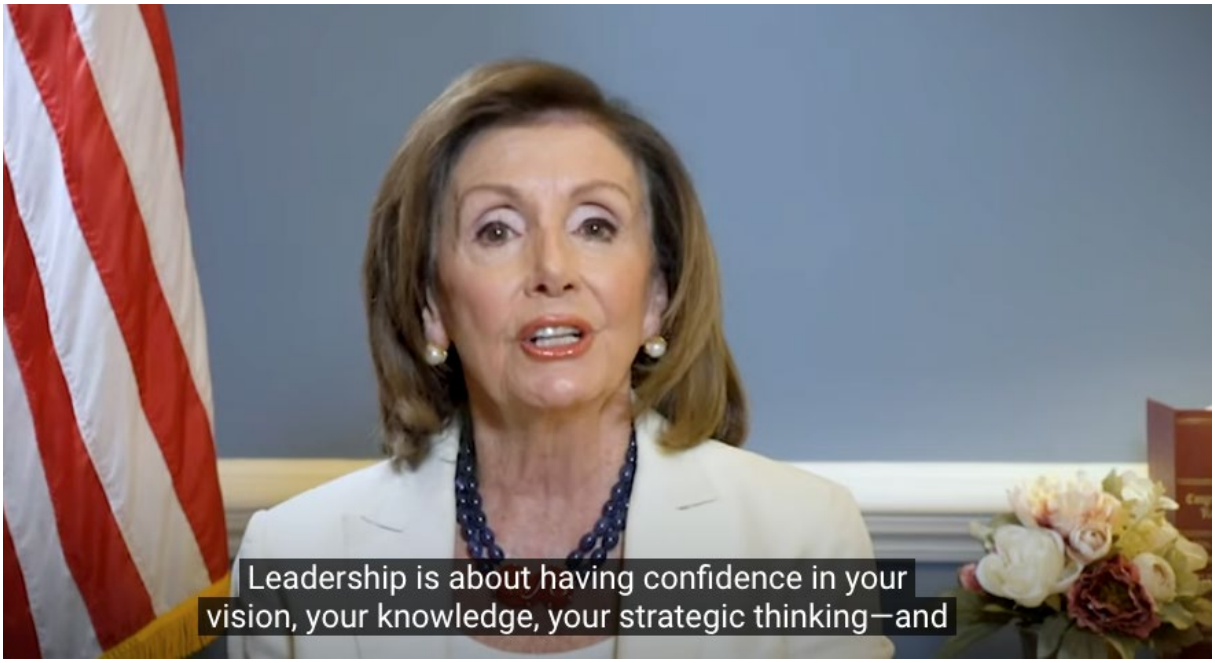
Online Appendices for Distorting the Truth versus Blatant Lies: The Effects of Different Degrees of Deception in Domestic and Foreign Political Deepfakes

Appendix A - scripts and screenshots of the deepfakes and authentic video

Note. The quality of the deepfake screenshots has been degraded for ethical considerations. The watermark in the screenshots was not presented to participants. The deepfakes are only shown in the experimental set-up and a protected environment to minimize impact and control exposure to direct corrections and refutations. Due to ethical considerations and data protection, we do not make the videos publicly available to readers. The pre-registration is also under embargo to protect sensitive data on the political issues that were prominent at the time of data collection. The experimental materials are explicated in the supplemental materials included here. More information is available upon request.

a. Control condition (authentic political speech from Nancy Pelosi)

"Leadership is about having confidence in your vision, your knowledge, your strategic thinking and how that connects to the aspirations of others. Recognize that you are the unique person you are with your own special power to shape the future, for yourself, and our country. You never know what opportunities the futures hold but we must be ready. I had no intentions of running for office or leadership, but when the opportunity presented itself, I was ready. My strength sprang from my family and the comfort of good friends. So my closing advice is to stay close and to treasure your family and friends."



