Applied Cognitive Psychology

WILEY

SHORT PAPER OPEN ACCESS

An Average Joe, a Laptop, and a Dream: Assessing the Potency of Homemade Political Deepfakes

Gillian Murphy[1,2] | Didier Ching[1,2] | Eoghan Meehan[1] | John Twomey[1,2] | Aaron Bolger[1] | Conor Linehan[1,2]

[1]School of Applied Psychology, University College Cork, Cork, Ireland | [2]Lero, the Science Foundation Ireland Research Centre for Software, Limerick, Ireland

Correspondence: Gillian Murphy (gillian.murphy@ucc.ie)

Received: 4 November 2024 | **Revised:** 26 February 2025 | **Accepted:** 4 April 2025

Funding: This work was supported by the financial support from the Research Ireland grant 13/RC/2094_2 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero—the Research Ireland Centre for Software (www.lero.ie).

Keywords: deepfake | false memory | misinformation | political misinformation

ABSTRACT

Academic and media commentary suggests that deepfake videos are problematic because they are both more easily created and more potent than previous forms of misinformation. Surprisingly, there is little research that experimentally tests these claims. In this study, we tasked a first-year undergraduate student with quickly creating political deepfakes using easily available online tools. We experimentally compared the effectiveness of misinformation delivered through those deepfake videos against misinformation delivered through text and synthetic audio format ($N = 443$). Deepfakes were effective at planting false memories for fabricated political scandals and, in some cases, reduced reported voting intention by up to 20%. However, they were not consistently more effective than simple text. In a follow-up study ($N = 300$), we confirmed that we effectively debriefed participants and caused no lasting measurable changes to their beliefs or memories. We encourage further critical study of the novel properties of deepfake technology.

1 | Introduction

As machine learning-based technologies have increasingly proliferated in recent years, researchers have sought to understand and predict various harms that may come about through the increasing and everyday use of those technologies (Gamage et al. 2022; Lee et al. 2024; Mink et al. 2024). Deepfakes are one such technology, where deep learning methods can fabricate media of a person doing or saying almost anything. There is a general consensus among commentators that deepfake technology may present a grave political/societal threat (Gamage et al. 2022) and may sway the electorate (Diakopoulos and Johnson 2021), increase scepticism of government (Pawelec 2022), or even instigate or influence military conflicts (Twomey et al. 2023). Of course, the use of technology to manipulate images is not a new phenomenon, but the common consensus is that deepfakes are especially dangerous for two reasons (Twomey et al. 2024). First, they are comparatively accessible, typically created via open-source software that can be used without specialist training or equipment (Winter and Salter 2020). Second, deepfakes are considered especially realistic, with their realism heightening the epistemic threat they present (Fallis 2021). When you combine these two factors—more accessible and more convincing—you have a potentially powerful and concerning technological innovation (Kietzmann et al. 2020).

It is quite surprising to note that these concerns about accessibility and potency are not evidenced in existing experimental literature. An increasing number of studies have investigated the effects of deepfake impersonation of political figures, though many do not compare the effects of deepfakes to a non-deepfake alternative (Ching et al. Forthcoming). For

example, Dobber et al. (2021) found that viewing a deepfake of a politician making an insensitive joke negatively affected attitudes toward him. However, the study had only two conditions; participants either saw the offensive deepfake or a control video with real, neutral statements. Therefore, the use of deepfake technology to present this misinformation may have been immaterial. Without presenting the same misinformation in text format, we cannot know whether deepfakes are *uniquely* powerful in this regard. The relatively small number of studies that have presented the same misinformation about public figures and events in multiple formats typically find that deepfakes are no more potent than simple text (e.g., Hameleers et al. 2022; Murphy, de Saint Laurent, et al. 2023; Murphy, Maher, et al. 2023).

Furthermore, studies that assess the effects of deepfakes typically do not do so in a way that takes advantage of the alleged "accessibility" of existing deepfake tools, typically repurposing existing deepfakes taken from the internet. Choice is often limited here—for example, one deepfake made by Buzzfeed that featured former US President Barrack Obama calling Donald Trump a "dipshit" has featured in many studies (Appel and Prietzel 2022; Murphy and Flynn 2022; Vaccari and Chadwick 2020). This approach is not ideal and participants' familiarity with the video may compound the experiment. Alternatively, researchers create their own deepfakes by partnering with professional companies, a process that is often described as arduous and expensive. For example, Hameleers et al. (2024) described the efforts they expended to create deepfake materials: "we relied on an extensive collaboration with computer vision and AI-experts to create the deepfake from scratch. A professional voice actor was hired to enact the deceptive speech. Based on GANs [Generative Adversarial Networks] and weeks of training, a computer model was trained to create synthetic media of the depicted political actor." (p. 5). Incidentally and despite these extensive efforts, Hameleers et al. (2024) found no difference in credibility between disinformation presented as text or presented as a professionally made deepfake.

We argue that there is an inherent paradox here: Researchers frequently write papers that strongly argue for more research on deepfakes, given the accessibility of the technology and the potential for harm (Chesney and Citron 2019; Gamage et al. 2022). Yet, these same papers do not feature custom deepfakes prepared by the research team, as presumably, creating their own deepfakes was too difficult or time-consuming or just intimidating for nontechnically skilled individuals (though see Shin and Lee 2022). If the creation of a convincing deepfake is beyond the reach of professional researchers who may study technology and human–computer interaction for a living, then how can we claim that this technology is accessible to an average person? At present, we have no clear understanding of the harm that may be caused by "normal people" making political deepfakes.

