Do Content Warnings Help People Spot a Deepfake? Evidence from Two Experiments

Andrew Lewis*[1], Patrick Vu[2], Raymond M. Duch[1], and Areeq Chowdhury[3]

[1] *University of Oxford*
[2] *Brown University*
[3] *The Royal Society*

Abstract. The advent and rapid advancement of ‘deepfake’ videos — so named as they are fake videos made to look real with the use of deep learning artificial intelligence programs — pose serious challenges to our digital information environment. As the technology continues to improve and fake videos proliferate, there is uncertainty about how people will discern genuine from manipulated videos, and how this will affect trust in online content. This paper conducts a pair of experiments aimed at gauging the public's ability to detect deepfakes from ordinary videos, and the extent to which content warnings improve detection of inauthentic videos. In the first experiment, we consider capacity for detection in natural environments: that is, do people spot deepfakes when they encounter them without a content warning? In the second experiment, we present the first evaluation of how warning labels affect capacity for detection, by telling participants at least one of the videos they are to see is a deepfake and observing the proportion of respondents who correctly identify the altered content. Our results show that, without a warning, individuals are no more likely to notice anything out of the ordinary when exposed to a deepfake video of neutral content (32.9%), compared to a control group who viewed only authentic videos (34.1%). Second, warning labels improve capacity for detection from 10.7% to 21.6%; while this is a substantial increase, the overwhelming majority of respondents who receive the warning are still unable to tell a deepfake from an unaltered video. A likely implication of this is that individuals, lacking capacity to manually detect deepfakes, will need to rely on the policies set by governments and technology companies around content moderation.

*Corresponding author: andrew.lewis@univ.ox.ac.uk. This research was conducted with financial support from the The Royal Society. We thank participants at the Social Media and Experimentation Workshop at Oxford's Centre for Experimental Social Science for helpful comments. The views expressed herein are solely the responsibility of the authors, and all errors are our own. *This version:* January 18, 2022.

1. Introduction

Fitting in with wider trends of digital mis- and dis-information, deepfakes have the potential to further fracture the shared basis of truth in an already polarised society. Because of the still-nascent state of the technology, it is difficult to predict how deepfakes of high profile individuals or events may be deployed and received. The possibility of bad actors using this technology for any number of deceptive ends — from *revenge porn* and identity fraud to terrorism and election manipulation — has prompted both the FBI (2021) and Europol (2020) to issue warnings, with the latter calling deepfakes "perhaps the most immediately tangible, damaging application" of artificial intelligence (p.52).

These warnings underscore the dual threat of deepfakes: first, the technology provides a new and potentially persuasive means of spreading lies; second, the difficulty of manually detecting real from fake videos (i.e., with the naked eye) threatens to lower the information value of video media entirely. As people internalise deepfakes' capacity to deceive, they will rationally place less trust in all online videos, including authentic content (Fallis, 2021). This may also be exacerbated by what is known as the *Liar's Dividend* — that is, genuine videos being written off as deepfakes by those with an interest in discrediting them (Chesney and Citron, 2019). The result in both cases is increased uncertainty, which can bolster motivated information processing and belief formation (Kunda, 1990; Dieckmann et al., 2017). This is why Europol warns, somewhat ominously, that deepfakes could "undermine the possibility of a reliable shared 'reality'" (p.53).

As concerns about the veracity of videos spread, it will fall on regulators and technology companies to serve as moderators of authenticity. If individuals lack the ability to discern deepfakes from genuine videos manually, detection algorithms will be the only way to consistently and accurately flag fake content. Whilst the technical tools for detection are proving effective — with one such model achieving as high as 98.2% accuracy (Kaur et al., 2020) — it remains unclear what the best approach is for dealing with deepfakes on prominent social media sites. (For example, one might imagine that videos intended to incite violence would be treated rather differently than those used for entertainment or satire.) As with text- and photo-based misinformation, potential remedies include labeling deepfakes with content warnings or removing them from platforms entirely. The former approach has gained traction in recent years, with both Facebook and Twitter making efforts to flag fake or misleading content during the COVID-19 pandemic (Sharevski et al., 2022).

To date, however, little is understood about how content warnings may be deployed for deepfakes, and the extent to which they will affect individuals' capacity to detect deepfakes on their own. This is a critical question. If individuals cannot detect deepfakes even with content warnings, it will make faith in the judgments of content moderators essential. Such is the purpose of this paper: in a pair of experiments described below, we test how likely UK residents are to spot a deepfake from a genuine video in both natural contexts (i.e., without a warning) and when they have been given a content warning.

