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

Soubhik Barari[1] **Christopher Lucas**[2] **Kevin Munger**[3]

[1]Harvard University

[2]Washington University in St. Louis

[3]Pennsylvania State University

soubhikbarari.github.io/files/deepfakes.pdf

UCLA Political Psych Lab
April 23, 2021

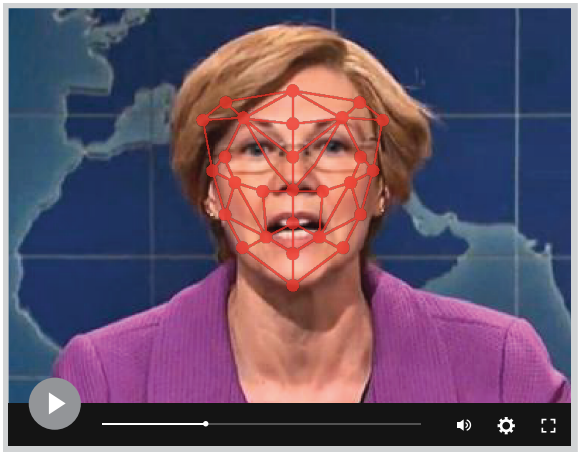
A New Frontier in Political Misinformation?

A New Frontier in Political Misinformation?

- Huge public/policy concern about political "deepfakes"

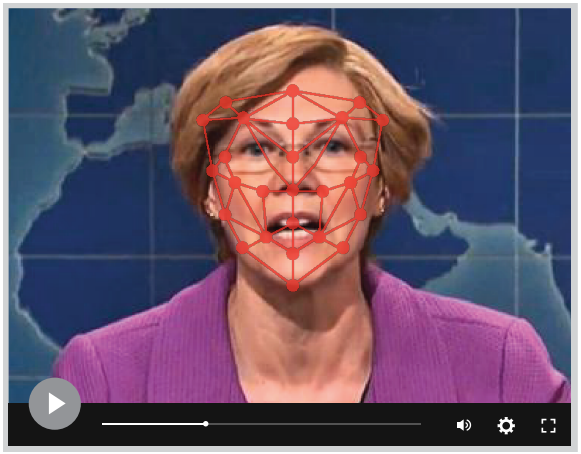
A New Frontier in Political Misinformation?

- 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?

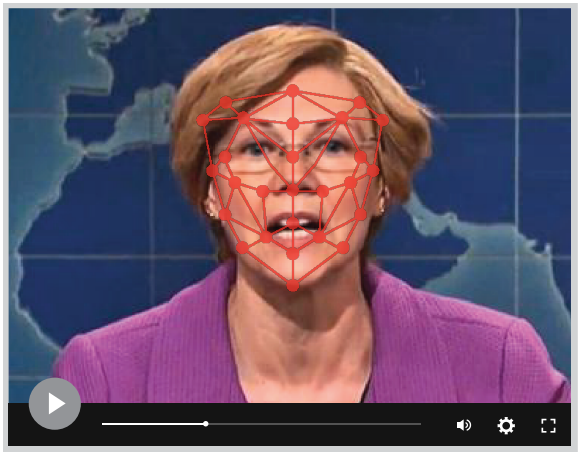
- Huge public/policy concern about political "deepfakes"
  - Low barriers ($ and skill) of entry
  - Deepfakes supposedly triggered government coups, sex scandals



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

A New Frontier in Political Misinformation?

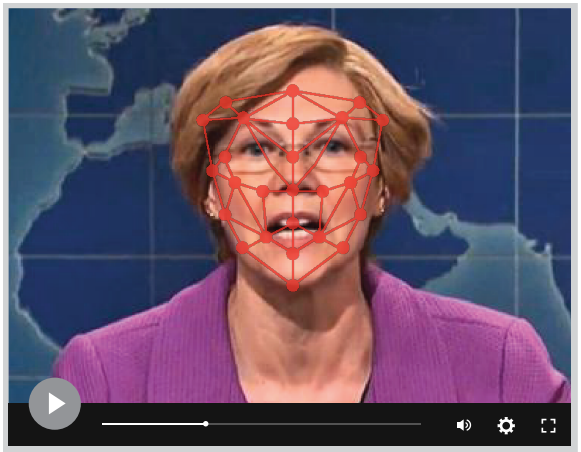
- Huge public/policy concern about political "deepfakes"
  - Low barriers ($ and skill) of entry
  - Deepfakes supposedly triggered government coups, sex scandals