1.1 | The Current Study

We recruited a first-year undergraduate student to serve as our "Average Joe"—meaning an ordinary person or an average member of the public. We gave him no training or equipment, but simply asked him to use the internet to create the best political deepfakes he could. We documented this process and then evaluated the effects of these deepfakes, relative to other forms of political misinformation, drawing from previous claims about what makes deepfakes so uniquely threatening:

i. False memories: Building on the idea that "seeing is believing," some have argued that "seeing is remembering"—that people might come to misremember deepfaked events as having really happened in the past (Liv and Greenbaum 2020; Murphy and Flynn 2022). There is a rich literature demonstrating how individuals can indeed form false memories for political events that never occurred (Murphy et al. 2019, 2021; Sacchi et al. 2007), and Murphy, de Saint Laurent, et al. (2023) and Murphy, Maher, et al. (2023) found that viewing a deepfake clip of Will Smith starring in a purported remake of The Matrix resulted in 39% of participants reporting false memories of this remake actually being released in cinemas. However, participants who simply read a text-based description of the remake were just as likely to form a false memory for this fabricated event (45%), indicating that there may be nothing especially potent about deepfakes in this regard- in line with older research on false memories for doctored photos (Garry and Wade 2005).

ii. Political opinions: Perhaps the most feared and commonly discussed potential harm of political deepfakes is that they will be used to sway voter opinions (Diakopoulos and Johnson 2021). Yet non-deepfake studies have demonstrated that video is not necessarily more persuasive than text (Wittenberg et al. 2021) and there is limited evidence that deepfakes are uniquely effective at changing attitudes (Dobber et al. 2021).

iii. Voting intention: In misinformation research generally, behavior (or the antecedent behavioral intention) is almost never measured (Murphy, de Saint Laurent, et al. 2023), and this holds true for deepfake research also. Recent work has highlighted the limited impact of a single exposure to fake news stories on attitudes and behavioral intentions (Aftab and Murphy 2022; de Saint Laurent et al. 2022; Greene and Murphy 2021). However, despite this, there is a tendency in the literature to suggest that deepfakes may be used to influence elections by defaming candidates, creating false statements of support, or to undermine trust in the entire election process (Diakopoulos and Johnson 2021).

iv. Suspicion: Because deepfake media is often indistinguishable for many viewers from real media (Mai et al. 2023), individuals may be slower to suspect they are consuming misinformation and to engage in critical thinking when they view misinformation as a deepfake rather than as simple text. This can be assessed by measuring participant suspicions about the purpose of a study before they are told they may have seen deepfakes, but to our knowledge no study has assessed this outcome across formats.

v. Detection: Perhaps, difficulty in prompted detection rates is a unique characteristic of deepfakes that might make them more threatening as misinformation. Tahir et al. (2021) found that when presented with real videos, 88% of participants correctly identified them as

authentic, but when presented with deepfakes made by DeepFaceLab, just 25% of participants correctly identified them as manipulated.

Here we report a study assessing the potency of deepfakes created by an Average Joe (Experiment 1) and the effectiveness of our debriefing procedures (Experiment 2).

2 | Experiment 1 Method

Materials and data for both experiments are available at https://osf.io/aw68t/?view_only=b6a7d0f8e39c471f8a4f3c26a41bb6fe.

2.1 | Participants

Of the 443 participants, 379 ($n = 379$) were recruited via Prolific and compensated for their time, whereas others ($n = 64$) were recruited via university student emails. Participants ranged in age from 18 to 88 ($M = 38.30$, SD = 11.83). Women made up 60% of the sample ($n = 270$), whereas the remaining participants were men ($n = 171$) or nonbinary ($n = 2$). All were native English speakers and were living in Ireland.

2.2 | The Journey of an Average Joe

We recruited a student intern (Eoghan Meehan, a co-author of this publication) to assume the role of an Average Joe. We provided him with some fake news stories the research team had developed and asked him to create an audio and video deepfake of the target politicians saying these things. Eoghan is a first-year university student studying both Psychology and Computing. He had never made a deepfake before but knew what a deepfake was. Although he had taken some Computer Science modules, none were related to any skills required to make deepfakes.

Eoghan started by browsing Google and Reddit for about an hour. He opted to use browser-based tools, as he found them more accessible than desktop apps. It is notable that at no point did Eoghan need to download any software or write or edit any code whatsoever. Below are the details of how he created each form of media.

2.2.1 | Audio

Eoghan used the generative AI software PlayHT (https://play.ht/) to create the audio files. This was one of the first search results to appear when Eoghan Googled "voice clone" and unlike some other similar services, it does not require users to submit an audio clip of the person whose voice you are cloning agreeing to their voice being used. Eoghan uploaded 30 s of audio of the target and typed the fake quote. He then continuously modified the audio by experimenting with the punctuation of the text until the voice sounded more realistic. The entire process, including sourcing and uploading the real audio clips, took less than an hour and incurred no fees.