Our findings in Experiment 1 confirm past research showing people struggle to identify deepfakes from genuine videos without a warning. Experiment 2 presents, to our knowledge, the first evidence on the impact of content warnings on manual detection. We find that participants who are issued a direct warning that one of the videos they will see is a deepfake correctly identify the deepfake in 21.6% of cases, compared to 10.7% in the control group. Thus even with a direct warning, the vast majority of people (78.4%) still cannot distinguish the deepfake from authentic content. Notably, we find that correct detection of deepfakes is uncorrelated with almost all characteristics we observe, including self-reported confidence in detection abilities, familiarity with the actor depicted in the deepfake, gender, and level of social media use. The only characteristic which significantly correlates with detection is age, with older participants better able to identify the deepfake.

These experiments show that human discernment is largely inadequate in detecting deepfakes, even when participants are directly warned that the content they view may have been altered. A practical interpretation of Experiment 2 is that — unlike how accuracy prompts and other interventions can help individuals better spot textual misinformation — warning labels do not enable individuals to simply look closer and see the irregularities on their own. As such, successful content warnings on deepfakes will rely on trust in moderators' judgments, raising concerns that any such warnings may be written off as politically motivated or biased.

The rest of this paper proceeds as follows. Section 2 reviews the literature on manual deepfake detection. Section 3 develops a theoretical model examining the impact of content warnings on detection. Section 4 outlines our experimental design, while Section 5 examines the results thereof. Section 6 concludes with a brief discussion of policy implications.

2. Literature Review

A small but growing number of studies have aimed to assess people's capacity to detect deepfakes from genuine videos. Overall, this research shows that the public is not very adept at detecting deepfakes. For example, using a now-famous deepfake in which former President Obama calls then-President Trump a "total and complete dipshit," Vaccari and Chadwick

(2020) find that the manipulated clip deceives about 15 percent of respondents into believing Mr Obama actually made this (rather inflammatory) statement, while only half of the sample recognise the statement as untrue. Thus, roughly half of people are deceived by or uncertain about the veracity of the deepfake, despite the fact that the content of the video itself is, as the authors write, 'highly improbable'.

Using a pair of clips of a fictional US politician (one with a real actor, one with a deepfake rendering of the actor), Ternovski et al. (2021) similarly observe that participants are unable to tell which is the manipulated clip, and generally believe the content of the statement irrespective of whether they see the real video or the deepfake. Moreover, even when participants are informed about the existence of deepfakes before seeing the videos, they are not any more likely to correctly identify the fake video. Instead, the effect of this information is only to reduce trust in all video content uniformly, irrespective of the veracity of an individual clip. This is in line with Fallis's (2021) model of how deepfakes will undermine the information value placed in videos writ large.

Studying how deepfakes may be deployed by political operatives to undermine their opponents, (Dobber et al., 2020) find deepfakes that are tailored to the "susceptibilities of the receiver" are not only likely to be perceived as genuine, but negatively affect respondents' perceptions of and attitudes toward the politician in question. Similarly, Barari et al. (2021) find that deepfakes "can convince the American public of scandals that never occurred." The authors argue, however, that deepfakes are no more likely to deceive than other types of mis- and dis-information.

Our paper builds on this literature by examining the role of content warnings in improving detection. It is, to our knowledge, the first study that experimentally estimates the impact of direct content warnings on detection of deepfakes. Additionally, in contrast to most previous research, the content of the deepfake we use is relatively neutral, to avoid participants dismissing a video out of hand because of its improbable content, rather than its quality.

3. Theoretical Model

In this section, we develop a simple Bayesian model for understanding the potential impact of content warnings. Consider a viewer watching a video V that is either authentic ($V = 1$) or inauthentic ($V = 0$). Before watching the video, the viewer has a prior subjective probability π_F that the video is inauthentic. After watching a video, the viewer forms beliefs about whether the video is authentic ($B = 1$) or inauthentic ($B = 0$). This judgement may be based on the characteristics of the specific viewer and video in question, and vary, for example, with the quality of the video, whether it is in fact authentic or inauthentic, or whether it contains a content warning.

