfake false memories. *Memory*, 1–13. https://doi.org/10.1080/09658211.2021.1919715

**Nash RA** (2018) Changing beliefs about past public events with believable and unbelievable doctored photographs. *Memory* **26**, 439–450. https://doi.org/10.1080/09658211.2017.1364393

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

**Nash RA and Wade KA** (2009) Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence. *Applied Cognitive Psychology* **23**, 624–637. https://doi.org/10.1002/acp.1500

**Nash RA, Wade KA and Lindsay DS** (2009) Digitally manipulating memory: Effects of doctored videos and imagination in distorting beliefs and memories. *Memory & Cognition* **37**, 414–424. https://doi.org/10.3758/MC.37.4.414

**Neisser U** (1979) The control of information pickup in selective looking. In Pick H (ed.), *Perception and Development: A Tribute to Eleanor Gibson*. New York: Halsted Press, pp. 201–219.

**Neisser U and Becklen R** (1975) Selective looking: Attending to visually specified events. *Cognitive Psychology* **7**, 480–494. https://doi.org/10.1016/0010-0285(75)90019-5

**Nightingale SJ, Agarwal S and Farid, H** (2021) Perceptual and computational detection of face morphing. *Journal of Vision* **21**, 1–18. https://doi.org/10.1167/jov.21.3.4

**Nightingale SJ and Farid H** (2022) AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America* **119**, e2120481119. https://doi.org/10.1073/pnas.2120481119

**Nightingale SJ, Wade KA and Watson DG** (2017) Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications* **2**, 30. https://doi.org/10.1186/s41235-017-0067-2

**Nightingale SJ, Wade KA, Farid H and Watson DG** (2019) Can people detect errors in shadows and reflections? *Attention, Perception, & Psychophysics* **81**, 2917–2943. https://doi.org/10.3758/s13414-019-01773-w

**Nightingale SJ, Wade KA and Watson DG** (2022) Investigating age-related differences in ability to distinguish between original and manipulated images. *Psychology and Aging* **37**, 326–337. https://doi.org/10.1037/pag0000682

**O'Sullivan D** (2020) A high school student created a fake 2020 candidate. Twitter verified it. *CNN Business*, 28 February. Available at https://edition.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html (accessed 1 May 2022).

**O'Sullivan D, Duffy C, Subramaniam T and Boxer S** (2021) Facebook is having a tougher time managing vaccine misinformation than it is letting on, leaks suggest. *CNN Business*, 27 October. Available at https://edition.cnn.com/2021/10/26/tech/facebook-covid-vaccine-misinformation/index.html (accessed 1 May 2022).

**Pennycook G and Rand DG** (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

**Pennycook G, Bear A, Collins ET and Rand GD** (2020) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* **66**, 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

**Pikoulis EV, Ioannou ZM, Paschou M and Sakkopoulos E** (2021) Face morphing, a modern threat to border security: Recent advances and open challenges. *Applied Sciences* **11**, 3207. https://doi.org/10.3390/app11073207

**RD** (2020) "Welcome to Chechnya" uses deepfake technology to protect its subjects. *The Economist*, 9 July. Available at https://www.economist.com/prospero/2020/07/09/welcome-to-chechnya-uses-deepfake-technology-to-protect-its-subjects (accessed 1 May 2022).

**Rensink RA, O'Regan JK and Clark JJ** (1997) To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* **8**, 368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x

**Robertson DJ, Kramer RS and Burton AM** (2017) Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PLoS One* **12**, e0173319. https://doi.org/10.1371/journal.pone.0173319

**Robertson DJ, Mungall A, Watson DG, Wade KA, Nightingale SJ and Butler S** (2018) Detecting morphed passport photos: A training and individual differences approach. *Cognitive Research: Principles and Implications* **3**, 1–11. https://doi.org/10.1186/s41235-018-0113-8

**Sacchi DLM, Agnoli F and Loftus EF** (2007) Changing history: Doctored photographs affect memory for past public events. *Applied Cognitive Psychology* **21**, 1005–1022. https://doi.org/10.1002/acp.1394

**Selimbegović L and Chatard A** (2015) Single exposure to disclaimers on airbrushed thin ideal images increases negative thought accessibility. *Body Image* **12**, 1–5. https://doi.org/10.1016/j.bodyim.2014.08.012

**Shen C, Kasra M, Pan W, Bassett GA, Malloch Y and O'Brien JF** (2019) Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society* **21**, 438–463. https://doi.org/10.1177/1461444818799526

**Silbey J and Hartzog W** (2019) The upside of deep fakes. *Maryland Law Review* **78**, 960–966.

**Simons DJ and Chabris CF** (1999) Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* **28**, 1059–2074. https://doi.org/10.1068/p281059

https://doi.org/10.1017/mem.2022.8 Published online by Cambridge University Press