2.2.2 | Videos

Eoghan explored the subreddit r/aivideo (https://www.reddit.com/r/aivideo/), which has detailed recommendations and guides for creating deepfake videos. The software he used was found in 'TOOLS LIST' on the landing page of the subreddit—though note that these lists are updated so often that within 2 weeks, the tools we used were no longer included in this list. Eoghan found two promising options, one of which was free-to-use and one which required payment. We included both in the study:

2.2.2.1 | Paid. The paid service Eoghan utilized was Sync Labs, with a reasonable monthly fee of $19 (https://synclabs.so/). Eoghan uploaded a real video of each of the targets, along with the fake audio tracks, and created two deepfake videos. Eoghan spent some time trying out different settings and tinkering with the videos, so in all, this process took approximately three and a half hours.

2.2.2.2 | Free. Eoghan also used a free service called Gooey (https://gooey.ai/Lipsync/). This worked very similarly to Sync Labs and so the entire process of making two videos with Gooey took around 40 min.

2.3 | Materials

2.3.1 | Fake Stories

We developed and piloted multiple stories about Irish politicians, seeking to achieve a balance of plausibility and negative perceptions. We selected two fake stories, one relating to Simon Harris (current Irish Taoiseach (head of state) and leader of the government party Fine Gael) and Mary Lou McDonald (leader of Sinn Féin, a party currently in opposition). Both stories related to housing, which has been consistently ranked amongst Irish voters' top concerns, as Ireland is currently experiencing a severe housing crisis. The Simon Harris story built on common criticisms of Fine Gael as being dismissive of the working-class reality, and the Mary Lou McDonald story built on common criticisms of Sinn Féin as being too socialist (Figure 1).

2.3.2 | True Stories

All participants saw one true filler story about another leader of a major political party. The story concerned Micheál Martin visiting Gaza in April 2024 and criticizing Israel, accompanied by a real image of the tour.

All participants also saw one true control story about either Harris or McDonald—whichever politician that did not feature in the fake story shown to that participant. The purpose of the control story was to serve as a neutral baseline to measure opinions toward the politician, to compare to those who saw the misinformation. The control story was therefore chosen to be as neutral as possible. It simply stated that "[politician] is to lead [party] into the next election" and described the politician confirming they will continue leading the party into the next

FIGURE 1 | The two fabricated news stories that were used in the study. Participants saw one of the two stories. The stories were always shown with the headline and underneath the headline was either the story shown above (text condition), an audio file (audio condition), or a video (deepfake condition—screenshots shown above).

Irish General Election (projected to take place approximately 6–9 months after this study was conducted). The story was presented as text, with no accompanying media.

2.3.3 | Dependent Variables

False memory was assessed via the question "Do you remember this event?" after each news story. Participants could respond (a) Yes, I have a clear memory of this, (b) Yes, I have a vague memory of this, (c) No, I don't remember this occurring, or (d) I'm not sure. As in similar previous work (Murphy et al. 2019; Greene et al. 2021), answers (a) and (b) were coded as remembering.

Political opinions were assessed via a novel four-item scale, as in previous work (Hameleers and Marquart 2023). The items were "I like [politician] personally," "I think [politician] would make a good Taoiseach," "I would vote for [politician] if I could," and "I would like to see [politician's party] in power." The scale was scored on a Likert scale where 1 = *strongly disagree* and 7 = *strongly agree*. The scale had excellent reliability ($\alpha = 0.94$).

Voting intention was assessed using the scale item "I would vote for [politician] if I could," rated on a 1–7-Likert scale.

Suspicion was measured at the end of the study, prior to debriefing. Participants were asked to tell us what they believed the study was about, from a list that included the correct answer and eight credible possibilities that did not involve misinformation or doctored media.

Fake news identification was measured after telling participants about the real purpose of the study. Participants were asked whether they believed they had seen any doctored media and were presented with a list of all the stories they had seen.

2.4 | Procedure

The survey was conducted online using Qualtrics survey software. After providing consent, participants completed some demographic questions. Participants then viewed three news stories presented in random order—one true filler story, one fake story about their target politician, and one neutral story about their nontarget politician. Each story was followed by an attention check asking participants what the news story was about, and then the memory, political opinions, and voting intention items. Participants then reported their suspicions about the study, before attempting to identify any fake stories they may have seen and were finally debriefed. Participation took around 8 min in total.

2.5 | Ethics

Ethical approval was granted for this study by our Institutional Ethics Committee. We employed debriefing methods that have been found to be effective in prior research (Greene and Murphy 2023; Greene et al. 2024), which included having participants engage in identifying the fake story in an active way, clearly identifying the fake story, explaining the purpose of the study, and

reassuring participants that false memories are entirely normal. Furthermore, we followed up with participants to assess whether the debriefing had been effective (see Study 2 in Section 4). Previous research has suggested that participants often enjoy experiments where they are misled or have false memories planted, despite the deception involved (Murphy, de Saint Laurent, et al. 2023; Murphy, Maher, et al. 2023; Murphy et al. 2020). Notably, at the end of Study 1, when participants were asked if they wanted to tell us anything else, a number of participants commented on how they had enjoyed the study ("really enjoyed this study") and how they had appreciated the careful debriefing ("the reassurance you provided is excellent"). No participants reported any negative feedback and none wished to withdraw their consent.

3 | Experiment 1 Results

Most participants (96%) passed the attention checks that followed the presentation of misinformation. Those that failed the attention check ($n = 16$) were removed from the analyses, leaving 437 participants in total.

3.1 | Did Synthetic Media Induce False Memories for Fake Political Stories?