- 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)	• Directional motivated reasoning • Accuracy motivated reasoning
	Sexists (w/female targets)	• 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	• Inability to evaluate plausibility of political events • Inability to recognize real facial features of target
	Low digital literacy	• 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)	• Directional motivated reasoning • Accuracy motivated reasoning
	Sexists (w/female targets)	• 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	• Inability to evaluate plausibility of political events • Inability to recognize real facial features of target
	Low digital literacy	• 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

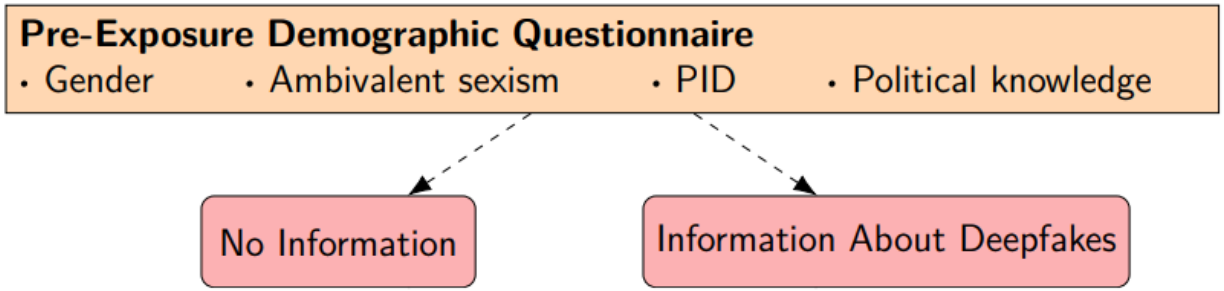
	Exposure(s)	Pre-Exposure Interventions	Outcomes
① Incidental Exposure	1. Authentic coverage of 2020 D candidates 2. **Randomized to `text`, `audio`, `video`, `skit` clip of E. Warren scandal, attack `ad`, or `control` (no stimuli)** 3. Authentic coverage of 2020 D candidates	• Info about deepfakes	• Credibility of clips • Affect towards candidates

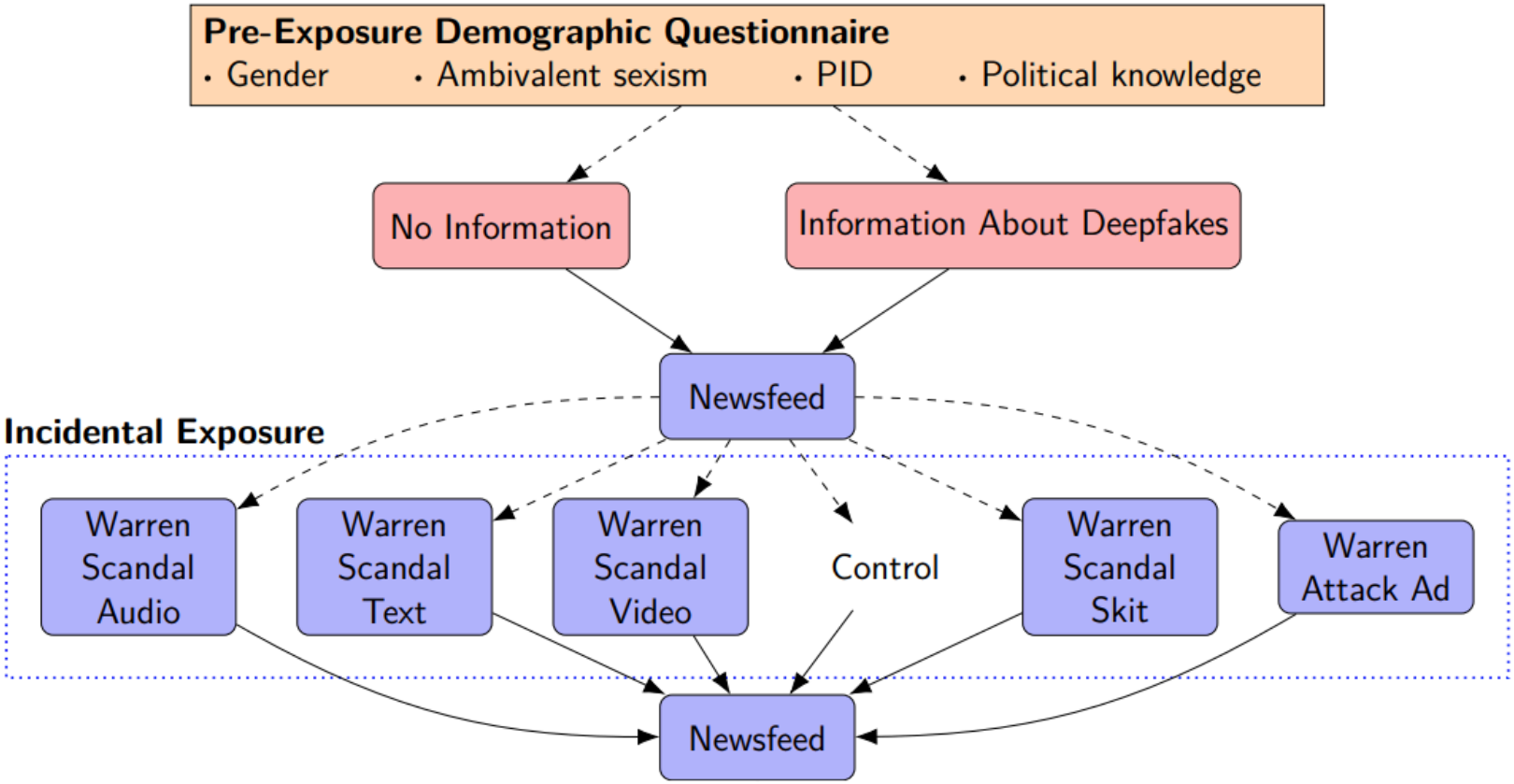
Overview of Experiments Embedded in Survey

