

Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media

Soubhik Barari¹ Christopher Lucas² Kevin Munger³

¹Harvard University

²Washington University in St. Louis

³Pennsylvania State University

soubhikbarari.github.io/files/deepfakes.pdf

UCLA Political Psych Lab

April 23, 2021

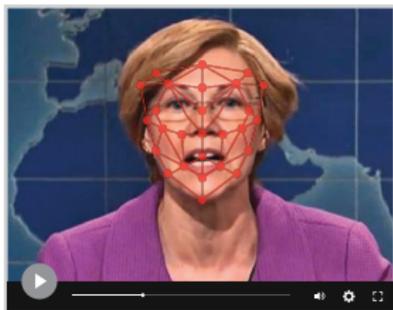
A New Frontier in Political Misinformation?

A New Frontier in Political Misinformation?

- ▶ Huge public/policy concern about political “deepfakes”

A New Frontier in Political Misinformation?

- ▶ Huge public/policy concern about political “deepfakes”
 - ▶ Low barriers (\$ and skill) of entry
 - ▶ Deepfakes supposedly triggered government coups, sex scandals



A New Frontier in Political Misinformation?

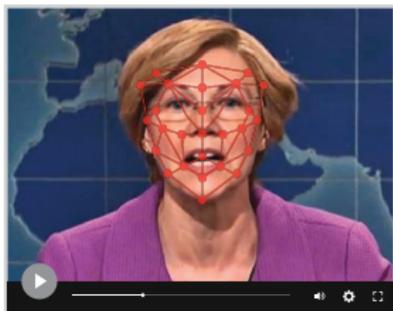
- ▶ Huge public/policy concern about political “deepfakes”
 - ▶ Low barriers (\$ and skill) of entry
 - ▶ Deepfakes supposedly triggered government coups, sex scandals



- ▶ Debate: video often assumed to be superior format of political communication (persuasion, affective appeal)

A New Frontier in Political Misinformation?

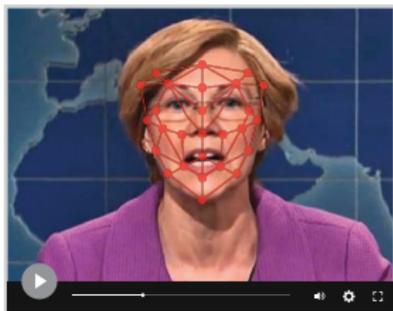
- ▶ Huge public/policy concern about political “deepfakes”
 - ▶ Low barriers (\$ and skill) of entry
 - ▶ Deepfakes supposedly triggered government coups, sex scandals



- ▶ Debate: video often assumed to be superior format of political communication (persuasion, affective appeal) \rightsquigarrow but, many recent studies document minimal persuasive effects (ads, news)

A New Frontier in Political Misinformation?

- ▶ Huge public/policy concern about political “deepfakes”
 - ▶ Low barriers (\$ and skill) of entry
 - ▶ Deepfakes supposedly triggered government coups, sex scandals



- ▶ Debate: video often assumed to be superior format of political communication (persuasion, affective appeal) \rightsquigarrow but, many recent studies document minimal persuasive effects (ads, news)
- ▶ So are these concerns warranted?

First-Order Questions

First-Order Questions

RQ1 Are deepfakes of political elites more **credible** or **affectively appealing** relative to equivalent information in extant media (text, audio)?

First-Order Questions

RQ1 Are deepfakes of political elites more **credible** or **affectively appealing** relative to equivalent information in extant media (text, audio)?

RQ2 Are these credibility perceptions or appeals **heterogeneous** across subgroups?

First-Order Questions

- RQ1** Are deepfakes of political elites more **credible** or **affectively appealing** relative to equivalent information in extant media (text, audio)?
- RQ2** Are these credibility perceptions or appeals **heterogeneous** across subgroups?
- RQ3** Are deepfakes of political elites **discernible** from authentic videos?

First-Order Questions

- RQ1** Are deepfakes of political elites more **credible** or **affectively appealing** relative to equivalent information in extant media (text, audio)?
- RQ2** Are these credibility perceptions or appeals **heterogeneous** across subgroups?
- RQ3** Are deepfakes of political elites **discernible** from authentic videos?

One survey ($n = 5,750$, U.S.), two experiments (Aug. 2020):

First-Order Questions

- RQ1** Are deepfakes of political elites more **credible** or **affectively appealing** relative to equivalent information in extant media (text, audio)?
- RQ2** Are these credibility perceptions or appeals **heterogeneous** across subgroups?
- RQ3** Are deepfakes of political elites **discernible** from authentic videos?

One survey ($n = 5,750$, U.S.), two experiments (Aug. 2020):

- 1** Incidental exposure: fake scandal planted in news feed \rightsquigarrow randomize medium (leaked video, text headline, audio hot mic)

First-Order Questions

- RQ1** Are deepfakes of political elites more **credible** or **affectively appealing** relative to equivalent information in extant media (text, audio)?
- RQ2** Are these credibility perceptions or appeals **heterogeneous** across subgroups?
- RQ3** Are deepfakes of political elites **discernible** from authentic videos?

One survey ($n = 5,750$, U.S.), two experiments (Aug. 2020):

- 1** Incidental exposure: fake scandal planted in news feed \rightsquigarrow randomize medium (leaked video, text headline, audio hot mic)
- 2** Detection task: discern deepfakes from authentic clips \rightsquigarrow randomize number of deepfakes in task environment

Who is Susceptible? (RQ2)

	Subgroup	Mechanisms of Credibility (though we don't test these)
Non-Intervenable in Survey	Partisans (w/out-partisan targets)	<ul style="list-style-type: none">• Directional motivated reasoning• Accuracy motivated reasoning
	Sexists (w/female targets)	<ul style="list-style-type: none">• Consistency w/prior hostile beliefs• Consistency w/prior benevolent beliefs
	Older adults	Inability to evaluate accuracy of digital info
	Low cognitive reflection	Overreliance on intuition in judgment
	Low political knowledge	<ul style="list-style-type: none">• Inability to evaluate plausibility of political events• Inability to recognize real facial features of target
	Low digital literacy	<ul style="list-style-type: none">• Inability to evaluate accuracy of digital info• Limited recognition of deepfake technology
Intervenable	Low accuracy salience	Limited attn. to factual accuracy of media
	Uninformed about deepfakes	Limited recognition of deepfake technology