We took a speech of Pelosi on leadership from YouTube (4:56 – 5:37). This was a fitting video because the video is shot in ultra-high definition, and because Pelosi speaks to the camera rather than to the House of Representatives (similarly to the deepfake conditions).

b. Experimental condition (plausible depolarizing political speech)

"Democrats and Republicans should come closer together. We should not see the world from opposite truths. Take the Capitol riots as example: We tend to be hostile to those who were violent, but we fail to listen to their fears and don't take the effort to understand what drove them. This will not get us any further"

c. Experimental condition (implausible polarizing political speech containing an in-group attack)

"To be honest, I am ashamed of my party. The Democrats always pretend to be the do-gooders, but they are hopelessly unrealistic, naïve, and unwilling to compromise. They always blame the other party, but fail to be reflective and see things through the eyes of those who hold different views."

d. Experimental condition (implausible polarizing political speech expressing out-group sympathy)

"In hindsight, I sympathize with those who went out and occupied the Capitol back in January. People felt the need to fight to get their country back. We need to go back to a country where we make our own rules. I can see how Donald Trump was responding to a feeling held by many US citizens about protecting values and democracy. I can only have respect for that. I see now that I went too far in trying to put our former President on trial with impeachment. I regret that, but I cannot make it undone anymore."

Appendix B – Cognitive Reflection Test items

A cookie and an egg cost $1.10 in total. The cookie costs $1.00 more than the egg. How much does the egg cost?

50 cents
1 dollar
5 cents
10 cents

If it takes 5 machines 5 minutes to make 5 cans, how long would it take 100 machines to make 100 cans?

50 minutes
100 minutes
25 minutes
5 minutes

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

24 days
44 days
47 days
15 days

Appendix C – preregistration plan

Preregistration Template from AsPredicted.org

1. Data collection
Have any data been collected for this study already? Note: 'Yes' is a discouraged answer for this preregistration form.

No, no data have been collected for this study yet.

2. Hypotheses

The main question we explore is whether deepfakes with different relationships to the political and social reality result in different effects on credibility and evaluations of the depicted political actor. To explore this question, we postulate different hypotheses:

H1: Deepfakes that are closer to authentic political statements are rated as more credible than deepfakes that deviate further from the truth.

H2: Deepfakes that deviate further from the truth have stronger effects on de-legitimizing the depicted political actor than deepfakes that stay closer to authentic statements.

H3: Motivated reasoning (accuracy and defensive motivations) moderates the effects of deepfake exposure on credibility and the de-legitimization of the depicted political actor.

We additionally introduce two research questions:

RQ1: How does the credibility rating of deepfakes differ from authentic audiovisual information?

RQ2: To what extent are the effects of deepfakes on credibility and the delegitimization of the depicted political actor similar or different in domestic and foreign contexts?

3. Dependent variables

To assess the potential delegitimizing impact of different degrees of manipulation in deepfakes, we look at two central outcome variables: the credibility and perceived authenticity of the deepfake and its impact on support for the political actor that was delegitimized. We measure credibility in two different ways. First of all, after seeing the video, participants could indicate their first responses to the video without being primed on its trustworthiness or credibility: “In general, what are your thoughts about Nancy Pelosi's video you just saw and the content of the video?” Second, we asked participants to rate various characteristics of the video on a scale from 1 (this does not apply at all) to 7 (this completely applies). The following keywords were used to assess the rating of the video: credible, authentic, informative, accurate, important and objective. This scale of perceived credibility was adopted from earlier research on the effects of disinformation (e.g., Hameleers et al., 2020; Schaewitz et al., 2020). With this measure, we aimed to circumvent the priming of evaluations directly related to the authenticity and credibility of the video. As a second dependent variable, we assessed how exposure to the deepfakes (versus the authentic video) affected the evaluation of the depicted political actor. In the actual survey, this measure was placed before the credibility items to avoid priming effects of credibility

ratings. We specifically asked: “Based on the video you just watched, please indicate how you feel about Nancy Pelosi by indicating to what extent you agree with the following statements.” (1 = completely disagree, 7 = completely agree). The statements included the following evaluations of Pelosi related to different traits: (1) Nancy Pelosi is sympathetic; (2) Nancy Pelosi is competent; (3) Nancy Pelosi is honest; (4) Nancy Pelosi knows what’s going on in our country; (5) Nancy Pelosi sells out her principles; (6) Nancy Pelosi deceives the public; (7) Nancy Pelosi is incapable of representing the people; (8) Nancy Pelosi is a weak politician. We used a balanced mix of positively and negatively phrases traits, and recoded items 1-4 to compute an average scale of negative evaluations of Pelosi. The items come from different sources, and are informed by existing multi-component measures of political candidates’ evaluations that distinguish between competence, leadership, integrity and empathy (e.g., McGraw, 2011).

