iScience, Volume *24*

Supplemental information

Fooled twice: People cannot detect deepfakes but think they can

Nils C. Köbis, Barbora Doležalová, and Ivan Soraperra

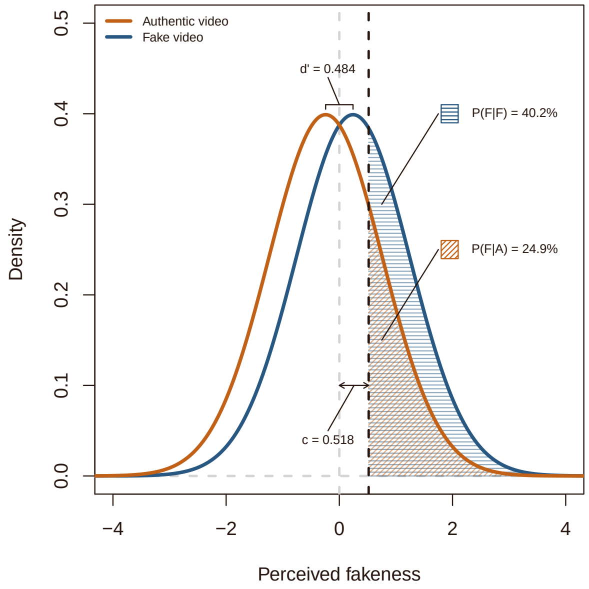
Supplemental Information

Table SI.1: Sensitivity analysis of the tests reported in the paper, related to STAR Methods

F tests - ANOVA: Fixed effects, omnibus, one-way			
Analysis:	Sensitivity: Compute required effect size		
Input:	α err prob	=	0.05
	Power $(1 - \beta)$	=	0.80
	Total sample size	=	210
	Number of groups	=	3
Output:	Noncentrality parameter λ	=	9.775
	Critical F	=	3.040
	Numerator df	=	2
	Denominator df	=	207
	Effect size f	=	0.216
t tests - Means: Difference from constant (one sample case)			
Analysis:	Sensitivity: Compute required effect size		
Input:	Tail(s)	=	Two
	α err prob	=	0.05
	Power $(1 - \beta errprob)$	=	0.80
	Total sample size	=	210
Output:	Noncentrality parameter δ	=	2.815
	Critical t	=	1.971
	Df	=	209
	Effect size d	=	0.194

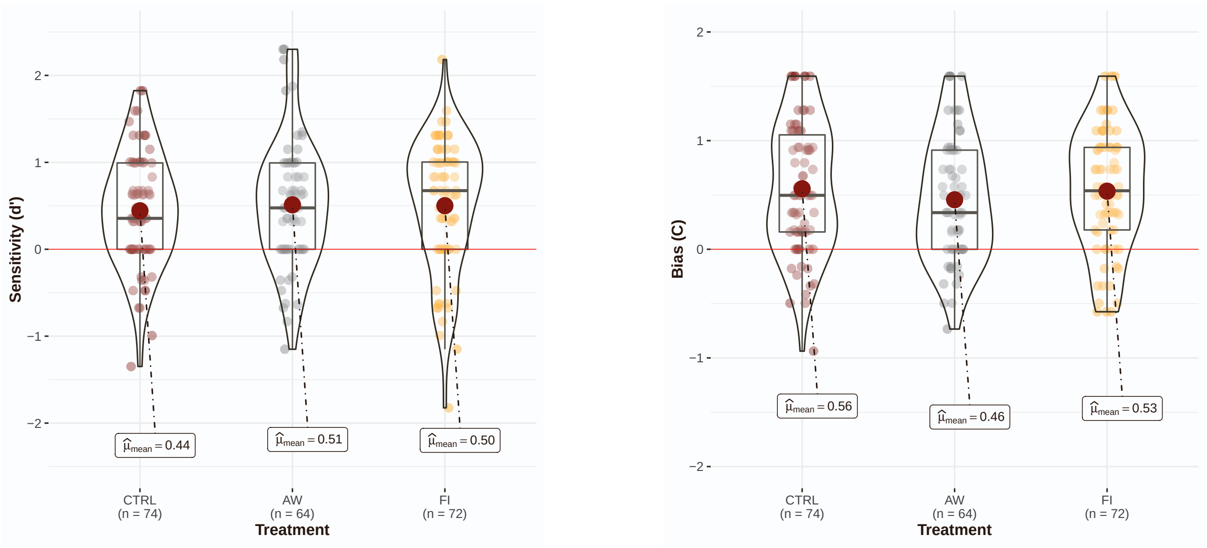
Notes: The top half of the table displays the sensitivity analysis for a one-way ANOVA indicating that a sample size of 210 participants divided in 3 groups and a power of 0.8, allows detecting a small to medium effect size (Cohen's f = 0.216). This sensitivity analysis applies to all the analyses comparing treatments—i.e., all the one-way ANOVAs—reported in the main text. The bottom half of the table displays the sensitivity analysis for a one-sample t-test indicating that a sample size of 210 participants and power of 0.8 allows detecting a small effect size (Cohen's d = 0.194). This sensitivity analysis applies to the analyses testing accuracy, bias, and overconfidence—i.e., all the one-sample t-tests—reported in the main text. Calculations were performed with G*Power version 3.1.9.6.