Who is Susceptible? (RQ2)

	Subgroup	Mechanisms of Credibility (though we don't test these)
Non-Intervenable in Survey	Partisans (w/out-partisan targets)	<ul style="list-style-type: none">• Directional motivated reasoning• Accuracy motivated reasoning
	Sexists (w/female targets)	<ul style="list-style-type: none">• Consistency w/prior hostile beliefs• Consistency w/prior benevolent beliefs
	Older adults	Inability to evaluate accuracy of digital info
	Low cognitive reflection	Overreliance on intuition in judgment
	Low political knowledge	<ul style="list-style-type: none">• Inability to evaluate plausibility of political events• Inability to recognize real facial features of target
	Low digital literacy	<ul style="list-style-type: none">• Inability to evaluate accuracy of digital info• Limited recognition of deepfake technology
Intervenable	Low accuracy salience	Limited attn. to factual accuracy of media
	Uninformed about deepfakes	Limited recognition of deepfake technology

If popular concerns true, these “at-risk” subgroups might find deepfakes *more* credible than audio, text, etc.

Overview of Experiments Embedded in Survey

Overview of Experiments Embedded in Survey

Exposure(s)	Pre-Exposure Interventions	Outcomes
-------------	----------------------------	----------

Overview of Experiments Embedded in Survey

	Exposure(s)	Pre-Exposure Interventions	Outcomes
① Incidental Exposure	<ol style="list-style-type: none">1. Authentic coverage of 2020 D candidates2. Randomized to text, audio, video, skit clip of E. Warren scandal, attack ad, or control (no stimuli)3. Authentic coverage of 2020 D candidates	<ul style="list-style-type: none">• Info about deepfakes	<ul style="list-style-type: none">• Credibility of clips• Affect towards candidates

Overview of Experiments Embedded in Survey

	Exposure(s)	Pre-Exposure Interventions	Outcomes
① Incidental Exposure	<ol style="list-style-type: none">1. Authentic coverage of 2020 D candidates2. Randomized to text, audio, video, skit clip of E. Warren scandal, attack ad, or control (no stimuli)3. Authentic coverage of 2020 D candidates	<ul style="list-style-type: none">• Info about deepfakes	<ul style="list-style-type: none">• Credibility of clips• Affect towards candidates
② Detection Task	<p>Random video feed:</p> <ul style="list-style-type: none">• No-fake: 8 authentic• Low-fake: 6 authentic, 2 deepfakes• High-fake: 2 authentic, 6 deepfakes	<ul style="list-style-type: none">• Debrief deepfakes in ①• Acc prime	<ul style="list-style-type: none">• Acc• FPR• FNR

Pre-Exposure Demographic Questionnaire

- Gender
- Ambivalent sexism
- PID
- Political knowledge

Pre-Exposure Demographic Questionnaire

- Gender
- Ambivalent sexism
- PID
- Political knowledge

No Information

Information About Deepfakes

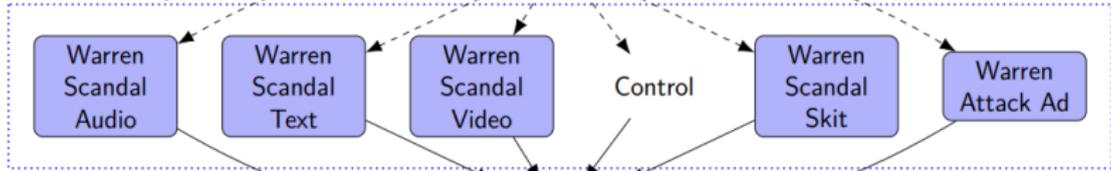
Pre-Exposure Demographic Questionnaire

- Gender
- Ambivalent sexism
- PID
- Political knowledge

No Information Information About Deepfakes

Newsfeed

Incidental Exposure



Post-Exposure Questionnaire

- To what extent (1-5) do you think clipping of [event in each clip] was:
funny / offensive / fake or doctored / informative
- Rate how warmly you feel (1-100) towards each candidate:
Biden / Klobuchar / Warren / Sanders
- Digital literacy

Example video exposure:



The screenshot shows a YouTube video player. At the top left, there is a profile picture of a woman and the name "Michelle Obama" with a verified badge. Below the name is the text "Michelle Obama" and "March 28, 2019 · 1:08 · 54 ·". To the right of the name is a three-dot menu icon. The video frame shows a woman with short blonde hair and glasses, wearing a blue blazer, holding a white smartphone and speaking. A subtitle at the bottom of the video frame reads "because he's a sexist piece of shit". Below the video frame, the video progress bar shows "0:05 / 0:08". Below the progress bar, the text "YOUTUBE.COM" is followed by the video title: "Leak: Elizabeth Warren calls Donald Trump 'a piece of sh**' and a pedophile in 2019 campaign call". Below the title are icons for YouTube, Facebook, and Twitter, and a "Comments" button.

Michelle Obama
March 28, 2019 · 1:08 · 54 ·

0:05 / 0:08 because he's a sexist piece of shit

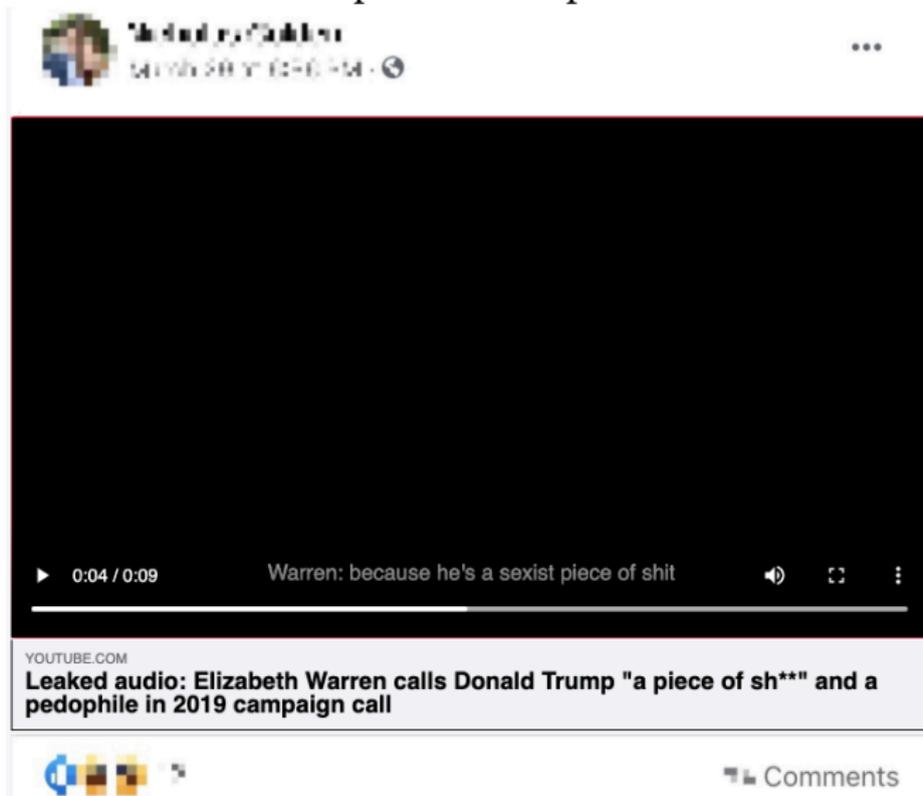
YOUTUBE.COM
Leak: Elizabeth Warren calls Donald Trump "a piece of sh" and a pedophile in 2019 campaign call**