We are interested in the probability that a video is inauthentic when it is believed to be inauthentic by the viewer – that is, the probability that the viewer has correctly detected the deepfake. Define $d_{T|T} \equiv P(B = 1|V = 1)$ as the probability that the viewer believes the video is authentic after viewing it, conditional on it being authentic; and $d_{F|F} \equiv P(B = 0|V = 0)$ as the probability that the viewer believes the video is inauthentic when it is in fact inauthentic. The quantities $d_{T|T}$ and $d_{F|F}$ measure the ability of the viewer to accurately detect authentic and inauthentic content. Suppose that these detection rates are functions of whether or not the video contains a content warning, which is denoted by a binary variable C. Then the probability of being correct when believing a video to be inauthentic for a Bayesian decision-maker is given by

$$P(V = 0|B = 0) = \frac{d_{F|F}(C)\pi_F}{d_{F|F}(C)\pi_F + (1 - d_{T|T}(C))(1 - \pi_F)}$$

From this simple model, we can derive a number of predictions about the impact of deepfakes and the effectiveness of content warnings.

First, the model highlights the importance of detection abilities. The probability of holding correct beliefs about an inauthentic video is increasing in both the improved detection of authentic videos ($\partial P(V = 0|B = 0)/\partial d_{T|T} > 0$) and the improved detection of inauthentic videos ($\partial P(V = 0|B = 0)/\partial d_{F|F} > 0$). When detection is as good as random, $d_{F|F} = d_{T|T} = 1/2$, the probability that a video is inauthentic given one's belief that it is reduces to π_F. That is, the viewer must rely entirely on prior beliefs that the video is true. It follows that improvements in detection can lead to more accurate beliefs among viewers. In Experiment 1, we estimate the detection rate of false videos, that is $d_{F|F} = P(B = 0|V = 0)$.

Second, the model can be used to highlight both the potential benefits and pitfalls of content warnings. Consider a social media platform that decides to implement content warnings C on uploaded videos that are deemed likely to be false. Suppose further that these warnings are imperfect (or perceived to be imperfect): while inauthentic videos are more likely to be flagged as deepfakes, authentic videos are now sometimes flagged as potential deepfakes. Content warnings therefore improve one margin of detection ($\partial d_{F|F}/\partial C > 0$) while degrading the other ($\partial d_{T|T}/\partial C < 0$). The net effect of content warnings on the probability of the viewer holding correct beliefs is therefore ambiguous, and depends on whether improvements from the former outweigh the costs of the latter.

In Experiment 2, we test whether content warnings improve detection. That is, we test if

$d_{F|F}(C = 1) > d_{F|F}(C = 0)$. We also provide more suggestive evidence that increased knowledge of deepfakes is correlated with more distrust in online content, in the sense that more authentic videos are erroneously identified as inauthentic i.e. $d_{T|T}(C = 1) < d_{T|T}(C = 0)$. Note that the presence of content warnings could in principle decrease detection abilities for authentic videos for two reasons: (i) the specific video in question erroneously has a content warning, or (ii) the presence of content warnings on the platform reduces detection abilities for authentic videos more generally, that is, even for videos without content warnings. The second effect highlights the possibility that content warnings could engender wider distrust in all online content.

4. Experimental Design

Our design focuses on two key questions surrounding the public's capacity to detect deepfakes with and without warnings. First, do people notice anything out of the ordinary when they encounter a deepfake in a natural environment? Second, when directly warned they will see at lease one deepfake in a set of videos, are people able to tell if a specific video is real or fake? We address these questions using a pair of survey experiments with three experimental conditions.

4.1. Experiment 1.

In the control (C1), participants watch five unaltered videos and are asked whether they have noticed anything out of the ordinary. In the first treatment (T1), participants watch four of these same five unaltered videos plus a deepfake, then answer the same series of questions. In both cases, if individuals answer "yes" to noticing something out of the ordinary, they are further asked to indicate in which specific video(s) they thought something amiss, and to provide a short explanation of why. The comparison between C1 and T1 serves to answer the first question about deepfake detection without warnings: if people are alert to irregularities in deepfake videos in natural settings, we would expect to see a statistically significant difference between these two groups in the proportion of participants reporting something out of the ordinary. We refer to this comparison between C1 and T1 throughout the paper as Experiment 1.