Participants formed false memories at rates of 6% for text, 14% for audio, 14% for the free deepfake, and 25% for the paid deepfake, X^2 (3, $N = 427$) = 15.21, $p = 0.002$, $V = 0.19$. Looking at the individual stories separately, there were slightly different patterns of results, as shown in Figure 2. For the Simon Harris story, the rates of false memory were 7% for text, 17% for audio, 8% for the free deepfake, and 30% for the paid deepfake, X^2 (3, $N = 213$) = 13.67, $p = 0.003$, $V = 0.25$. For the Mary Lou McDonald story, the rates of false memory were 6% for text, 11% for audio, 20% for the free deepfake, and 20% for the paid deepfake, X^2 (3, $N = 214$) = 7.32, $p = 0.062$, $V = 0.19$.

3.2 | Did Synthetic Media Sway Voter Attitudes?

For the Simon Harris story, there was a significant but small effect of media condition on attitudes, $F(4, 425) = 4.76$, $p < 0.001$, $\eta^2 = 0.04$. Relative to the control condition ($M = 14.21$, SD = 6.94), the audio condition reported significantly worse opinions of Harris ($M = 10.27$, SD = 6.62) ($p = 0.002$). The other conditions were not significantly different from the controls; text ($M = 12.03$, SD = 7.45, $p = 0.212$), free deepfake ($M = 12.15$, SD = 6.74, $p = 0.307$), paid deepfake ($M = 11.40$, SD = 6.54, $p = 0.075$). For the Mary Lou McDonald story, there was no significant effect of condition $F(4, 420) = 1.17$, $p = 0.323$, $\eta^2 = 0.01$. Similar opinions were reported among controls ($M = 13.48$, SD = 7.35), text ($M = 11.40$, SD = 7.49), audio ($M = 12.54$, SD = 7.17), free deepfake ($M = 12.04$, SD = 6.91), and paid deepfake ($M = 13.17$, SD = 6.65).

3.3 | Did Synthetic Media Sway Voting Intentions?

For the Simon Harris story, there was a significant effect of condition on voting intention $F(4, 426) = 4.93$, $p < 0.001$, $\eta^2 = 0.05$. Relative to the control condition ($M = 3.40$, SD = 1.94), the audio ($M = 2.35$, SD = 1.74) and paid video ($M = 2.60$, SD = 1.68) conditions reported significantly lower intentions to vote for Simon Harris. This equates to an approximate reduction of voting intention of 31% and 23% for audio and paid deepfake, respectively. The text ($M = 2.79$, SD = 1.93) and free deepfake ($M = 2.79$, SD = 1.80) conditions were not significantly different to controls.

For the Mary Lou McDonald story, there was no significant effect of condition on voting intention $F(4, 421) = 0.66$, $p = 0.616$, $\eta^2 = 0.01$. Similar opinions were reported across the conditions; controls ($M = 3.18$, SD = 2.03), text ($M = 2.76$, SD = 2.06), audio ($M = 3.04$, SD = 2.14), free deepfake ($M = 2.84$, SD = 1.83), and paid deepfake ($M = 2.98$, SD = 1.86) (Figure 2).

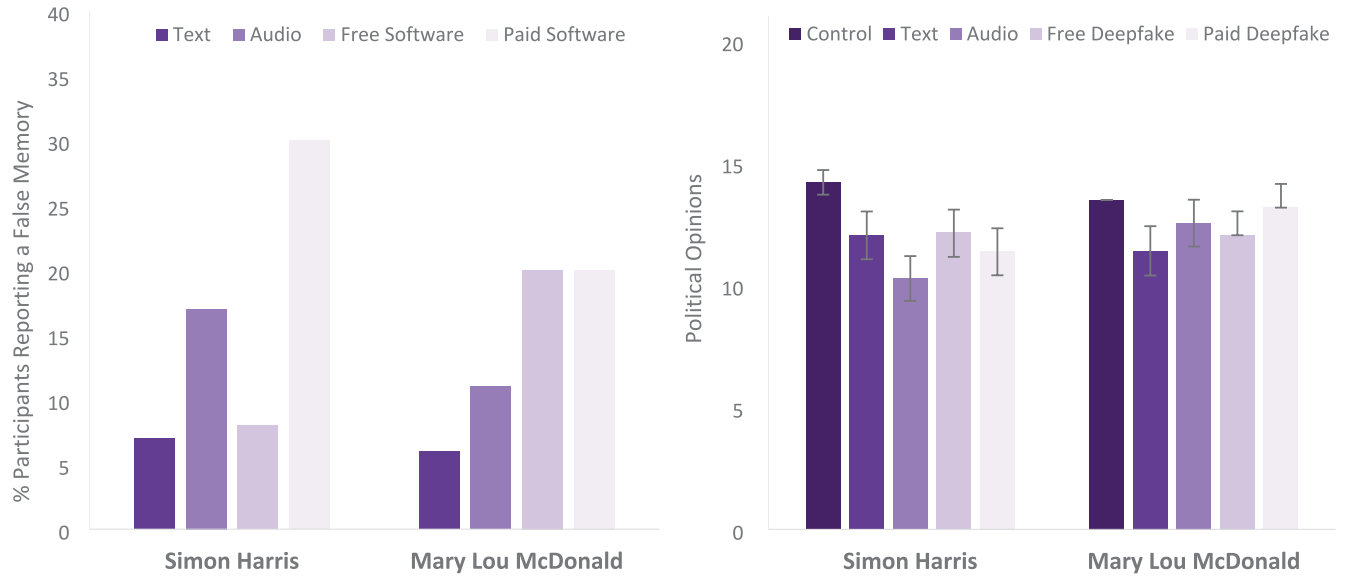

FIGURE 2 | False memory rates (left) and political opinions (right) across the conditions, for the two target politicians. Error bars represent standard error (for voting opinions only, as the false memory rates are count data).