Comments

Watch Video



Example audio exposure:



The screenshot shows a YouTube video player interface. At the top left is a profile picture of a woman and the channel name "The Daily Caller". Below the video player, the text "Warren: because he's a sexist piece of shit" is visible. The video player controls show a progress bar at 0:04 / 0:09. Below the video player, the text "YOUTUBE.COM" is followed by the title "Leaked audio: Elizabeth Warren calls Donald Trump 'a piece of sh***' and a pedophile in 2019 campaign call". At the bottom right, there is a "Comments" button.

Warren: because he's a sexist piece of shit

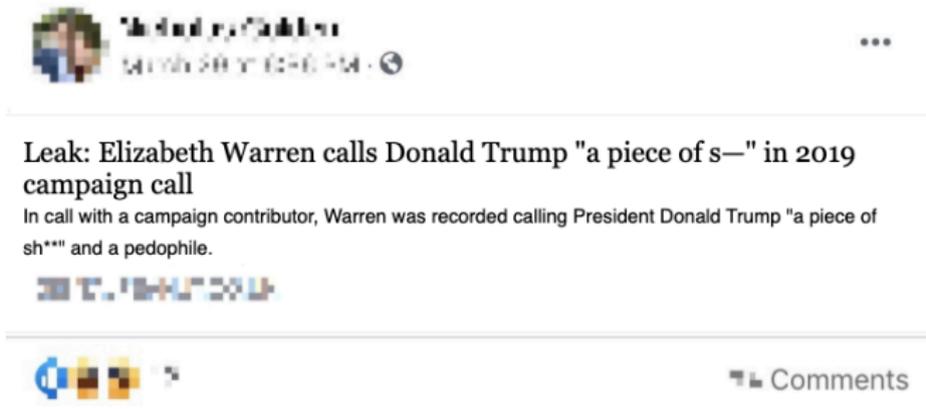
0:04 / 0:09

YOUTUBE.COM

Leaked audio: Elizabeth Warren calls Donald Trump "a piece of sh*" and a pedophile in 2019 campaign call**

Comments

Example text exposure:



The screenshot shows a social media post from 'The Daily Caller'. The profile picture is a circular icon with a person's face. The name 'The Daily Caller' is displayed in bold, with 'March 28, 2019 12:46 PM' and a globe icon below it. A three-dot menu icon is in the top right. The main text reads: 'Leak: Elizabeth Warren calls Donald Trump "a piece of s—" in 2019 campaign call'. Below this is a paragraph: 'In call with a campaign contributor, Warren was recorded calling President Donald Trump "a piece of sh***" and a pedophile.' There is a redacted area with a grey background and a white border. At the bottom, there are icons for share, like, and comment, and the text 'Comments' with a comment icon.

The Daily Caller
March 28, 2019 12:46 PM

Leak: Elizabeth Warren calls Donald Trump "a piece of s—" in 2019 campaign call

In call with a campaign contributor, Warren was recorded calling President Donald Trump "a piece of sh***" and a pedophile.

Comments

Reference affective exposure (skit):



Reference affective exposure (ad):



Sen. Liz Warren is pushing legislation to let the Mashpee Wampanoag Tribe get into the casino business with a \$1 billion resort...

But Senator Elizabeth Warren is now pushing

0:04 / 0:30

YOUTUBE.COM

Tell Senator Warren: No Faux Casino, Pocahontas! | Ad

Post-Exposure Questionnaire

- To what extent (1-5) do you think clipping of [event in each clip] was: funny / offensive / fake or doctored / informative
- Rate how warmly you feel (1-100) towards each candidate: Biden / Klobuchar / Warren / Sanders
- Digital literacy

Post-Exposure Questionnaire

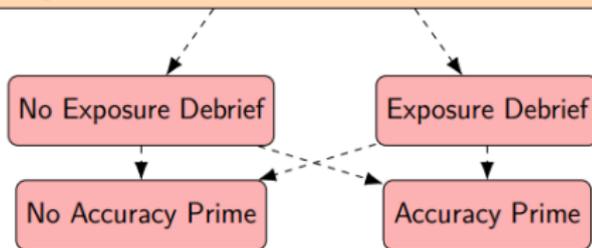
- To what extent (1-5) do you think clipping of [event in each clip] was: funny / offensive / fake or doctored / informative
- Rate how warmly you feel (1-100) towards each candidate: Biden / Klobuchar / Warren / Sanders
- Digital literacy