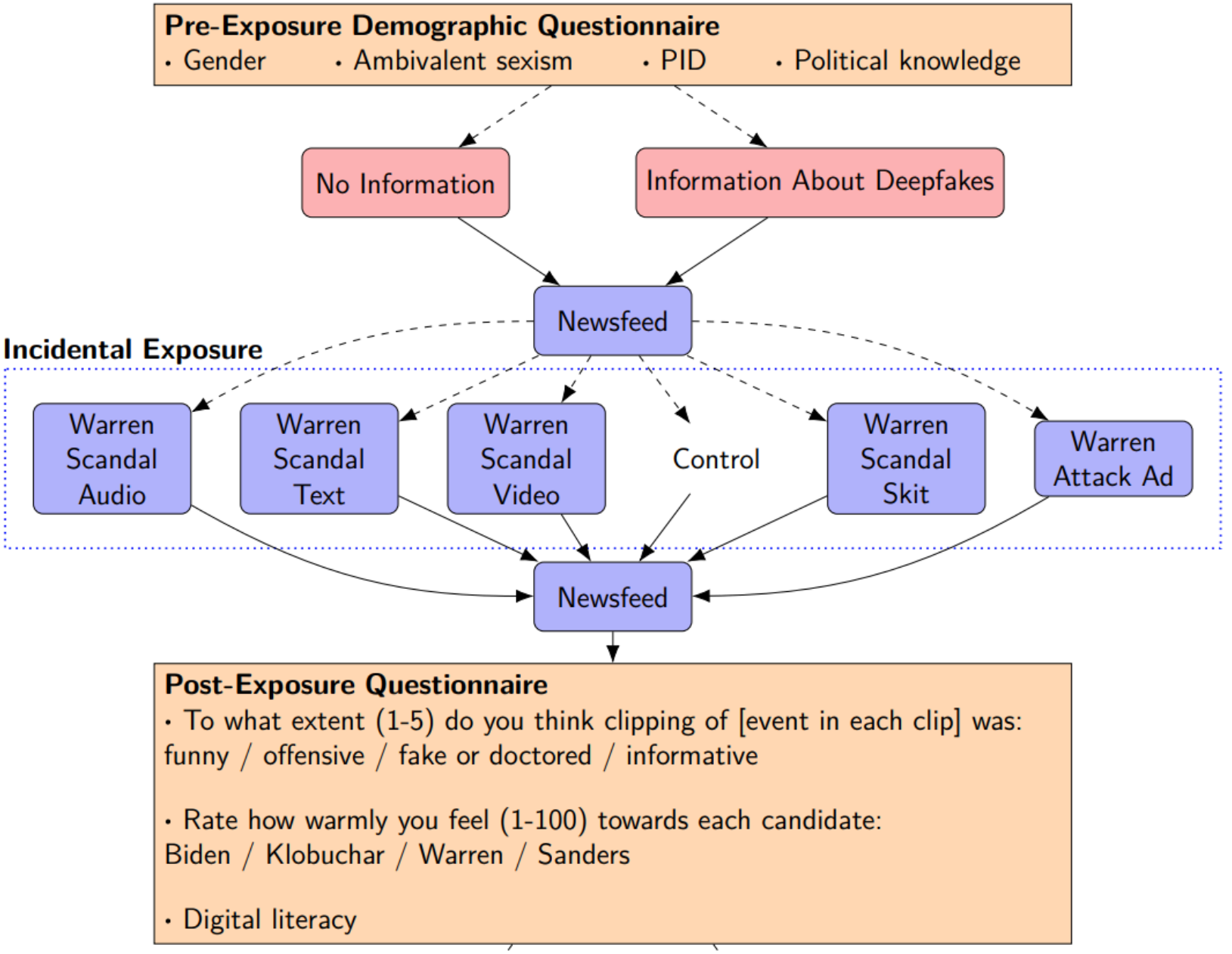
	Exposure(s)	Pre-Exposure Interventions	Outcomes
① Incidental Exposure	1. Authentic coverage of 2020 D candidates 2. **Randomized to `text`, `audio`, `video`, `skit` clip of E. Warren scandal, attack `ad`, or `control` (no stimuli)** 3. Authentic coverage of 2020 D candidates	• Info about deepfakes	• Credibility of clips • Affect towards candidates
② Detection Task	Random video feed: • **No-fake**: 8 authentic • **Low-fake**: 6 authentic, 2 deepfakes • **High-fake**: 2 authentic, 6 deepfakes	• Debrief deepfakes in ① • Acc prime	• Acc • FPR • FNR