4.2. Experiment 2.

In the second experiment, we examine how content warnings affect capacity for detection of false content. We compare two groups who see the same five videos, one of which is a deepfake. The difference between the groups is that one receives a content warning (T2), while the other does not (T1). (Note that the group which does not receive the content warning, T1, is the treatment group in Experiment 1.) Participants are first briefly informed what deepfakes are ("manipulated videos that use deep learning artificial intelligence to make fake videos that appear real") and then told that at least one of the five videos they will see is a deepfake. They are then asked to select which video(s) they believe are fake. Only those participants who select the deepfake and only the deepfake are counted as having correctly distinguished the fake video from the genuine content. We then compare the proportion of participants correctly detecting the deepfake in T2, who receive a content warning, with the proportion who detect it in T1, who view the same videos but who receive no content warning. In T1, only those who detect something out of the ordinary are given the choice to select which video (or videos) are suspect. Those who do not detect something out of the ordinary, despite having seen a deepfake, are counted as not detecting the fake video. For ease, we refer to the comparison between T1 and T2 as Experiment 2.

4.3. Videos. We make use of a deepfake video of the American actor Tom Cruise created and made public by the by VFX artist Chris Ume. The clip is shown alongside a series of authentic video clips of Mr. Cruise from publicly available YouTube channels. To control for past familiarity, all participants also watch a one minute excerpt of an interview with Mr. Cruise to provide a baseline acquaintance with the actor's appearance and speech patterns. All six videos can be viewed in our Supplementary Materials.

4.4. Procedure. We recruit a sample (n=1,093) of UK-based participants through Lucid Marketplace, which provides subject pools balanced on key demographics. Past research evaluating Lucid's data quality suggests the platform is in line with other online sample providers and is "suitable for evaluating many social scientific theories" (Coppock and McClellan, 2019). Participants complete a Qualtrics-based survey. We exclude any potential participants who are under 18, not residents of the UK, or are unwilling to provide informed consent to participate. Participants are randomised into one of the three experimental treatments. In line with our pre-registration plan, all participants must pass an attention check to be included in the study. We also remove subjects who have seen the deepfake video previously. This removes 5 respondents from Experiment 1 and 8 from Experiment 2.

4.5. Descriptive statistics across both experiments. Descriptive statistics of the experiment participants from all three treatment groups are presented in Table 1. The first column reports the mean value of various characteristics in C1. Randomisation implies that participants in all three groups should be similar in both observed (and unobserved) attributes. Columns 2 and 3 provides a check on the randomisation process by showing the difference in means for observed characteristics between the T1 and T2 relative to C1. None of the means is statistically different at the 5 per cent level, which is consistent with successful randomisation. The descriptive statistics show that the average age of participants is around 45, roughly half are female, and around 36% were previously aware of deepfakes.

Participants in T2 viewed the set of 5 videos where one was a deepfake. They were told at least one of the videos was fake and asked if it was obvious which video (or videos) were inauthentic. The proportion of participants who said it was obvious was 31.9%. A greater share responded that it was not obvious, at 38.3%, and the remaining 29.8% were not sure. As we show later, those who stated it was obvious were no more likely than others to select the correct video.

Table 1: Descriptive statistics

	C1 mean	T1 vs. C1	T2 vs. C1
Age	45.82	-0.512	-1.108
	[51.916]	[-0.409]	[-0.898]
Female	0.507	0.069	0.049
	[19.238]	[1.863]	[1.329]
Aware of deepfakes	0.357	0.004	0.065
	[14.148]	[0.118]	[1.786]
Proficient with social media (1-10)	6.025	-0.138	0.148
	[42.606]	[-0.706]	[0.78]
Internet use (1-10)	6.416	-0.19	-0.109
	[49.606]	[-1.021]	[-0.615]
Familiar with actor (1-10)	7.219	0.021	-0.05
	[53.57]	[0.111]	[-0.272]
Observations	361	354	365

Note: Difference in means in T1 and T2 relative to C1. Results are from an OLS regression of the row variables on an indicator for T1 and T2 for columns 2 and 3 respectively. Below each estimate in square brackets are t-values.

5. Results

The main results are summarised in Figure 1. The first graph shows that, without a content warning, individuals are no likelier to spot something out of the ordinary when exposed to a deepfake compared with a control group who see only authentic videos. The second graph shows the effect of content warnings on propensity for manual detection: individuals who receive no content warning correctly identify the deepfake in about 10% of cases, while those who are warned the content they see may be altered are twice as likely to detect it. Detection rates are nonetheless quite low in both groups, with a substantial majority of participants failing to detect the deepfake, irrespective of whether or not they are warned. The following sections elaborate on these results and present additional findings.