4. Conditions

How many and which conditions will participants be assigned to?

we use a between-subjects factorial design with four conditions: Participants are randomly exposed to either (1) an authentic political speech; (2) a deepfake that is close to existing viewpoints of the depicted politician without an in-group delegitimization cue (depolarization message without in-group incrimination); (3) a deepfake that offers an attack on the Democrats and delegitimizes fellow partisans (a polarizing message with in-group incrimination); (4) a deepfake with a more radical issue position that legitimizes the opposed political party and political violence combined with a strong in-group party attack (a polarizing message with both in-group incrimination and out-group sympathy).

5. Analyses

To test the research questions and hypotheses, we use ANOVAs and linear regression models. For the ANOVAs, we look at the mean score comparison of credibility ratings and evaluations of the depicted political actor across all conditions. With corrected pairwise mean score comparisons, we explore the significance of the mean difference between the control and the deepfake conditions together, as well as between the control condition and all individual deepfake conditions. We herewith explore whether deepfakes differ significantly from authentic videos, and whether the extremity of the deepfake plays a role in its effects. With the regression models, we dummy-coded the conditions variable into different k-1 binary variables. The control condition (the authentic video) is used as a reference group. The regression models include an assessment of the direct effects of the different deepfakes (Model I), the moderating variables (Model II) and the interaction effect between the deepfake conditions and the moderators (Model III).

6. Outliers and Exclusions

We will first of all do a basic data quality check. Extreme short and extreme long responses will be deleted. The rule of thumb is that completion times under 5 minutes should be removed as the minimal reading time of all questions and stimuli is higher than 5 minutes. In that sense, these response times will be speedsters. We will also check for straight lining behavior and extreme long response times. A response time above 4 hours is considered extremely long as this extends a regular online sessions. Hence, although participants may take a 'break' during the survey, too much distraction and multiple sessions may bias the findings. We will always conduct additional

analyses with the problematic cases included as a robustness checks, so we can assess the influence of these cases on the final results.

Appendix D – details manipulation checks

All manipulations succeeded. Participants exposed to an implausible deepfake that expressed sympathy with out-group partisans were significantly and substantially more likely to identify out-group liking and support (79.0%) compared to participants in the in-group delegitimizing deepfake (7.4%) the plausible deepfake (7.8%) or the control condition (5.8%). The same applied to the depolarizing and plausible deepfake: 65.3% correctly identified the depolarizing narrative, whereas substantially and significantly less participants falsely attributed this to the implausible deepfake that expressed sympathy with out-group partisans (10.2%), the in-group attack deepfake (12.1%) or the control condition (12.4%). We also find comparable results for the deepfake emphasizing negative sentiments toward in-group partisans: 77.4% correctly identified the deceptive statements, whereas they were falsely associated with the implausible deepfake that expressed sympathy with out-group partisans (9.3%), the plausible deepfake (6.9%) or the control condition (6.5%).

Appendix E – Regression Tables

Table E1. The effects of deepfakes on credibility and delegitimization moderated by scores on the cognitive reflection test

	Dependent variable				
	Credibility assessment of the message		Negative evaluation Pelosi (delegitimization)		
	(1)	(2)	(3)	(4)	(5)
Condition: Plausible deepfake[a]	0.18	0.45**	-0.15	-0.41***	-1.49****
	(0.15)	(0.19)	(0.12)	(0.15)	(0.23)
Condition: In-group attack implausible deepfake[a]	-0.47***	-0.19	0.02	-0.15	-0.67***
	(0.15)	(0.19)	(0.12)	(0.15)	(0.26)
Condition: Emphasizing out-group sympathy implausible deepfake[a]	-1.07****	-0.56***	0.46****	0.10	-1.84****
	(0.15)	(0.19)	(0.12)	(0.16)	(0.23)
Cognitive reflection		0.16		-0.27****	-0.16**
		(0.10)		(0.08)	(0.07)
Condition: Plausible deepfake * cognitive reflection		-0.33**		0.33***	0.03
		(0.14)		(0.12)	(0.10)
Condition: In-group attack implausible deepfake * cognitive reflection		-0.33**		0.20*	0.12
		(0.14)		(0.12)	(0.10)
Condition: Emphasizing out-group sympathy implausible deepfake * cognitive reflection		-0.59****		0.42****	0.23**
		(0.14)		(0.12)	(0.1)
Credibility assessment					-0.69***
					(0.04)
Condition: Plausible deepfake * Credibility assessment					0.34****
					(0.05)