Figure SI.1: Signal Detection Theory — Average sensitivity and bias, related to Figure 2



Notes: The figure shows the results of a signal detection theory analysis. It displays the distribution of the perceived "fakeness" of both the authentic and fake videos. The average sensitivity d' measures the ability to discriminate between fake and authentic videos. The average bias c measures how conservative participants are when stating that a video is fake. Unbiased participants have a $c = 0$. The orange shaded area is the probability to declare "fake" when the video is authentic $P(F|A)$—i.e., the probability of a False Alarm—and the blue shaded area is the probability to declare "fake" when the video is fake $P(F|F)$—i.e., the probability of a correct hit.

Figure SI.2: Signal Detection Theory — Distribution of individual sensitivity and bias by Treatment, related to Figure 4

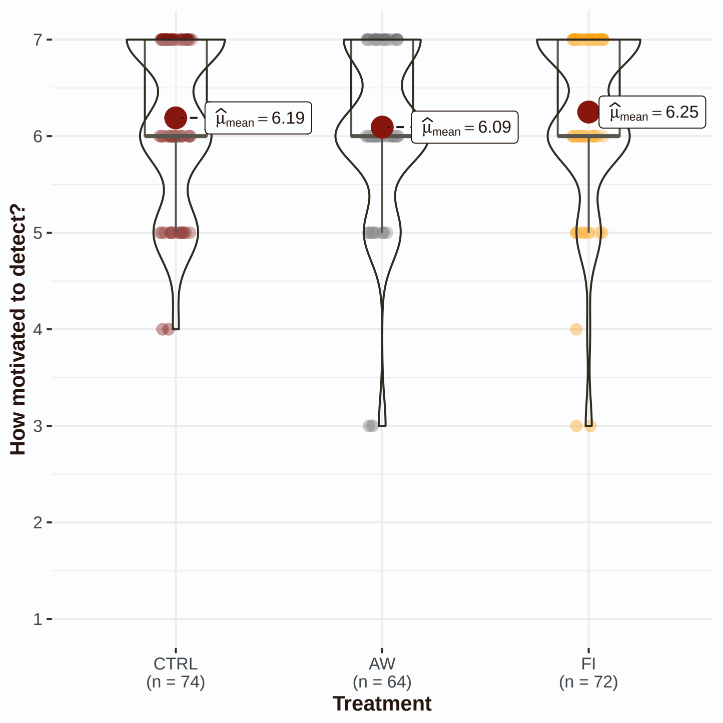


(a) Sensitivity (d')

(b) Bias (c)

Notes: Panel (a) shows violin plots of the distribution of the individual sensitivity d' (y axis) by the treatment (x axis). Panel (b) shows violin plots of the distribution of the individual bias c (y axis) by the treatment (x axis). Sensitivity and bias are calculated using the psycho package (Makowski, 2018). While black lines represent medians, the dark red dots represent means. Boxes indicate the interquartile range, each dot shows a raw data point. Plots created with the Ggstatsplot package (Patil, 2018).

Figure SI.3: Distribution of the stated motivation to correctly detect videos, related to Figure 4



Notes: Box-and-whisker plots display the distribution of the answers to the question: "How motivated were you to detect videos?" (y axis) by treatment (x axis). Motivation was measured on a 7 points Likert scale where: 1 = "not at all" and 7 = "very much". Treatments labels are: CTRL = Control treatment, AW = Awareness treatment, FI = Financial Incentives treatment. Black lines represent the medians. Dots show the raw data. Plots created with the Ggstatsplot package (Patil, 2018).