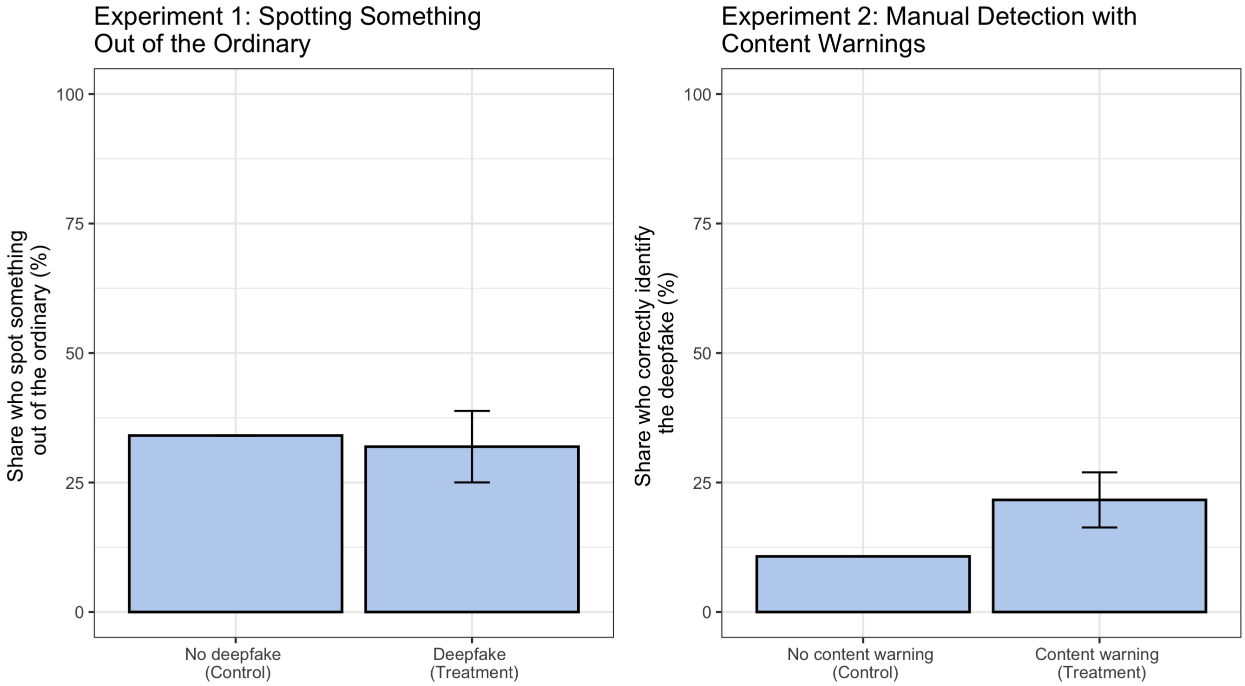

Figure 1: Summary of main results.

Notes: Whiskers show the 95 percent confidence interval calculated from a regression of the indicator outcome variable on an indicator for being in the treatment group using robust standard errors. Further details are available in Tables 2 and 4.

5.1. Experiment 1. The first experiment focuses on people's capacity to detect deepfake videos in a natural setting. In other words, do people spot something amiss when they encounter a deepfake without a content warning? Participants are randomised into two groups, C1 and T1. Participants in C1 watch five authentic videos, while those in T1 watch four of the same authentic video and a deepfake. All participants are then asked if they have found anything to be "out of the ordinary" in the videos they watched. If participants respond *Yes* to this question, they are then asked to indicate the video(s) in which they found something out of the ordinary.

5.1.1 Spotting something out of the ordinary. Table 2 present results from regressing an indicator for detecting something out of the ordinary from a set of videos on a treatment indicator for whether one of the 5 videos was a deepfake. Specifically, we regress

$$y_i = \alpha + \beta T_i + x_i'\delta + \epsilon_i$$

where y_i is an indicator for reporting something is out of the ordinary in the set of videos viewed; T_i is an indicator for viewing a set of videos that include the deepfake; x_i is a vector of individual characteristics; and ϵ_i is an unobserved error term. The coefficient of interest, β, measures the adjusted mean difference between C1 and T1 in detecting something out of the ordinary. Standard errors for all regressions are robust to heteroscedasticity.

The results in column 1 show that participants are not able to detect something out of the ordinary when videos include a deepfake. In the control group, C1, where all videos were authentic, only 34.1% of participants report having noticed anything out of the ordinary. In the treatment group, T1, where one video was a deepfake, the fraction of participants reporting anything out of the ordinary was nearly identical, at 32.9%. The difference is statistically insignificant at the 5 per cent level.