3.4 | Did Synthetic Media Raise Voter Suspicions?

When asked what the study was about, 14% of participants correctly guessed the study was investigating misinformation or doctored media. Detection was more common in the free deepfake condition (23% detected) than in the other conditions: text (11%) audio (14%), paid deepfake (9%); X^2 (3, $N = 427$) = 9.69, $p = 0.021$, $V = 0.15$.

3.5 | Did Synthetic Media Affect Prompted Fake News Identification?

When prompted, the Simon Harris story was correctly selected as fake by 78% of participants, whereas the Mary Lou McDonald story was correctly selected as fake by 76% of voters. The (true) filler story about a third politician visiting Gaza was also selected as fake by 33% of participants.

There was no effect of condition on detection rates, though for both stories, the lowest detection rates were seen in the paid deepfake condition. Identification rates in the Simon Harris story were: text (81%), audio (77%), free deepfake (87%), and paid deepfake (69%), X^2 (3, $N = 213$) = 5.22, $p = 0.157$, $V = 0.16$. Identification rates in the Mary Lou McDonald story were: text (78%), audio (79%), free deepfake (76%), and paid deepfake (71%), X^2 (3, $N = 214$) = 0.91, $p = 0.823$, $V = 0.07$.

4 | Study 2

We conducted a follow-up study to assess if our debriefing procedures were effective. We strongly feel that misinformation researchers have a responsibility to ensure that our practices are ethical and that responsibility does not end the moment ethical approval is granted. It has been argued that follow-up assessment is best practice for misinformation studies, though still rare in reality (Murphy and Greene 2023). A recent review indicated that less than 1% of empirical misinformation studies assess the effectiveness of their debriefing methods (Greene et al. 2022). Follow-up is important to ensure participants are not left misinformed after taking part in a deepfake study, but equally, because our deepfakes featured real public figures, we feel we have an ethical obligation to ensure that their reputation is not damaged as a result of being a target of one of our studies.

5 | Experiment 2 Method

5.1 | Participants

All Prolific participants were invited to complete the follow-up study 1 week after Study 1. The study remained open for 5 days. In total, 300 participants completed the follow-up. Their ages ranged from 18 to 76 ($M = 39.34$, SD = 10.84). Women made up the majority of the sample ($n = 191$, 64%), whereas the remaining participants were men ($n = 107$) or nonbinary (2). Those who completed Study 2 had similar false memory rates in Study 1 (14.6%) as those who declined to participate (14%), X^2 (1, $N = 443$) = 0.04, $p = 0.850$, $V = 0.01$.

5.2 | Materials and Procedure

The study was almost identical to Study 1, with the same outcome measures, except that participants in the follow-up study saw both of the fake stories, along with the repeated filler story about a politician visiting Gaza and a novel filler story (about Green Party leader Roderic O'Gorman). The full survey can be seen in our OSF files.

6 | Experiment 2 Results

6.1 | Did Participants Retain False Memories of the Fake Stories?

A McNemar's test found no significant difference, with similarly low false memory rates for repeated fake stories ($M = 2\%$) and novel fake stories ($M = 2\%$) ($p = 0.99$). For the true stories, however, there was an effect of repetition, with higher memory rates for the repeated true story (33%) than for the novel true story (13%) ($p < 0.001$).

6.2 | Did Misinformation Exposure Affect Political Opinions & Voting Intentions?

At follow-up, we found no significant difference in opinions of Simon Harris between those who had seen misinformation about him in Study 1 ($M = 12.74$, SD = 6.77) and those who had not ($M = 13.83$, SD = 7.21), $t(298) = 1.35$, $p = 0.178$, $d = 0.16$. Likewise, there was no significant difference in opinions of Mary Lou McDonald between those who had seen misinformation about her in Study 1 ($M = 12.19$, SD = 7.09) and those who had not ($M = 12.56$, SD = 7.09), $t(298) = 0.44$, $p = 0.661$, $d = 0.05$. We also found no significant difference in voting intention between those who had seen misinformation about Simon Harris ($M = 2.92$, SD = 1.84) and those who had not ($M = 3.25$, SD = 2.00), $t(298) = 1.50$, $p = 0.135$, $d = 0.17$, or between those who had seen misinformation about Mary Lou McDonald ($M = 2.89$, SD = 1.92) and those who had not ($M = 3.06$, SD = 2.04), $t(298) = 0.76$, $p = 0.450$, $d = 0.09$. Though these effects were all nonsignificant, it is worth noting that on every measure, those who had seen the misinformation in Study 1 reported feeling less favorable toward that candidate.

6.3 | Could Participants Now Identify the Fake Stories?

The repeated true story was still incorrectly identified as fake by a sizable proportion of the sample (32%). The new true story was identified as fake by 28% of participants. The fake stories were identified by most participants (Simon Harris 86%, Mary Lou McDonald 93%), with higher identification rates for the repeated fake stories (92%) than for the novel fake stories (86%) ($p = 0.024$).

7 | Discussion

The current study provides evidence that deepfake technology has become truly accessible to the average internet user.