No Exposure Debrief

Exposure Debrief

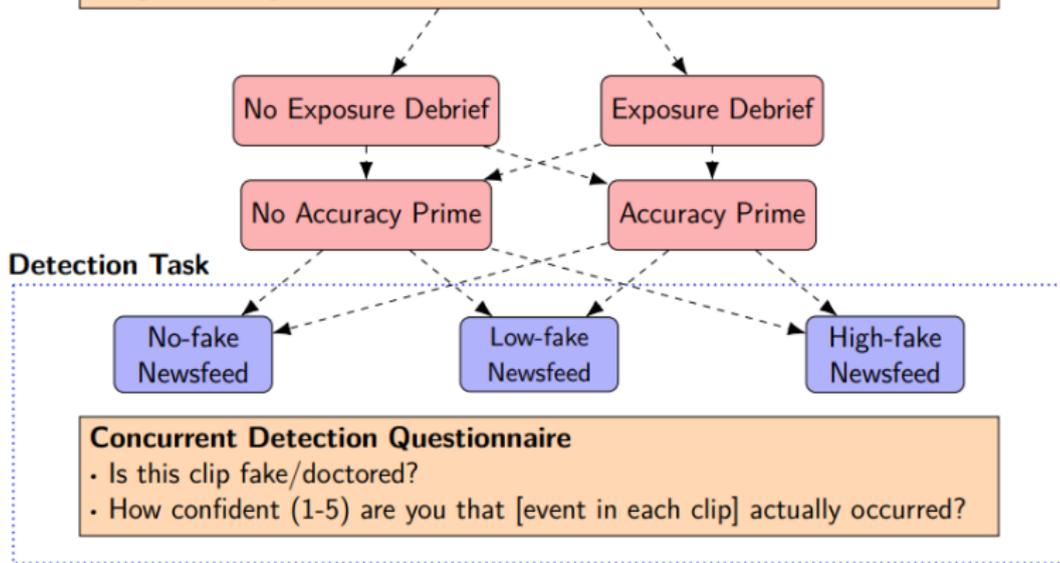
Post-Exposure Questionnaire

- To what extent (1-5) do you think clipping of [event in each clip] was: funny / offensive / fake or doctored / informative
- Rate how warmly you feel (1-100) towards each candidate: Biden / Klobuchar / Warren / Sanders
- Digital literacy



Post-Exposure Questionnaire

- To what extent (1-5) do you think clipping of [event in each clip] was: funny / offensive / fake or doctored / informative
- Rate how warmly you feel (1-100) towards each candidate: Biden / Klobuchar / Warren / Sanders
- Digital literacy



Example detection clips:



(a) Is this clipping fake/doctored?

(b) Is this clipping fake/doctored?

Many Trade-Offs Considered in Our Design

- ▶ Why Warren in the incidental exposure?

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
 - ~> prime target for deepfake video: controversies, salience, gender, supply of impersonators

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
~> prime target for deepfake video: controversies, salience, gender, supply of impersonators
- ▶ **Why those clips in detection task?**

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
 - ↪ prime target for deepfake video: controversies, salience, gender, supply of impersonators
- ▶ **Why those clips in detection task?**
 - ↪ highest quality deepfakes we could find matched to real clips of same elites, hard to know exact populations

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
~> prime target for deepfake video: controversies, salience, gender, supply of impersonators
- ▶ **Why those clips in detection task?**
~> highest quality deepfakes we could find matched to real clips of same elites, hard to know exact populations
- ▶ **Why credibility (“is this real?”) and not deception (“did this happen”?)**

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
~> prime target for deepfake video: controversies, salience, gender, supply of impersonators
- ▶ **Why those clips in detection task?**
~> highest quality deepfakes we could find matched to real clips of same elites, hard to know exact populations
- ▶ **Why credibility (“is this real?”) and not deception (“did this happen”?)**
~> responses theoretically could be different, some evidence they’re not in practice (Appendix G32-G33), useful future research

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
↪ prime target for deepfake video: controversies, salience, gender, supply of impersonators
- ▶ **Why those clips in detection task?**
↪ highest quality deepfakes we could find matched to real clips of same elites, hard to know exact populations
- ▶ **Why credibility (“is this real?”) and not deception (“did this happen?”)**
↪ responses theoretically could be different, some evidence they’re not in practice (Appendix G32-G33), useful future research
- ▶ **Are your 2019 deepfakes representative of \geq 2021 deepfakes?**

Many Trade-Offs Considered in Our Design

- ▶ **Why Warren in the incidental exposure?**
~> prime target for deepfake video: controversies, salience, gender, supply of impersonators
- ▶ **Why those clips in detection task?**
~> highest quality deepfakes we could find matched to real clips of same elites, hard to know exact populations
- ▶ **Why credibility (“is this real?”) and not deception (“did this happen?”)**
~> responses theoretically could be different, some evidence they’re not in practice (Appendix G32-G33), useful future research
- ▶ **Are your 2019 deepfakes representative of \geq 2021 deepfakes?**
~> no, *but*, if deepfakes are now indistinguishable from real videos, our findings hint that’s still a problem

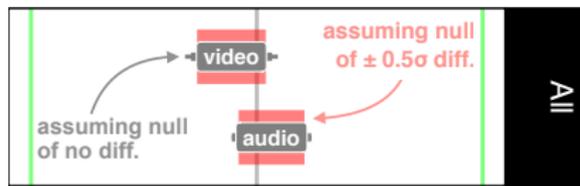
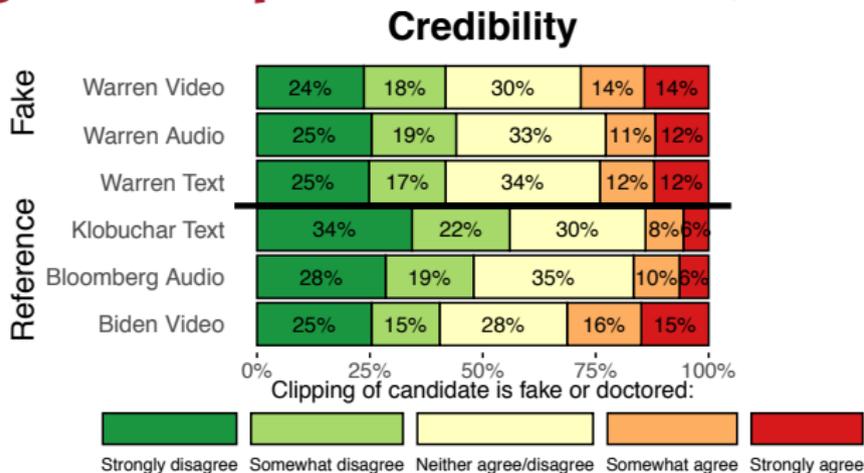
Findings

RQ1/2:

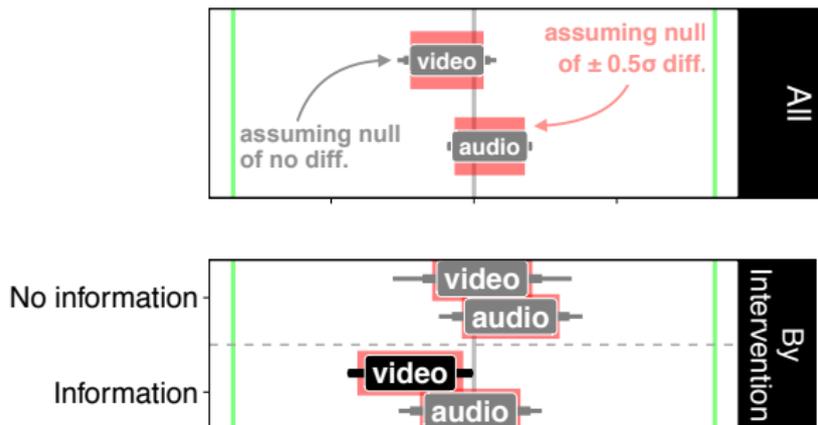
RQ1/2: No

RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)

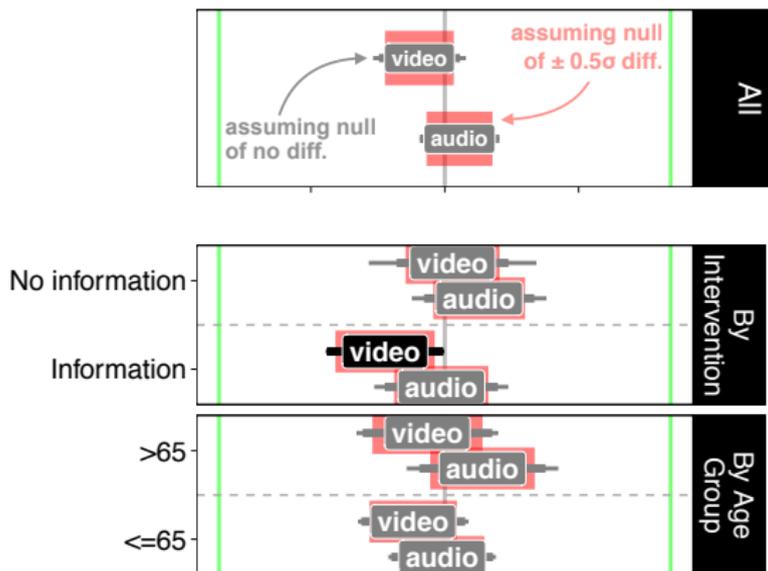
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



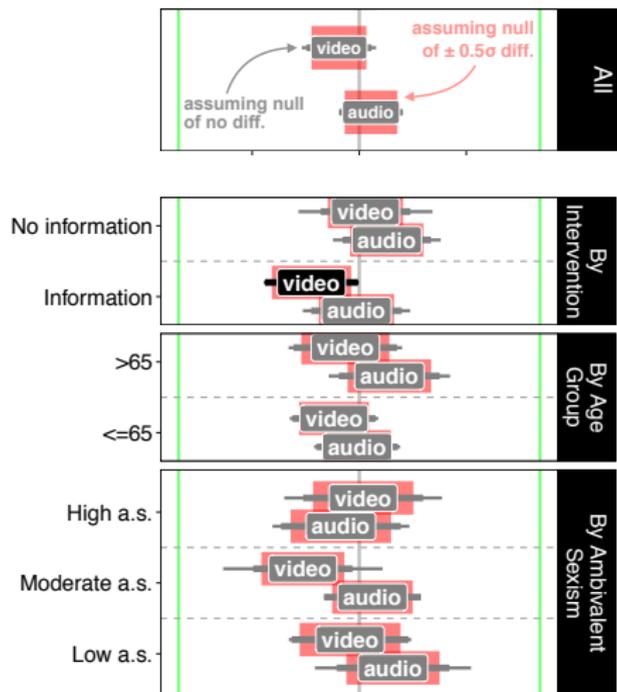
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



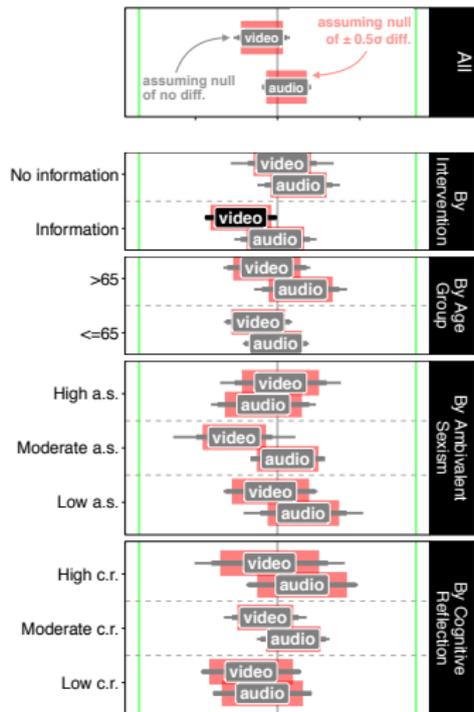
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



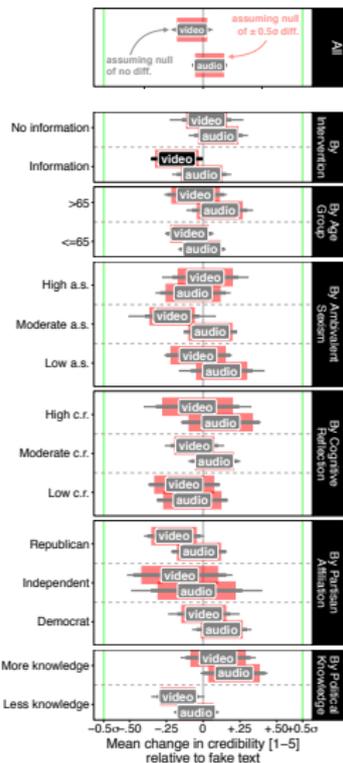
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



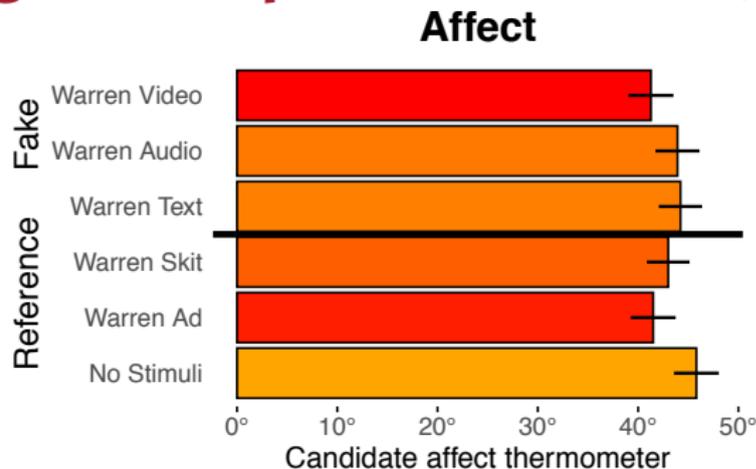
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