Column 2 includes a set of control variables in regression. The estimated treatment effect remains close to zero and statistically significant, as we would expect with randomisation. However, the probability of reporting something out of the ordinary does vary with observable characteristics. Younger participants and those who were previously aware of deepfakes are more likely to report something is out of the ordinary. In particular, a 10-year reduction in age is associated with a 5.3 percentage point increase in reporting something is out of the ordinary. Those who were previously aware of deepfake technology were 11.6 percentage points more likely to report something out of the ordinary.

Table 2: OLS Regressions: Detection without Content Warnings

	Dependent variable:	
	Out of the ordinary	
	(1)	(2)
Treatment (T1)	−0.022	−0.016
	(0.035)	(0.035)
Female		−0.035
		(0.036)
Age x 1/10		-0.053^{***}
		(0.011)
Aware of deepfakes		0.116^{***}
		(0.039)
Proficient with social media		0.0005
		(0.043)
High level of internet use		0.025
		(0.040)
Familiar with actor (0-10)		0.003
		(0.007)
Constant	0.341^{***}	0.531^{***}
	(0.025)	(0.090)
Observations	720	699
Adjusted R^2	−0.001	0.061

*p<0.1; **p<0.05; ***p<0.01

Note: *Aware of deepfakes* is a binary variable equal to 1 if the participant was aware of deepfakes prior to participating in the experiment. *Proficiency with social media* is a binary variable equal to one if the participant's self-report is higher than the sample median for the question: 'On a scale of 0-10 where 10 is very proficient and 0 is not proficient at all, how proficient do you consider yourself in navigating social media platforms?' Similarly, *High level of internet use* is a binary variable equal to one if the participant's self-report is higher than the sample median for the question: 'On a scale of 0-10 where 10 is a great deal and 0 is none at all, how much time do you spend on the Internet outside of work-related commitments on an average day?' Standard errors are robust to heteroscedasticity.

5.1.2 Heterogeneous treatment effects. Younger subjects and those aware of deepfake technology are more likely to report that something is out of the ordinary. This raises a natural question: are these groups better able to detect the presence of an inauthentic video, or do they simply exhibit a higher base-line level of skepticism for online content? To answer this question we regress:

$$y_i = \alpha + \gamma H_i + \beta(T_i \times H_i) + \epsilon_i$$

where H_i is an indicator variable for the dimension of heterogeneity (e.g. having prior awareness of deepfakes, or being younger than the median age). The coefficient β measures the difference in reporting something out of the ordinary between the subgroups in C1 and T1 for whom $H_i = 1$. If these subgroups are better able to detect something out of the ordinary, then $\beta > 0$. The coefficient γ measures the difference in means in C1 between those with $H_i = 0$ and $H_i = 1$ of reporting something out of the ordinary; this measures baseline difference in distrust of online content.

Table 4 presents the results. In column 1, H_i is a binary variable which equals 1 if i reports having a prior awareness of deepfakes. In column 2, $H_i = 1$ if participant i is below the median age in the sample, and 0 otherwise. The results show no statistical difference in detecting something out of the ordinary between those in C1 and T1 who were previously aware of deepfake technology and those who were below the median age. These subgroups, however, have a higher baseline probability of reporting something out of the ordinary, as was shown in Table 2. Overall, these results show that certain subgroups are more likely to express scepticism over the authenticity of online content, but they are no more or less likely to report this when actually encountering a deepfake.

Table 3: Heterogeneous Treatment Effect without Content Warnings

	Dependent variable:	
	Out of ordinary	
	(1)	(2)
Aware of deepfakes	0.144***	
	(0.048)	
Aware of deepfakes x Treatment (T1)	0.066	
	(0.062)	
Below median age		0.212***
		(0.043)
Below median age x Treatment (T1)		−0.020
		(0.053)
Constant	0.266***	0.229***
	(0.021)	(0.022)
Observations	699	699
Adjusted R^2	0.035	0.041

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Note: For definitions of the control variables, see notes in Table 2. Standard errors are robust to heteroscedasticity.