Table SI.2: Linear probability models of the likelihood to guess "fake" and to guess correctly, related to Figure 5

	Dependent variable:			
	d(guessed fake)		*d*(correct guess)	
	Mod.	Mod.	Mod.	Mod.
	(1)	(2)	(3)	(4)
d(AW tmt)	0.035	0.036	0.015	0.019
	(0.033)	(0.034)	(0.020)	(0.020)
d(FI tmt)	0.008	0.010	0.008	0.009
	(0.032)	(0.033)	(0.019)	(0.019)
d(video is fake)	0.153^{***}	0.151^{***}		
	(0.016)	(0.016)		
N. of views		−0.003		0.010
		(0.016)		(0.013)
d(verific. q.)		-0.076^{*}		0.002
		(0.040)		(0.039)
Motivation		−0.012		0.014^{*}
		(0.019)		(0.008)
Period N.		0.002		0.001
		(0.002)		(0.002)
Age		−0.002		-0.002^{***}
		(0.001)		(0.001)
d(Female)		0.004		−0.017
		(0.028)		(0.018)
d(Bachelor)		0.014		0.005
		(0.030)		(0.018)
d(Master or PhD)		−0.018		0.033
		(0.047)		(0.027)
Constant	0.236^{***}	0.546^{***}	0.569^{***}	0.639^{***}
	(0.024)	(0.127)	(0.013)	(0.067)
Videos' FE	No	Yes	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
Robust SE	Yes	Yes	Yes	Yes
Observations	3,360	3,360	3,360	3,360
Clusters	210	210	210	210
R^2	0.028^{***}	0.068^{***}	0.000	0.046^{***}

Notes: Video fixed effects (Video's FE) refer to dummy variables controlling for the characteristics of each video. Robust standard errors clustered at the individual level are in parentheses. *d* for dummy variables. The binary dependent variables "*d*(guessed fake)" and "*d*(correct guess)" are equal to 1 when the participant guesses that the video is fake and when the participant makes a correct guess, respectively. The variables "*d*(AW tmt)" and "*d*(FI tmt)" are equal to 1 in the awareness and the financial incentives treatments, respectively; "*d*(video is fake)" = 1 when the video is a deepfake; "N. of views" is how often the participant watched the video; "*d*(verific. q.)" = 1 when the participant correctly answered the question about video content; "Motivation" is the reported level of motivation to detect the videos on a scale 1-7; "Period N." refers to the round from 1-16; "Age" reported in years, "*d*(Female)" = 1 when the participant self identifies as female; "*d*(Bachelor)" & "*d*(Master or PhD)" indicate a dummy variable for education levels with High School being the reference category. Significance is coded as follows: *p<0.1; **p<0.05; ***p<0.01.

Table SI.3: Confidence in the guess made by the participant, related to Figure 6

	Dependent variable:	
	Confidence	
	Mod. (5)	Mod. (6)
d(wrong guess)	−0.543 (0.441)	−0.127 (0.451)
d(guessed fake)	−4.076*** (0.936)	−4.543*** (0.904)
d(wrong guess)×d(guessed fake)	−4.582*** (0.950)	−3.824*** (0.895)
d(AW tmt)		−1.135 (1.745)
d(FI tmt)		−0.391 (1.631)
N. of views		−4.030*** (0.798)
d(verific. q.)		−0.509 (1.286)
Motivation		3.157*** (0.993)
Period N.		0.090** (0.037)
Age		−0.036 (0.061)
d(Female)		−1.297 (1.365)
d(Bachelor)		−4.502*** (1.522)
d(Master or PhD)		−3.873* (2.332)
Constant	79.786*** (0.815)	72.775*** (6.492)
Videos' FE	No	Yes
Clustered SE	Yes	Yes
Robust SE	Yes	Yes
Observations	3,360	3,360
Clusters	210	210
R^2	0.044***	0.153***

Notes: Video fixed effects (Video's FE) refer to dummy variables controlling for the characteristics of each video. Robust standard errors clustered at the individual level are in parentheses. d for dummy variables. The dependent variable "Confidence" is the believed probability to have guessed correctly the video from 50% (random) to 100% (guessed for sure). The variables "d(wrong guess)" = 1 when the participant's guess is wrong; "d(guessed fake)" = 1 when the participant guesses that the video is a fake; "d(AW tmt)" and "d(FI tmt)" are equal to 1 in the awareness and the financial incentives treatments, respectively; "N. of views" is how often the participant watched the video; "d(verific. q.)" = 1 when the participant correctly answered the question about video content; "Motivation" is the reported level of motivation to detect the videos on a scale 1-7. "Period N." refers to the round from 1-16; "Age" reported in years, "d(Female)" = 1 when the participant self identifies as female; "d(Bachelor)" & "d(Master or PhD)" indicate a dummy variable for education levels with High School being the reference category. Significance is coded as follows: *p<0.1; **p<0.05; ***p<0.01.