Condition: In-group attack implausible deepfake * Credibility assessment					0.14*
					(0.06)
Condition: Emphasizing out-group sympathy implausible deepfake * Credibility assessment					0.49**
					(0.6)
Constant	3.86****	3.73****	3.80****	4.02****	6.58****
	(0.10)	(0.13)	(0.08)	(0.11)	(0.17)
R^2	0.07	0.09	0.02	0.04	
Adjusted R^2	0.06	0.08	0.02	0.03	
Residual Std. Error	1.81 (df = 1183)	1.79 (df = 1179)	1.47 (df = 1183)	1.46 (df = 1179)	
F Statistic	28.10**** (df = 3; 1183)	15.80**** (df = 7; 1179)	9.25**** (df = 3; 1183)	6.19**** (df = 7; 1179)	

Note:
[a]Reference group is control condition (authentic video)
*p<0.1; **p<0.05; ***p<0.01

Table E2. The effects of deepfakes on credibility and delegitimization moderated by prior evaluations of the depicted politician

	Dependent variable:					
	Credibility assessment of the message			Negative evaluation Pelosi (delegitimization)		
	(1)	(2)	(3)	(4)	(5)	(6)
Condition: Plausible deepfake[a]	0.18	0.51**	-0.02	-0.15	-0.44**	0.05
	(0.15)	(0.24)	(0.20)	(0.12)	(0.18)	(0.15)
Condition: In-group attack implausible deepfake[a]	-0.47***	0.06	-1.08****	0.02	-0.23	0.51****
	(0.15)	(0.25)	(0.20)	(0.12)	(0.18)	(0.15)
Condition: Emphasizing out-group sympathy implausible deepfake[a]	-1.07****	0.09	-1.96****	0.46****	-0.12	0.97****
	(0.15)	(0.24)	(0.20)	(0.12)	(0.18)	(0.15)
Support for Pelosi		0.04****			-0.04****	
		(0.003)			(0.002)	
Condition: Plausible deepfake * support for Pelosi		-0.01*			0.01***	
		(0.005)			(0.003)	
Condition: In-group attack implausible deepfake * support for Pelosi		-0.01***			0.01**	
		(0.005)			(0.003)	
Condition: Emphasizing out-group sympathy implausible deepfake * support for Pelosi		-0.03****			0.01****	
		(0.004)			(0.003)	
Support Trump			-0.01****			0.02****
			(0.003)			(0.002)
Condition: Plausible deepfake * support for Trump			0.004			-0.002
			(0.004)			(0.003)
Condition: In-group attack implausible deepfake * support for Trump			0.02****			-0.01****
			(0.004)			(0.003)
Condition: Emphasizing out-group sympathy implausible deepfake * support for Trump			0.03****			-0.01****
			(0.004)			(0.003)
Constant	3.86****	2.20****	4.37****	3.80****	5.49****	2.92****
	(0.10)	(0.17)	(0.14)	(0.08)	(0.12)	(0.11)

R^2	0.07	0.28	0.11	0.02	0.44	0.22
Adjusted R^2	0.06	0.27	0.10	0.02	0.43	0.21
Residual Std. Error	1.81 (df = 1183)	1.62 (df = 1005)	1.77 (df = 1160)	1.47 (df = 1183)	1.16 (df = 1005)	1.32 (df = 1160)
F Statistic	28.10^{****} (df = 3; 1183)	55.40^{****} (df = 7; 1005)	20.37^{****} (df = 7; 1160)	9.25^{****} (df = 3; 1183)	111.01^{****} (df = 7; 1005)	46.41^{****} (df = 7; 1160)