**Slater A, Tiggemann M, Firth B and Hawkins K** (2012) Reality check: An experimental investigation of the addition of warning labels to fashion magazine images on women's mood and body dissatisfaction. *Journal of Social and Clinical Psychology* **31**, 105–122. https://doi.org/10.1521/jscp.2012.31.2.105

**Sleigh S** (2021) MP demands deepfake porn and 'nudifying' images are made sex crimes. *HuffPost*, 2 December. Available at https://www.huffingtonpost.co.uk/entry/ban-rape-deepfake-nudifying-tech_uk_61a79734e4b0f398af1aeeb1 (accessed 1 May 2022).

**Strange D, Gerrie MP and Garry M** (2005) A few seemingly harmless routes to a false memory. *Cognitive Processing* **6**, 237–242. https://doi.org/10.1007/s10339-005-0009-7

**Sütterlin S, Lugo RG, Ask TF, Veng K, Eck J, Fritschi J, Özmen T, Bärreiter B and Knox BJ** (2022) The role of IT background for metacognitive accuracy, confidence and overestimation of deep fake recognition skills. In Schmorrow DD and Fidopiastis CM (eds), *Augmented Cognition, Lecture Notes in Computer Science*, vol. **13310**. Springer, Cham. https://doi.org/10.1007/978-3-031-05457-0_9

**Thomson TJ, Angus D and Dootson P** (2020) 3.2 billion images and 720,000 hours of video are shared online daily. Can you sort real from fake? *The Conversation*, 3 November. Available at https://theconversation.com/3-2-billion-images-and-720-000-hours-of-video-are-shared-online-daily-can-you-sort-real-from-fake-148630

**Tiggerman M, Slater A, Bury B, Hawkins K and Firth B** (2013) Disclaimer labels on fashion magazine advertisements: Effects on social comparison and body dissatisfaction. *Body Image* **10**, 45–53. https://doi.org/10.1016/j.bodyim.2012.08.001

**van den Bosch A, Bogers T and de Kunder M** (2016) Estimating search engine index size variability: A 9-year longitudinal study. *Scientometrics* **107**, 839–856. https://doi.org/10.1007/s11192-016-1863-z

**Venkatesh S, Ramachandra R, Raja K, Spreeuwers L, Veldhuis R and Busch C** (2020) Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 280–289.

**Wade KA, Garry M, Read JD and Lindsay DS** (2002) A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin and Review* **9**, 597–603. https://doi.org/10.3758/bf03196318

**Wade KA, Green SL and Nash RA** (2010) Can fabricated evidence induce false eyewitness testimony? *Applied Cognitive Psychology* **24**, 899–908. https://doi.org/10.1002/acp.1607

**Wakefield J** (2022) Deepfake presidents used in Russia-Ukraine war. *BBC News*, 18 March. Available at https://www.bbc.co.uk/news/technology-60780142 (accessed 1 May 2022).

**Walter N and Tukachinsky RH** (2020) A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it?. *Communication Research* **47**, 155–177. https://doi.org/10.1177/0093650219854600

**Wang C** (2019) Deepfakes, revenge porn, and the impact on women. *Forbes*, 1 November. Available at https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and-the-impact-on-women/ (accessed 1 May 2022).

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

**Wright DS, Wade KA and Watson DG** (2013) Delay and déjà vu: Timing and repetition increase the power of false evidence. *Psychonomic Bulletin and Review* **20**, 812–818. https://doi.org/10.3758/s13423-013-0398-z

**Yu N, Skripniuk V, Abdelnabi S and Fritz M** (2021) Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14448–14457. https://doi.org/10.48550/arXiv.2007.08457

**Zhou Y and Shen L** (2022) Confirmation bias and the persistence of misinformation on climate change. *Communication Research* **49**, 500–523. https://doi.org/10.1177/00936502211028049

**Sophie Nightingale** is a Lecturer in Psychology at Lancaster University. Her main research interest concerns the intersection of technology with human cognition, particularly in security, legal, and forensic contexts. Her work includes drawing on psychological and computational techniques to examine the manipulation of content and improve its detection.

**Kimberley Wade** is a Professor of Psychology at the University of Warwick. Her research examines episodic and autobiographical memory distortions, and the implications for practitioners working in legal and clinical settings. Her research is published in many high-impact journals, and appears frequently in the media, undergraduate texts, and popular books.

**Cite this article:** Nightingale SJ, Wade KA (2022). Identifying and minimising the impact of fake visual media: Current and future directions. *Memory, Mind & Media* **1**, e15, 1–13. https://doi.org/10.1017/mem.2022.8