With just a few hours and no training or equipment, a first-year undergraduate student was capable of generating deceptive deepfake misinformation. Participants were mostly unaware of the true purpose of the study, and even when prompted, were not more likely to identify the deepfake version of the fake story than the text or audio versions. This is an important contribution, as previous research has mostly utilized publicly available deepfakes (Appel and Prietzel 2022; Murphy and Flynn 2022; Vaccari and Chadwick 2020) or worked with professional companies or paid voice actors to generate new materials (Groh et al. 2024; Hameleers et al. 2024). Our findings suggest that the paid deepfake service was more convincing in some respects than the free service, but both were effective.

To date, most deepfake studies have assessed attitudes and beliefs rather than other measures like voting intention or memory—a criticism that applies to the field of misinformation research generally (Murphy, de Saint Laurent, et al. 2023; Murphy, Maher, et al. 2023). Most studies have found that deepfakes are generally no more potent at swaying political opinions than less technically advanced forms of misinformation (Hameleers 2024) and our findings are in line with this consensus. However, in a novel finding, when we assessed voting intentions, we found that one of our deepfake videos reduced willingness to vote for the target by 23%, and indeed, the fabricated audio alone reduced voting intentions by 31%. Behavioral intentions are not the same as actual behaviors (Greene and Murphy 2021), but this provides some evidence that homemade deepfakes could potentially affect elections—a longstanding concern (Chesney and Citron 2019; Diakopoulos and Johnson 2021).

Relative to text, fake stories presented as deepfakes more than tripled the false memory rate, with participants coming to remember that these fabricated political events really occurred. This is at odds with previous work that has found similar false memory rates for fake stories presented as text or deepfakes (Murphy and Flynn 2022; Murphy, de Saint Laurent, et al. 2023; Murphy, Maher, et al. 2023). Both of those studies utilized publicly available deepfakes taken from Youtube, which presents obvious issues when trying to ascertain whether they can mislead participants' memory, as participants may have seen them before. The current study suggests further research is warranted to better understand the effects of entirely novel deepfakes on memory.

The findings of our debriefing follow-up were reassuring in many ways and in line with previous research (Murphy and Greene 2023)—participants reported fewer false memories and a better ability to identify the fake stories at follow-up—but there was a nonsignificant trend for reduced support for the target politicians. It is possible that a larger sample size would have allowed us to detect what may have been a small but persistent negative effect of participation. This may be due to sleeper effects, where the misinformation effects grew stronger over time (as in Murphy et al. 2021), or continued-influence effects, where the misinformation continued to affect opinion despite debriefing (as in Lewandowsky et al. 2012). Further careful research is required to ensure that research in this field is conducted ethically and any harm to participants (or the deepfake target) is minimized (De Ruiter 2021).

7.1 | Implications

For researchers, this study has two major implications. Firstly, that creating novel, effective deepfake materials is not beyond the capabilities of any research team, and experimenters should move beyond reusing publicly available video (Appel and Prietzel 2022; Vaccari and Chadwick 2020; Murphy, de Saint Laurent, et al. 2023; Murphy, Maher, et al. 2023). Where researchers do make claims on the effectiveness of deepfakes (e.g., how deceptive they are, how they sway opinions, and so forth) they should clearly distinguish whether they are describing accessible deepfake technology or professional-grade materials. Secondly, researchers should take care to conduct deepfake research in an ethical manner, ideally publishing some evidence that their debriefing procedures were effective (Murphy and Greene 2023).

For science communicators, journalists and policy makers, this study has some more troubling implications. Here we provide clear evidence that creating reasonably potent deepfakes is now a very simple task. Although previous studies have encouraged raising awareness of deepfakes and training individuals to detect them (Tahir et al. 2021), others have shown that raising awareness of deepfakes can result in greater distrust of real media (Ternovski et al. 2022), also known as "the liar's dividend" (Twomey et al. 2023). We argue that we must also consider the possibility that raising awareness may direct individuals to engage with this technology, where they can easily create and spread convincing misinformation. This is significant, as a recent international survey indicated that just 28% of participants knew at least a little bit about deepfakes (Umbach et al. 2024). Obviously, the solution here is not to attempt to keep deepfake technology a secret, but we recommend researchers incorporate values-driven interventions into any awareness efforts, as there is some tentative evidence that this approach can encourage more responsible use of deepfake technology (Naffi et al. 2023).

7.2 | Strengths and Limitations

This study's strengths include the novelty and inclusion of a follow-up to assess the ethical procedures employed. However, this study has important limitations and is just a single study in an Irish context only. Participants only viewed the fake stories once, whereas multiple presentations can increase belief (illusory truth effect; Fazio 2020). Importantly, participants were also forced to view these stories. In a natural environment, after our Average Joe made convincing materials, he would also have had to effectively share them to get them to have a broad impact on voters, a task with many challenges (Chen et al. 2023). We also acknowledge that our data speaks only to the current state of deepfakes and it is possible that better and more effective deepfake technology may be "just around the corner." However we argue that AI is a permanently not-yet-realized technology (Reeves 2022) and deepfakes first emerged in 2017 (Twomey et al. 2024), thus it is surely not premature to evaluate the effectiveness of the technology that is currently available, even

if future iterations may (or may not) be more potent (Salovaara et al. 2017).