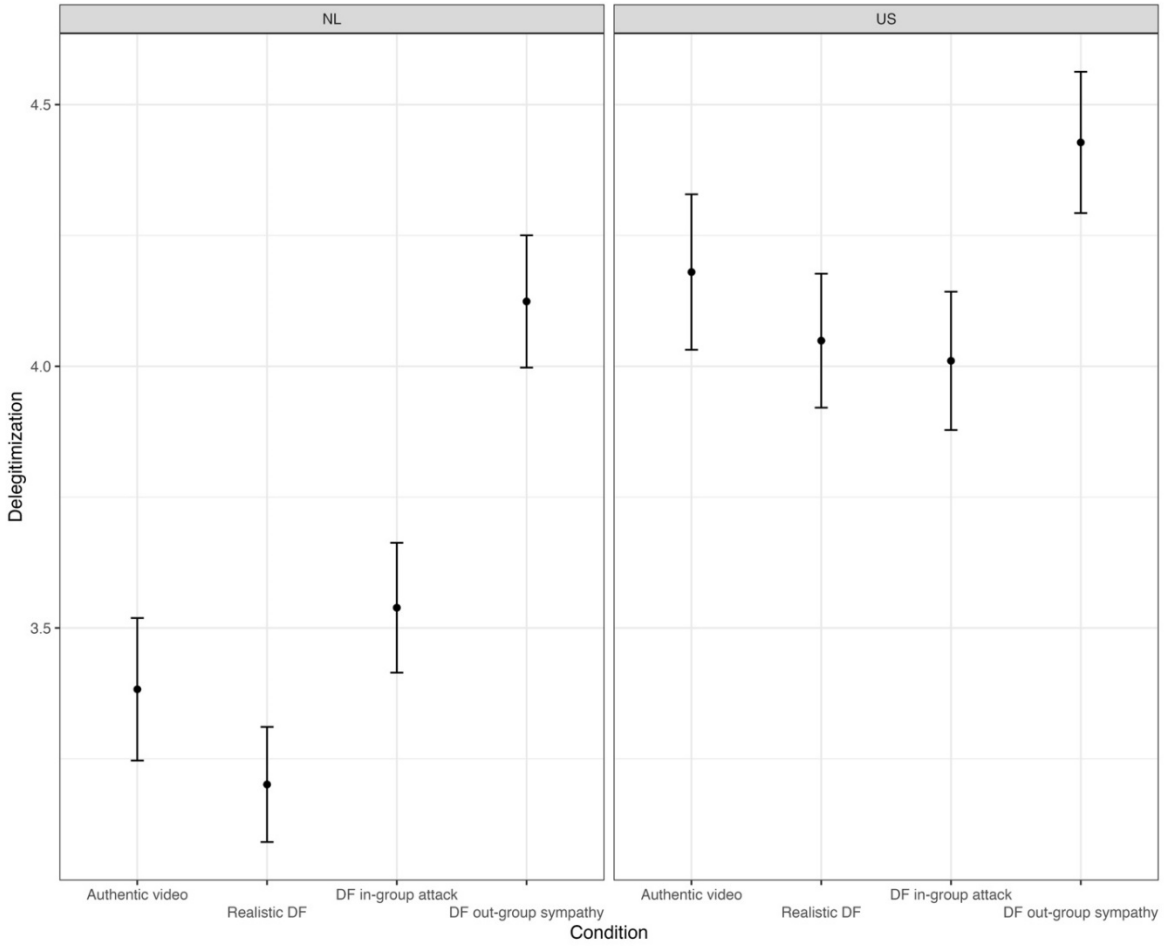
Note:
[a]Reference group is control condition (authentic video)
*p<0.1; **p<0.05; ***p<0.01