Pre-Exposure Demographic Questionnaire

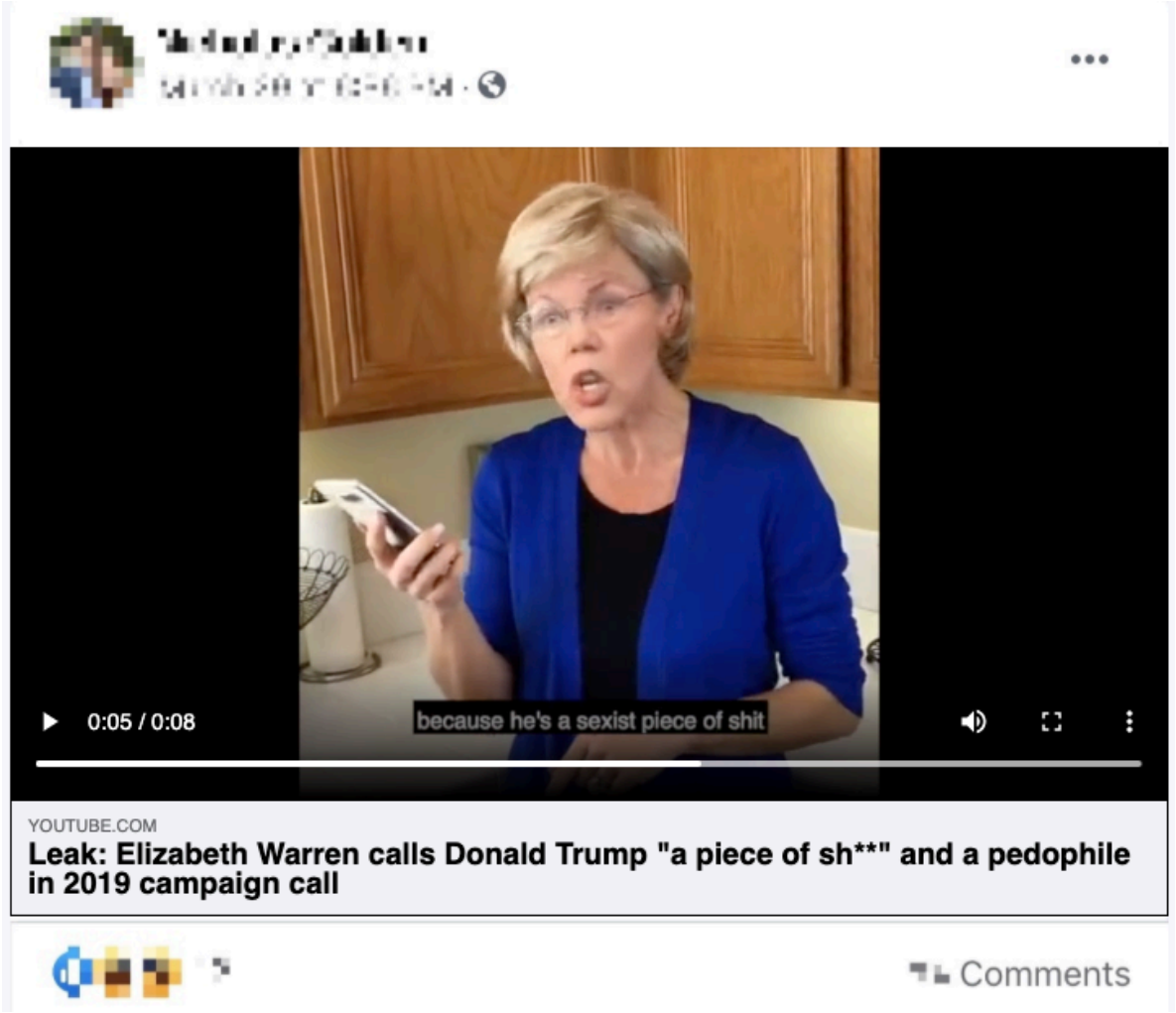
- Gender
- Ambivalent sexism
- PID
- Political knowledge