7.3 | Conclusion

Overall, the current study serves as a litmus test for the present-day accessibility and potency of deepfake technology. We encourage other researchers to remain critical and evidence-based in their claims about emerging technologies and resist dystopian narratives about what an emerging technology "might" do in the near future, when they are making claims about what the technology may do today (Broinowski 2022).

Author Contributions

Gillian Murphy: conceptualization, investigation, funding acquisition, writing – original draft, methodology, formal analysis, supervision. **Didier Ching:** conceptualization, methodology, writing – review and editing. **Eoghan Meehan:** methodology, investigation, writing – review and editing. **John Twomey:** conceptualization, methodology, writing – review and editing. **Aaron Bolger:** conceptualization, writing – review and editing, methodology. **Conor Linehan:** writing – review and editing, conceptualization.

Ethics Statement

The study reported in this manuscript received ethical approval from the Social Research Ethics Committee—University College Cork, Ireland.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All materials and data for the current project are available at https://osf.io/aw68t/?view_only=b6a7d0f8e39c471f8a4f3c26a41bb6fe.

References

Aftab, O., and G. Murphy. 2022. "A Single Exposure to Cancer Misinformation May Not Significantly Affect Related Behavioural Intentions." *HRB Open Research* 5, no. 82: 82.

Appel, M., and F. Prietzel. 2022. "The Detection of Political Deepfakes." *Journal of Computer-Mediated Communication* 27, no. 4: zmac008.

Broinowski, A. 2022. "Deepfake Nightmares, Synthetic Dreams: A Review of Dystopian and Utopian Discourses Around Deepfakes, and Why the Collapse of Reality May Not Be Imminent—Yet." *Journal of Asia-Pacific Pop Culture* 7, no. 1: 109–139.

Chen, S., L. Xiao, and A. Kumar. 2023. "Spread of Misinformation on Social Media: What Contributes to It and How to Combat It." *Computers in Human Behavior* 141: 107643.

Chesney, B., and D. Citron. 2019. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107: 1753.

Ching, D., J. Twomey, M. Aylett, M. Quayle, C. Linehan, and G. Murphy. Forthcoming. "Can Deepfakes Manipulate Us? Assessing the Evidence via a Critical Scoping Review." *PLoS One*.

De Ruiter, A. 2021. "The Distinct Wrong of Deepfakes." *Philosophy and Technology* 34, no. 4: 1311–1332.

de Saint Laurent, C., G. Murphy, K. Hegarty, and C. M. Greene. 2022. "Measuring the Effects of Misinformation Exposure and Beliefs on Behavioural Intentions: A COVID-19 Vaccination Study." *Cognitive Research: Principles and Implications* 7, no. 1: 87.

Diakopoulos, N., and D. Johnson. 2021. "Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections." *New Media & Society* 23, no. 7: 2072–2098.

Dobber, T., N. Metoui, D. Trilling, N. Helberger, and C. De Vreese. 2021. "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" *International Journal of Press/Politics* 26, no. 1: 69–91.

Fallis, D. 2021. "The Epistemic Threat of Deepfakes." *Philosophy and Technology* 34, no. 4: 623–643.

Fazio, L. K. 2020. "Repetition Increases Perceived Truth Even for Known Falsehoods." *Collabra: Psychology* 6, no. 1: 38.

Gamage, D., P. Ghasiya, V. Bonagiri, M. E. Whiting, and K. Sasahara. 2022. "Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

Garry, M., and K. A. Wade. 2005. "Actually, a Picture Is Worth Less Than 45 Words: Narratives Produce More False Memories Than Photographs Do." *Psychonomic Bulletin and Review* 12, no. 2: 359–366.

Greene, C. M., C. de Saint Laurent, G. Murphy, T. Prike, K. Hegarty, and U. K. Ecker. 2022. "Best Practices for Ethical Conduct of Misinformation Research." *European Psychologist* 28: a000491.

Greene, C. M., and G. Murphy. 2021. "Quantifying the Effects of Fake News on Behavior: Evidence From a Study of COVID-19 Misinformation." *Journal of Experimental Psychology: Applied* 27, no. 4: 773–784.

Greene, C. M., and G. Murphy. 2023. "Debriefing Works: Successful Retraction of Misinformation Following a Fake News Study." *PLoS One* 18, no. 1: e0280295.

Greene, C. M., R. A. Nash, and G. Murphy. 2021. "Misremembering Brexit: Partisan Bias and Individual Predictors of False Memories for Fake News Stories Among Brexit Voters." *Memory* 29, no. 5: 587–604.

Greene, C. M., K. M. Ryan, L. Ballantyne, et al. 2024. "Unringing the Bell: Successful Debriefing Following a Rich False Memory Study." *Memory and Cognition* 52: 1–14.

Groh, M., A. Sankaranarayanan, N. Singh, D. Y. Kim, A. Lippman, and R. Picard. 2024. "Human Detection of Political Speech Deepfakes Across Transcripts, Audio, and Video." *Nature Communications* 15, no. 1: 7629.

Hameleers, M. 2024. "Cheap Versus Deep Manipulation: The Effects of Cheapfakes Versus Deepfakes in a Political Setting." *International Journal of Public Opinion Research* 36, no. 1: edae004.

Hameleers, M., and F. Marquart. 2023. "It's Nothing but a Deepfake! The Effects of Misinformation and Deepfake Labels Delegitimizing an Authentic Political Speech." *International Journal of Communication* 17: 21.

Hameleers, M., T. G. van der Meer, and T. Dobber. 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, no. 3: 20563051221116346.