5.2. Experiment 2. Experiment 1 shows that people are not likely to spot something out of the ordinary when viewing deepfakes. Do content warnings increase detection rates? To answer this question we compare detection between T1, who received no content warning, and a treatment group T2 that did; both groups viewed the same set of 5 videos, of which one was a deepfake. The experimental variation comes from exposing only one group to the content warning. The content warning given to participants in T2 states: "On the following pages are a series of 5 additional videos of Mr Cruise, at least one of which is a deepfake video." Participants in T2 must select at least one video which they believe is a deepfake. In the control group, T1, participants receive no content warnings and are simply asked if, in the set of videos they viewed, anything appeared 'out of the ordinary'. Only those who answered in the affirmative were given the opportunity to choose which video aroused suspicion. Those who registered nothing out of the ordinary, despite having viewed false content, are counted as not having detected the deepfake. Correct detection is coded as selecting the deepfake and only the deepfake.

The first column in Table 4 shows that only 10.7% of participants in T1, who received no content warning, correctly identified the deepfake. The overwhelming majority of this group (68.1%) did not detect anything out of the ordinary to begin with. Conditional on saying something was out of the ordinary, only 33.6% identified the correct video as the deepfake. By contrast, the treatment group who received the content warning identified the deepfake in 21.6% of cases, making them about twice as likely to correctly detect the inauthentic video than the control group. The difference between these detection rates is statistically significant at the 5 per cent level. While this indicates a substantial improvement from a low base level of detection, the vast majority of participants (78.4%) who receive a content warning are still unable to correctly distinguish the deepfake from the authentic content.

How much of this increased detection is due to increased alertness to characteristics of the video, compared to simply being compelled to choose a video when receiving a content warning? A comparison which provides suggestive, albeit imperfect, evidence for this question is to compare the detection rates between T2 and the subset of participants in the T1 who detect something out of the ordinary. This is, importantly, a non-random sample who detect something out of the ordinary and are therefore likely more alert to the quality and veracity of the videos. Among this 'alert' subgroup in T1 (who do not receive a content warning) detection is in fact higher than in the treatment group T2 who receive a warning. The former identify the deepfake in 33.6% of cases compared to 21.6% in the latter, and the difference is statistically significant at the 5% level. This provides suggestive evidence that the increased overall detection from content warnings arises largely from being compelled to select a video because of the given knowledge that (at least) one is inauthentic.

The unfavourable comparison to the alert group, and the rather low detection rate overall, both provide suggestive evidence that content warnings are effective at increasing the detection rate of false videos largely by causing more general distrust in online content. An implication of this is a very large fraction of participants identify at least one of the authentic videos as inauthentic ('false positives') in the content warnings treatment (78.4%). The incidence of false positives in the control group T1 is, by comparison, much lower (21.2%) by virtue of the fact that most participants do not register anything out of the ordinary. Thus, while content warnings increase the detection rates of false videos (i.e. in the theoretical model: $d_{F|F}(C = 1) > d_{F|F}(C = 1)$), there is some suggestive evidence that this comes at the expense of decreasing detection rates of authentic videos ($d_{T|T}(C = 1) < d_{T|T}(C = 0)$).

Overall, content warnings increase detection rates of false videos from around 1 in 10 to 1 in 5. This is a sizable increase, but overall detection rates with content warnings still remain low. Moreover, the mechanism leading to higher detection rates appears to be by causing more general distrust in online content, which impairs the other side of detection, that is, correctly identifying authentic videos. Greater generalized distrust of online content is a possible outcome of content warnings that policymakers should take into account when assessing the costs and benefits of moderating online content.

5.2.1 Heterogeneity analysis. Are some groups of people better able to detect deepfakes than others? We restrict our analysis to participants in T2 because they were asked additional questions that allow us to examine more dimensions of heterogeneity.

Table 5 shows the results. Each column includes a different dimension of heterogeneity, and the final column includes all of them in a single specification. Remarkably, the results show very few observable characteristics are correlated with correct detection. The only characteristic which is positively correlated with detection is age, where *older* participants are more likely to identify the deepfake. In particular, an increase in 10 years of age is associated with an 3.3 percentage point increase in detection. By contrast, prior awareness of deepfakes, self-reported confidence in detection[1], answering whether the deepfake was 'obvious', proficiency with social media, high internet use, and familiarity with the actor, were not associated with better or worse detection. Overall, most people, regardless of their characteristics (at least the ones we observe), have low detection rates.

[1] We create a categorical variable from a 0–10 self-reported scale of confidence in detection with 3 possibilities: low confidence (0–2); medium confidence (3–7); and high confidence (8–10).