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
Warren Scandal Audio
Warren Scandal Text
Warren Scandal Video
Control
Warren Scandal Skit
Warren Attack Ad
Newsfeed

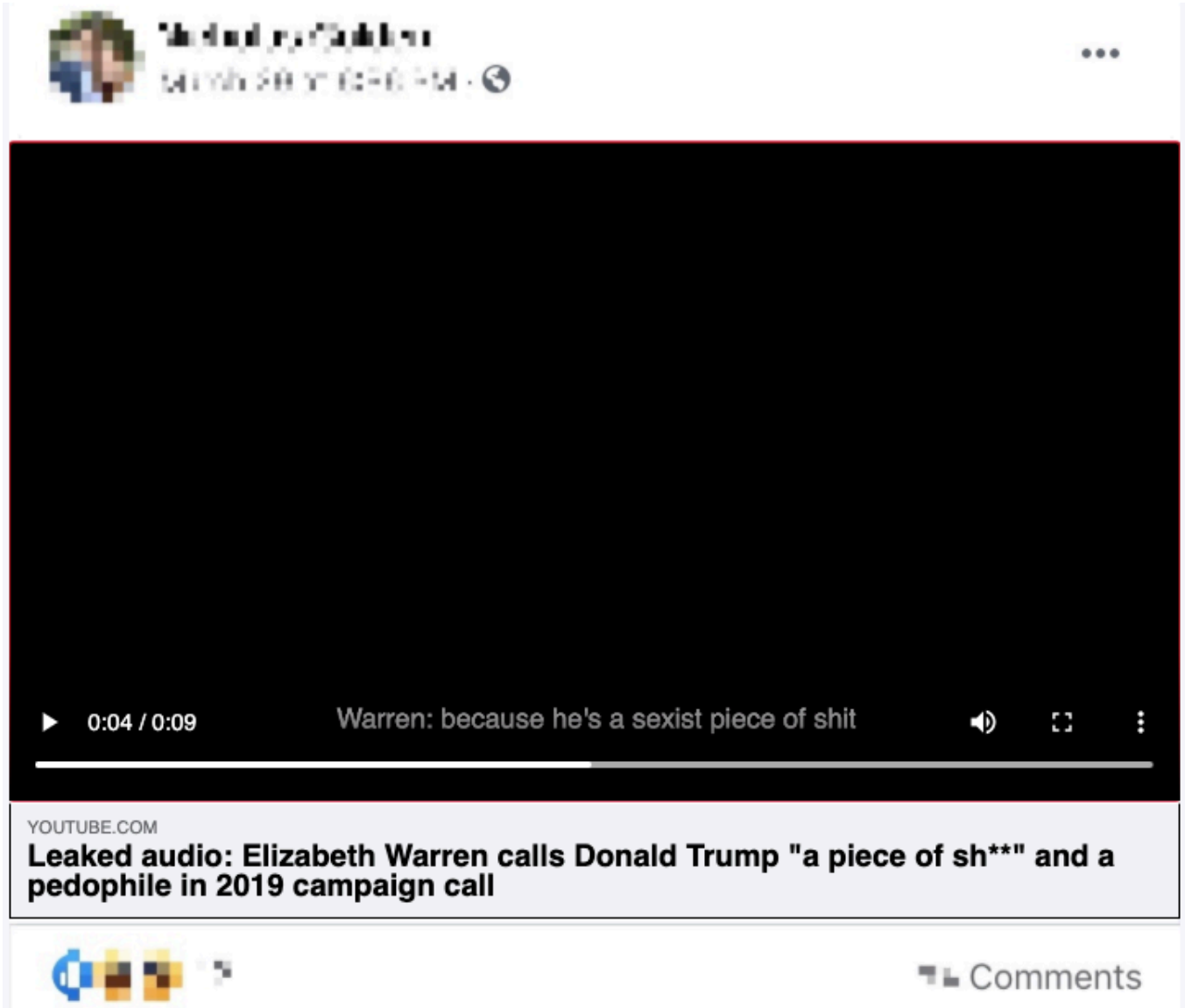

Pre-Exposure Demographic Questionnaire
• Gender
• Ambivalent sexism
• PID
• Political knowledge
No Information
Information About Deepfakes
Newsfeed
Incidental Exposure