Hameleers, M., T. G. van der Meer, and T. Dobber. 2024. "They Would Never Say Anything Like This! Reasons to Doubt Political Deepfakes." *European Journal of Communication* 39, no. 1: 56–70.

Kietzmann, J., L. W. Lee, I. P. McCarthy, and T. C. Kietzmann. 2020. "Deepfakes: Trick or Treat?" *Business Horizons* 63, no. 2: 135–146.

Lee, H. P., Y. J. Yang, T. S. Von Davier, J. Forlizzi, and S. Das. 2024. "Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks." In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

Lewandowsky, S., U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13, no. 3: 106–131.

Liv, N., and D. Greenbaum. 2020. "Deep Fakes and Memory Malleability: False Memories in the Service of Fake News." *AJOB Neuroscience* 11, no. 2: 96–104. https://doi.org/10.1080/21507740.2020.1740351.

Mai, K. T., S. Bray, T. Davies, and L. D. Griffin. 2023. "Warning: Humans Cannot Reliably Detect Speech Deepfakes." *PLoS One* 18, no. 8: e0285333.

Mink, J., M. Wei, C. W. Munyendo, et al. 2024. "It's Trying Too Hard to Look Real: Deepfake Moderation Mistakes and Identity-Based Bias." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20. Association for Computing Machinery.

Murphy, G., C. de Saint Laurent, M. Reynolds, et al. 2023. "What Do We Study When We Study Misinformation? A Scoping Review of Experimental Research (2016–2022)." *Harvard Kennedy School Misinformation Review*.

Murphy, G., and E. Flynn. 2022. "Deepfake False Memories." In *Memory Online*, 112–124. Routledge.

Murphy, G., and C. M. Greene. 2023. "Conducting Ethical Misinformation Research: Deception, Dialogue, & Debriefing." *Current Opinion in Psychology* 54: 101713. https://doi.org/10.1016/j.copsyc.2023.101713.

Murphy, G., E. Loftus, R. H. Grady, L. J. Levine, and C. M. Greene. 2020. "Fool Me Twice: How Effective Is Debriefing in False Memory Studies?" *Memory* 28, no. 7: 938–949.

Murphy, G., E. F. Loftus, R. H. Grady, L. J. Levine, and C. M. Greene. 2019. "False Memories for Fake News During Ireland's Abortion Referendum." *Psychological Science* 30, no. 10: 1449–1459.

Murphy, G., L. Lynch, E. Loftus, and R. Egan. 2021. "Push Polls Increase False Memories for Fake News Stories." *Memory* 29, no. 6: 693–707.

Murphy, G., J. Maher, L. Ballantyne, et al. 2023. "How Do Participants Feel About the Ethics of Rich False Memory Studies?" *Memory* 31, no. 4: 474–481. https://doi.org/10.1080/09658211.2023.2170417.

Naffi, N., M. Charest, S. Danis, et al. 2023. "Empowering Youth to Combat Malicious Deepfakes and Disinformation: An Experiential and Reflective Learning Experience Informed by Personal Construct Theory." *Journal of Constructivist Psychology* 38: 1–22.

Pawelec, M. 2022. "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions." *Digital Society* 1, no. 2: 19.

Reeves, S. 2022. "Navigating Incommensurability Between Ethnomethodology, Conversation Analysis, and Artificial Intelligence." *arXiv Preprint* arXiv:2206.11899.

Sacchi, D. L., F. Agnoli, and E. F. Loftus. 2007. "Changing History: Doctored Photographs Affect Memory for Past Public Events." *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21, no. 8: 1005–1022.

Salovaara, A., A. Oulasvirta, and G. Jacucci. 2017. "Evaluation of Prototypes and the Problem of Possible Futures." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2064–2077.

Shin, S. Y., and J. Lee. 2022. "The Effect of Deepfake Video on News Credibility and Corrective Influence of Cost-Based Knowledge About Deepfakes." *Digital Journalism* 10, no. 3: 412–432.

Tahir, R., B. Batool, H. Jamshed, et al. 2021. "Seeing Is Believing: Exploring Perceptual Differences in Deepfake Videos." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

Ternovski, J., J. Kalla, and P. Aronow. 2022. "The Negative Consequences of Informing Voters About Deepfakes: Evidence From Two Survey Experiments." *Journal of Online Trust and Safety* 1, no. 2: 28.

Twomey, J., D. Ching, M. P. Aylett, M. Quayle, C. Linehan, and G. Murphy. 2023. "Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets That Discuss Deepfakes in the Russian Invasion of Ukraine." *PLoS One* 18, no. 10: e0291668.

Twomey, J., D. Ching, M. P. Aylett, M. Quayle, C. Linehan, and G. Murphy. 2024. "What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology." *IEEE Transactions on Technology and Society* 6: 64–79.

Umbach, R., N. Henry, G. F. Beard, and C. M. Berryessa. 2024. "Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries." In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1-20.

Vaccari, C., and A. Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media+ Society* 6, no. 1: 2056305120903408.

Winter, R., and A. Salter. 2020. "DeepFakes: Uncovering Hardcore Open Source on GitHub." *Porn Studies* 7, no. 4: 382–397.

Wittenberg, C., B. M. Tappin, A. J. Berinsky, and D. G. Rand. 2021. "The (Minimal) Persuasive Advantage of Political Video Over Text." *Proceedings of the National Academy of Sciences* 118, no. 47: e2114388118.