Table SI.4: Linear probability models: The effect of video characteristics on the likelihood to guess "fake" and to guess correctly, related to Figure 5

	Dependent variable:	
	d(guessed fake) Mod.	*d*(correct guess) Mod.
	(7)	(8)
d(AW tmt)	0.036 (0.034)	0.019 (0.020)
d(FI tmt)	0.010 (0.033)	0.009 (0.019)
d(video is fake)	0.154*** (0.016)	
d(Female in video)	0.021 (0.017)	−0.063*** (0.017)
Dark skin color (1-6)	0.011** (0.005)	0.017*** (0.005)
d(Glasses in video)	0.006 (0.018)	0.122*** (0.024)
Brightness (1-4)	0.024 (0.016)	−0.004 (0.023)
N. of views	0.001 (0.016)	0.012 (0.012)
d(verific. q.)	−0.078* (0.041)	0.005 (0.038)
Motivation	−0.013 (0.019)	0.014* (0.008)
Period N.	0.002 (0.002)	0.0005 (0.002)
Age	−0.002 (0.001)	−0.002*** (0.001)
d(Female)	0.004 (0.028)	−0.017 (0.018)
d(Bachelor)	0.015 (0.030)	0.005 (0.018)
d(Master or PhD)	−0.017 (0.046)	0.033 (0.027)
Constant	0.310** (0.131)	0.486*** (0.094)
Videos' FE	No	No
Clustered SE	Yes	Yes
Robust SE	Yes	Yes
Observations	3,360	3,360
Clusters	210	210
R^2	0.036***	0.016***

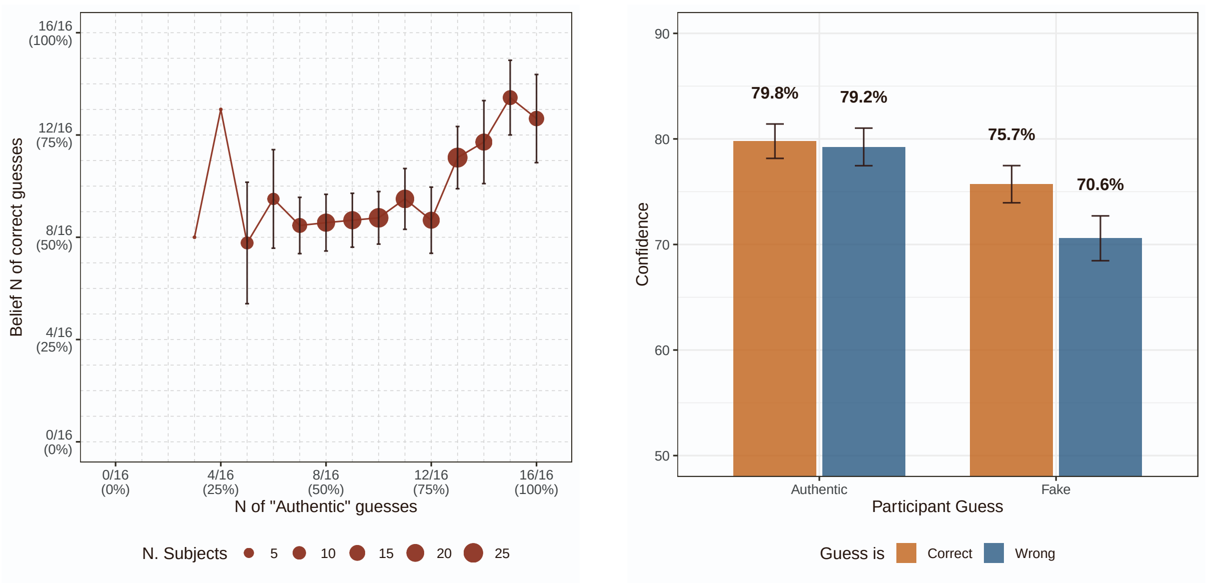
Notes: Robust standard errors clustered at the individual level are in parentheses. *d* for dummy variables. The binary dependent variables "*d*(guessed fake)" and "*d*(correct guess)" are equal to 1 when the participant guesses that the video is fake and when the participant makes a correct guess, respectively. The variables "*d*(AW tmt)" and "*d*(FI tmt)" are equal to 1 in the awareness and the financial incentives treatments, respectively; "*d*(video is fake)" = 1 when the video is a deepfake; to analyze other characteristics of the video, four independent coders naive to the purpose of the study coded all 32 videos; the coding scheme: "*d*(Female in video)" = 1 when the protagonist in the video is female (coders' rating); "Dark skin color" skin color of the protagonist in the video using Fitzpatrick Skin Type Scale (Buolamwini & Gebru, 2018) ranging from "Pale white skin" (=1) to "Dark brown or black skin" (=6), (coders' rating); '*d*(Glasses in video)" = 1 when the protagonist in the video wears glasses (coders' rating); "Brightness" how bright the video is "very dark" (=1) to "very bright" (=4), (coders' rating); "N. of views" is how often the participant watched the video; "*d*(verific. q.)" = 1 when the participant correctly answered the question about video content; participant level variables: "Motivation" is the reported level of motivation to detect the videos on a scale 1-7. 'Period N." refers to the round from 1-16; "Age" reported in years, "*d*(Female)" = 1 when the participant self identifies as female; "*d*(Bachelor)" & "*d*(Master or PhD)" indicate a dummy variable for education levels with High School being the reference category. Significance is coded as follows: *p<0.1; **p<0.05; ***p<0.01.