Table 4: OLS Regressions: Detection with Content Warnings

	Dependent variable:	
	Detection	
	(1)	(2)
Treatment (T2)	0.109^{***}	-0.120^{**}
	(0.027)	(0.049)
Constant	0.107^{***}	0.336^{***}
	(0.016)	(0.045)
Control sample	Full	Selected out of the ordinary
Observations	719	478
Adjusted R^2	0.020	0.012
Note:	*p<0.1; **p<0.05; ***p<0.01	

Note: Standard errors are robust to heteroscedasticity.

Table 5: Heterogeneity Analsysis with Content Warnings

	Dependent variable:								
	Detection								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Obvious: not	0.061								0.068
	(0.051)								(0.059)
Obvious: don't know	0.047								0.072
	(0.054)								(0.060)
Confidence: low		−0.009							−0.082
		(0.069)							(0.082)
Confidence: medium		−0.022							−0.087
		(0.052)							(0.057)
Male			0.033						0.019
			(0.044)						(0.046)
Aware of deepfakes				−0.015					0.018
				(0.044)					(0.050)
Age x 1/10					0.035***				0.033**
					(0.013)				(0.014)
Proficient with social media						−0.031			0.020
						(0.046)			(0.053)
High level of internet use							−0.099**		−0.074
							(0.043)		(0.051)
Familiar with actor (0-10)								0.001	0.004
								(0.044)	(0.047)
Constant	0.179***	0.231***	0.202***	0.223***	0.062	0.226***	0.249***	0.217***	0.088
	(0.036)	(0.044)	(0.028)	(0.029)	(0.058)	(0.026)	(0.028)	(0.031)	(0.084)
Observations	365	365	365	365	364	365	365	350	349
Adjusted R^2	−0.002	−0.005	−0.001	−0.002	0.017	−0.002	0.010	−0.003	0.007

*p<0.1; **p<0.05; ***p<0.01

Note: For definitions of the control variables, see notes in Table 2. Standard errors are robust to heteroscedasticity.

6. Conclusion

Our results make two main contributions: first, we confirm early research indicating people struggle to manually detect high quality deepfakes; second, we present the first evaluation of the effect of content warnings on detection, showing that the vast majority of individuals are still unable to spot a deepfake from a genuine video even when they are told the content they are viewing may have been altered. Successful content moderation — for example, with specific videos flagged as likely fake by social media platforms — will therefore depend not on enhancing the public's ability to detect irregularities in altered videos on their own, but instead on fostering trust in the judgements of moderators.

Additionally, greater awareness of the existence and deceptiveness of deepfakes (and false media more generally) is likely to increase distrust of all online videos. Improved detection of false videos may therefore come at the cost of the public distrusting a greater share of authentic content. This creates a tension. Successful moderation depends on trust, while more misinformation decreases the public's trust in media. Future research should examine how policy can best deal with this tension, and seek to advance our understanding of how individual characteristics such as ideological predisposition may affect trust in content moderation efforts.

References

Barari, Soubhik, Christopher Lucas, and Kevin Munger, "Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media," *OSF Preprints*, 2021.

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

Coppock, Alexander and Oliver A. McClellan, "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents," *Research & Politics*, 2019.

Dieckmann, Nathan F., Robin Gregory, Ellen Peters, and Robert Hartman, "Seeing What You Want to See: How Imprecise Uncertainty Ranges Enhance Motivated Reasoning," *Risk Analysis*, 2017, *37* (3), 471–486.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes Claes Vreese, "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?," *The International Journal of Press/Politics*, 2020.

Europol, "Malicious Uses and Abuses of Artificial Intelligence," 2020.

Fallis, Don, "The Epistemic Threat of Deepfakes," *Philosophy Technology*, 2021, *34*, 623–643.

FBI, "Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations," 2021.

Kaur, Sawinder, Parteek Kumar, and Ponnurangam Kumaraguru, "Deepfakes: Temporal Sequential Analysis to Detect Face-swapped Video Clips using Convolutional Long Short-term Memory," *Journal of Electronic Imaging*, 2020, *29* (3).

Kunda, Ziva, "The Case for Motivated Reasoning," *Psychological Bulletin*, 1990, *108* (3), 480–498.

Sharevski, Filipo, Raniem Alsaadi, Peter Jachim, and Emma Pieroni, "Misinformation Warnings: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes," *Computers and Security*, 2022, *114*.

Ternovski, John, Joshua Kalla, and P. M. Aronow, "Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments," *Working paper*, 2021.

Vaccari, Cristian and Andrew Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, 2020.