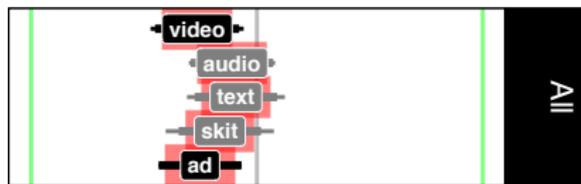
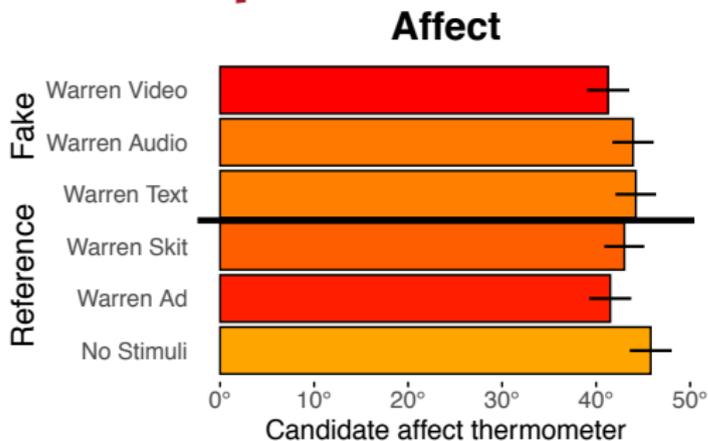
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



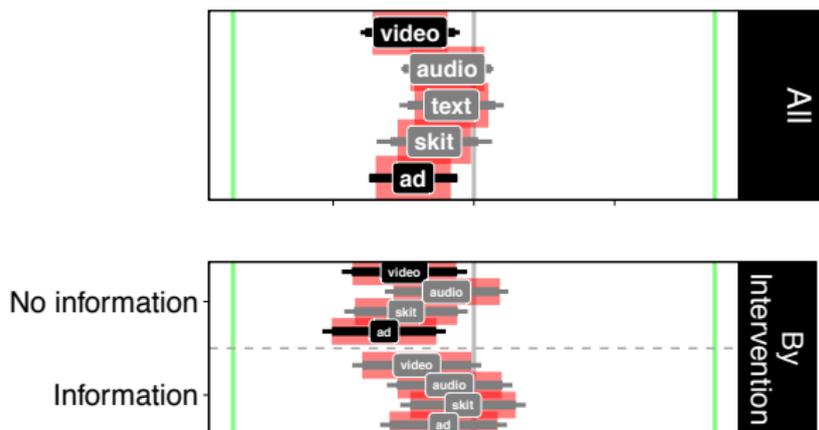
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



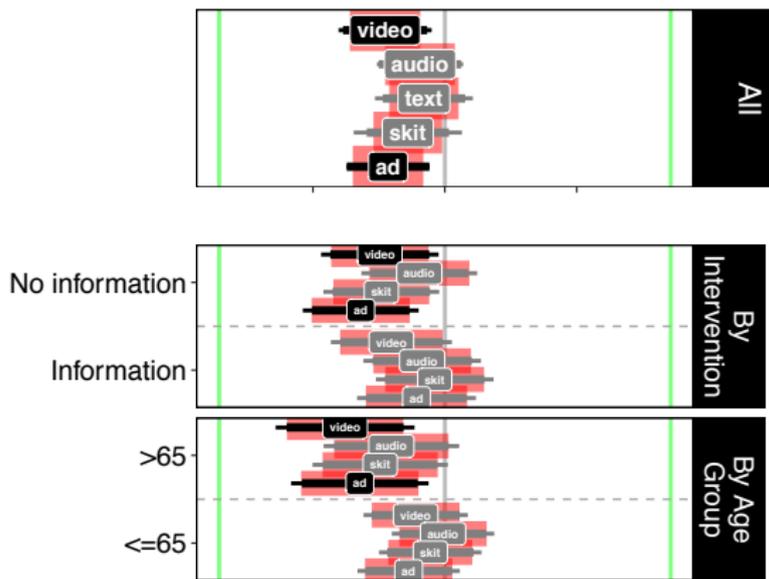
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



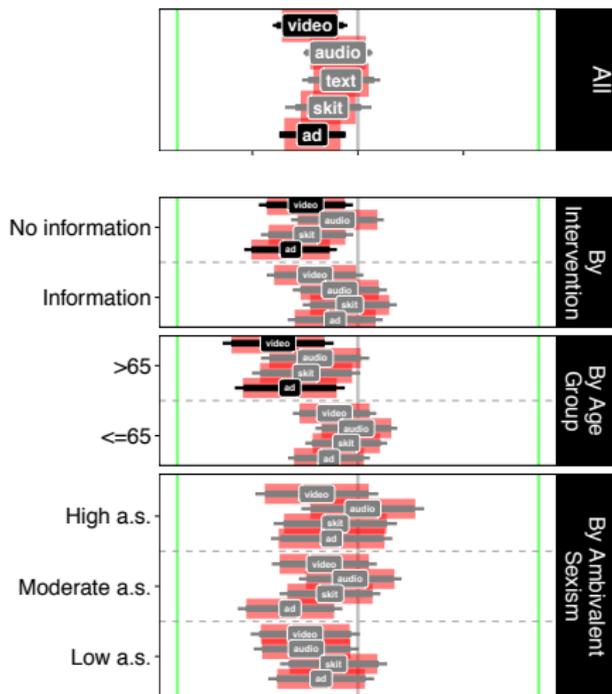
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