Table SI.5: The effect of video's characteristics on participants' confidence in the guess, related to Figure 6

	Dependent variable:
	Confidence Mod.
	(9)
d(wrong guess)	−0.538 (0.445)
d(guessed fake)	−4.069*** (0.912)
d(wrong guess)×d(guessed fake)	−4.201*** (0.942)
d(AW tmt)	−1.158 (1.746)
d(FI tmt)	−0.399 (1.634)
d(Female in video)	−0.638* (0.351)
Dark skin color (1-6)	−0.172* (0.100)
d(Glasses in video)	1.308*** (0.472)
Brightness (1-4)	0.304 (0.356)
N. of views	−4.026*** (0.776)
d(verific. q.)	−0.459 (1.284)
Motivation	3.149*** (0.992)
Period N.	0.085** (0.037)
Age	−0.034 (0.061)
d(Female)	−1.287 (1.367)
d(Bachelor)	−4.509*** (1.523)
d(Master or PhD)	−3.883* (2.337)
Constant	70.504*** (6.576)
Videos' FE	No
Clustered SE	Yes
Robust SE	Yes
Observations	3,360
Clusters	210
R^2	0.136***

Notes: Robust standard errors clustered at the individual level are in parentheses. d for dummy variables. The dependent variable "Confidence" as the believed probability to have guessed correctly the video from 50% (random) to 100% (guessed for sure). The variables "d(guessed fake)" = 1 when the guess was that the video is a fake; "d(wrong guess)" = 1 when the guess was wrong; "d(AW tmt)" and "d(FI tmt)" = 1 in the awareness and the financial incentives treatments, respectively; "d(video is fake)" = 1 when the video is a deepfake; to analyze other characteristics of the video, four independent coders naive to the purpose of the study coded all 32 videos; the coding scheme: "d(Female in video)" = 1 when the protagonist in the video is female (coders' rating); "Dark skin color" skin color of the protagonist in the video using Fitzpatrick Skin Type Scale (Buolamwini & Gebru, 2018) ranging from "Pale white skin" (=1) to "Dark brown or black skin" (=6), (coders' rating); 'd(Glasses in video)" = 1 when the protagonist in the video wears glasses (coders' rating); "Brightness" how bright the video is "very dark" (=1) to "very bright" (=4); "N. of views" how often the participant watched the video; "d(verific. q.)" = 1 when the participant correctly answered the question about video content; "Motivation" is the reported level of motivation to detect the videos on a scale 1-7; "Period N." refers to the round from 1-16; "Age" reported in years, "d(Female)" = to 1 when participant self identifies as female; "d(Bachelor)" & "d(Master or PhD)" indicate a dummy variable for education levels with High School being the reference category. Significance is coded as follows: *p<0.1; **p<0.05; ***p<0.01.

Figure SI.4: Confidence by number of “authentic” guesses and by correct guesses, related to Figure 7



(a) Beliefs by number of “Authentic” guesses

(b) Confidence by guess type

Notes: Panel (a) displays the average belief about the number of correct guesses (y axis) by the number of times the participant guessed “authentic” (x axis). The size of the bubble is proportional to the number of participants. Error bars represent confidence intervals. Panel (b) displays the average confidence (y axis)—i.e., the beliefs about the probability to have guessed correctly the video from 50% (random) to 100% (for sure)—by what the participant guessed and whether the guess was correct (x axis). Error bars represent the 95% confidence interval based on the robust standard errors clustered at the individual level reported in Table SI.3 Model 5.