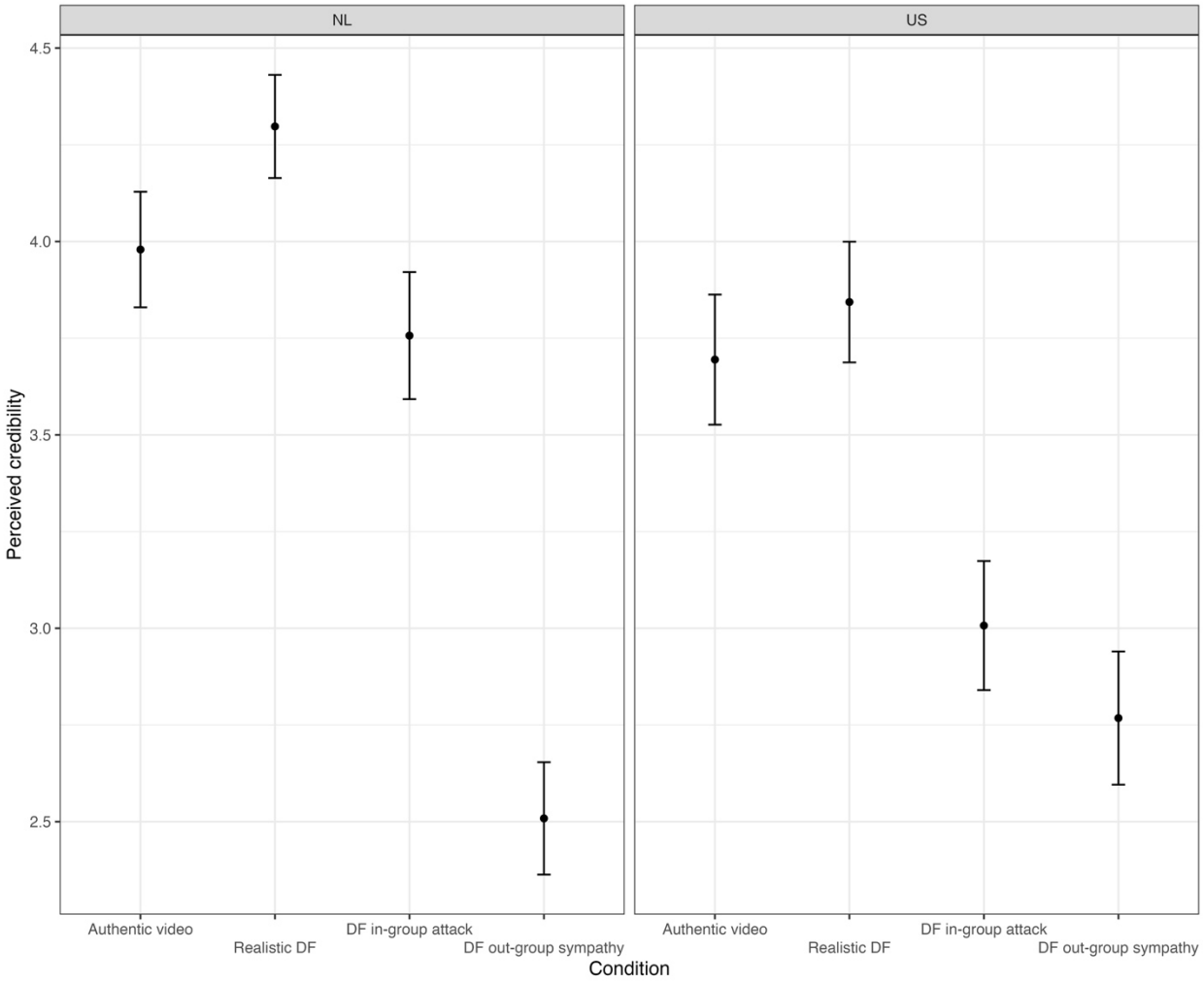
Table F1. The effects of deepfakes on credibility and delegitimization including all moderators

	Dependent variable:	
	Credibility assessment of the message	Negative evaluation Pelosi (delegitimization)
	(1)	(2)
Condition: Plausible deepfake[a]	-0.47**	0.64**
	(0.19)	(0.27)
Condition: In-group attack implausible deepfake[a]	-0.22	0.20
	(0.20)	(0.27)
Condition: Emphasizing out-group sympathy implausible deepfake[a]	-0.26	0.57**
	(0.20)	(0.27)
Cognitive reflection	-0.08	0.03
	(0.07)	(0.10)
Support for Pelosi	-0.04****	0.04****
	(0.002)	(0.003)
Country VS versus NL	0.21***	-0.06
	(0.08)	(0.11)
Condition: Plausible deepfake * cognitive reflection	0.06	-0.14
	(0.10)	(0.14)
Condition: In-group attack implausible deepfake * cognitive reflection	0.01	-0.17
	(0.10)	(0.14)
Condition: Emphasizing out-group sympathy implausible deepfake * cognitive reflection	0.18*	-0.48****
	(0.10)	(0.14)
Condition: Plausible deepfake * support for Pelosi	0.01**	-0.01*
	(0.003)	(0.005)
Condition: In-group attack implausible deepfake * support for Pelosi	0.01**	-0.01***
	(0.003)	(0.005)
Condition: Emphasizing out-group sympathy implausible deepfake * support for Pelosi		
Condition: Plausible deepfake * support for Pelosi	0.01****	-0.03****
	(0.003)	(0.004)
Constant	5.38****	2.23****
	(0.14)	(0.20)
R^2	0.44	0.29
Adjusted R^2	0.44	0.29
Residual Std. Error (df = 1000)	1.16	1.60
F Statistic (df = 12; 1000)	66.19****	34.77****

Note: [a]Reference group is control condition
*p<0.1; **p<0.05; ***p<0.01

Appendix G Mean score comparison for credibility and delegitimization (negative evaluations of depicted politician) per condition, per country, with only those respondents who indicated to be familiar with Pelosi




NL
US
4.5
4.0
3.5
3.0
2.5
Perceived credibility
Authentic video
Realistic DF
DF in-group attack
DF out-group sympathy
Condition