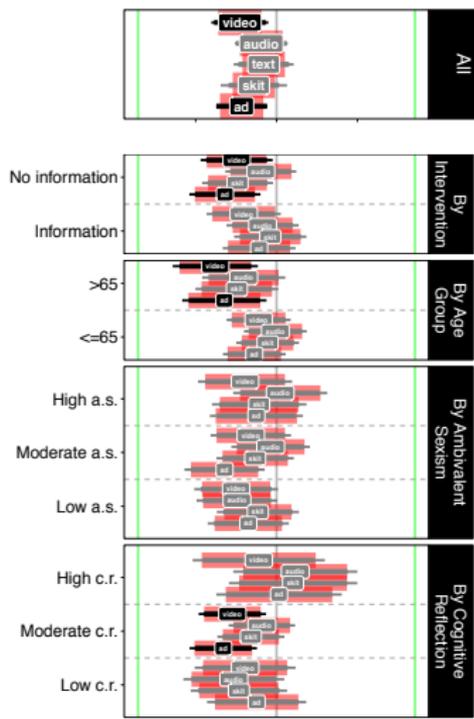
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



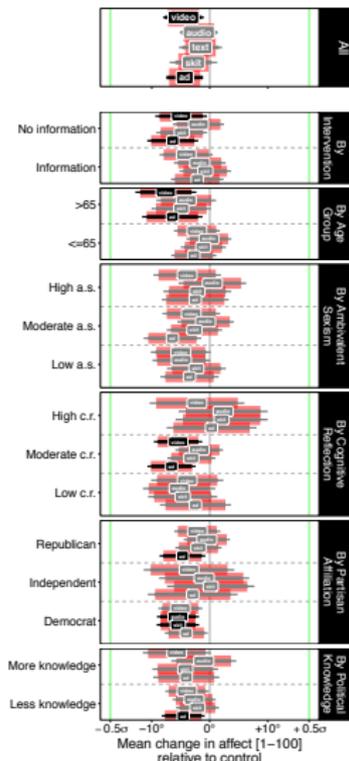
RQ1/2: No (deepfakes no more credible or affectively appealing than comparable fake media)



RQ1/2: No (deepfakes no more credible or affectively appealing than comparable fake media)



RQ1/2: No (deepfakes no more credible or affectively appealing than comparable fake media)

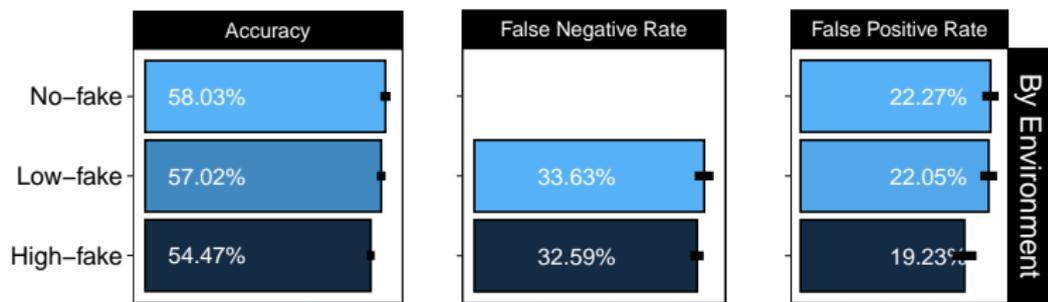


RQ3:

RQ3: Sorta

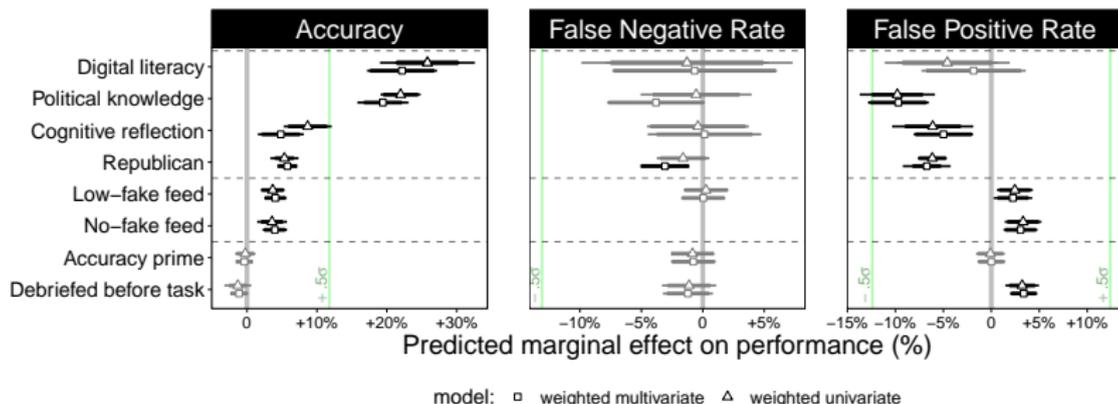
RQ3: Sorta (FNR higher than FPR)

RQ3: Sorta (FNR higher than FPR)



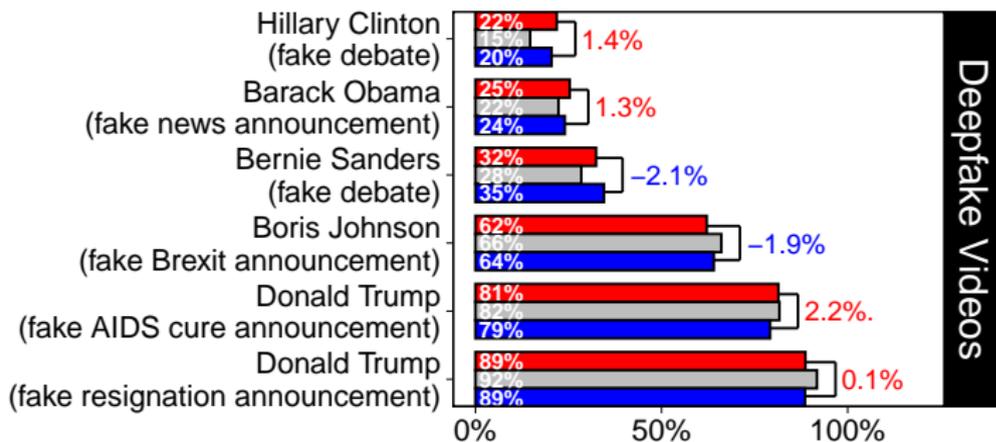
RQ3: Sorta (but, digital literacy and pol. knowledge improve FPR)

RQ3: Sorta (but, digital literacy and pol. knowledge improve FPR)

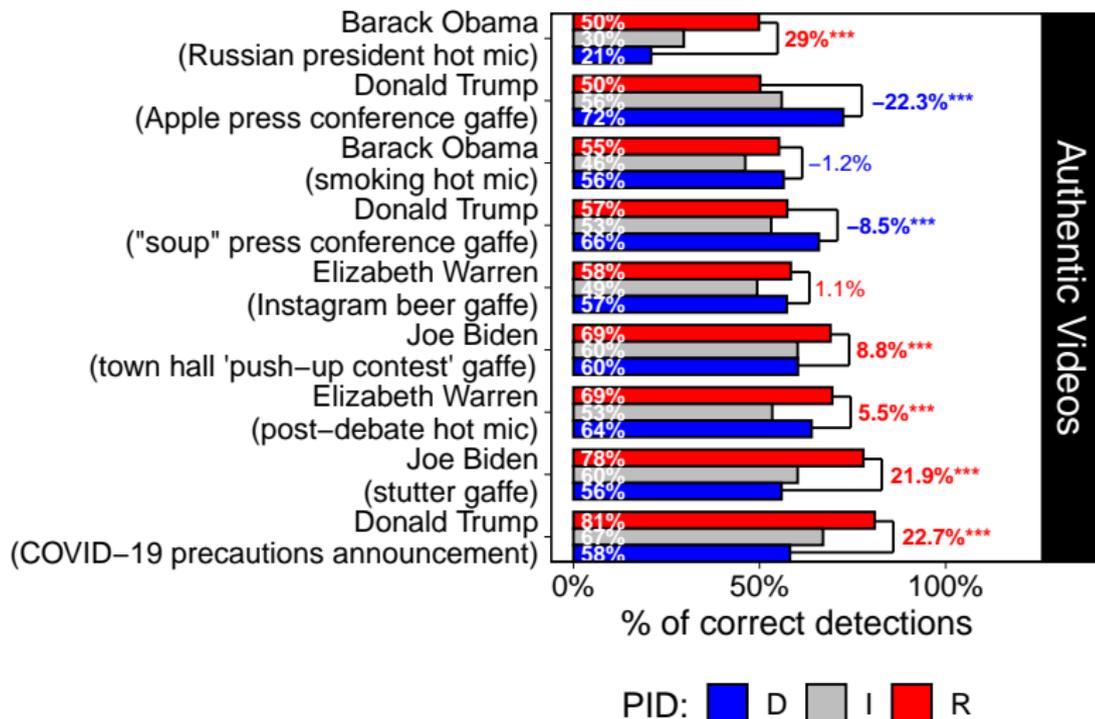


RQ3: Sorta (however, *significant* gap in FPR between Democrats and Republicans)

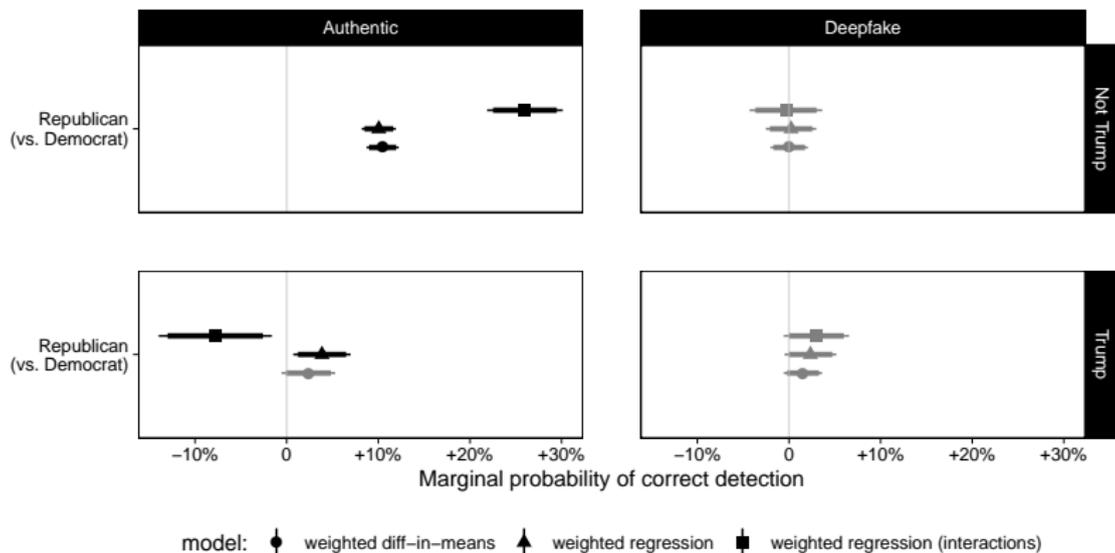
RQ3: Sorta (however, *significant* gap in FPR between Democrats and Republicans)



RQ3: Sorta (however, significant gap in FPR between Democrats and Republicans)



RQ3: Sorta (however, significant gap in FPR between Democrats and Republicans)



model: ◆ weighted diff-in-means ▲ weighted regression ■ weighted regression (interactions)

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020),

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)
- ▶ False positives in a “deepfake world” more concerning (Ternovski, Kalla, Aronow 2021),

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)
- ▶ False positives in a “deepfake world” more concerning (Ternovski, Kalla, Aronow 2021), but **digital + political literacy** help

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)
- ▶ False positives in a “deepfake world” more concerning (Ternovski, Kalla, Aronow 2021), but **digital + political literacy** help
- ▶ As deepfake technology approaches limits of realism,

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)
- ▶ False positives in a “deepfake world” more concerning (Ternovski, Kalla, Aronow 2021), but **digital + political literacy** help
- ▶ As deepfake technology approaches limits of realism, findings suggest partisanship may influence credibility assessments more

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)
- ▶ False positives in a “deepfake world” more concerning (Ternovski, Kalla, Aronow 2021), but **digital + political literacy** help
- ▶ As deepfake technology approaches limits of realism, findings suggest partisanship may influence credibility assessments more ~→ why?

Takeaways

- ▶ Contribute to a growing consensus that video communication not only has minimal effects (Coppock, Hill, Vavreck 2020), but also minimal *differential* effects (Wittenberg, Berinsky, Zong, Rand n.d.)
- ▶ False positives in a “deepfake world” more concerning (Ternovski, Kalla, Aronow 2021), but **digital + political literacy** help
- ▶ As deepfake technology approaches limits of realism, findings suggest partisanship may influence credibility assessments more ~→ why?
- ▶ Partisan cheerleading? Motivated reasoning? All mechanisms to explore in future work.

“If everybody lies to you, the consequence is not that you believe the lies, but rather that nobody believes anything any longer” – Hannah Arendt