Warren Scandal Audio
Warren Scandal Text
Warren Scandal Video
Control
Warren Scandal Skit
Warren Attack Ad
Newsfeed
Post-Exposure Questionnaire
• 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:



Watch Video

Example audio exposure:



Example text exposure:

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):

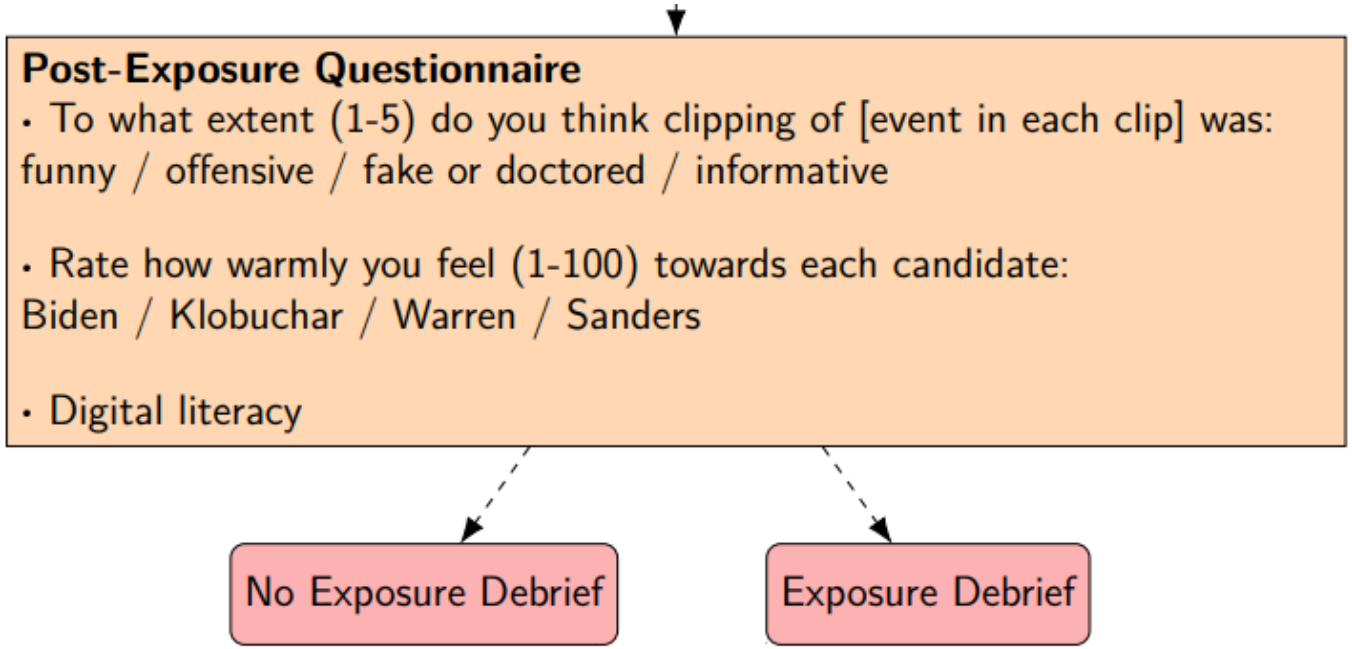
Reference affective exposure (ad):

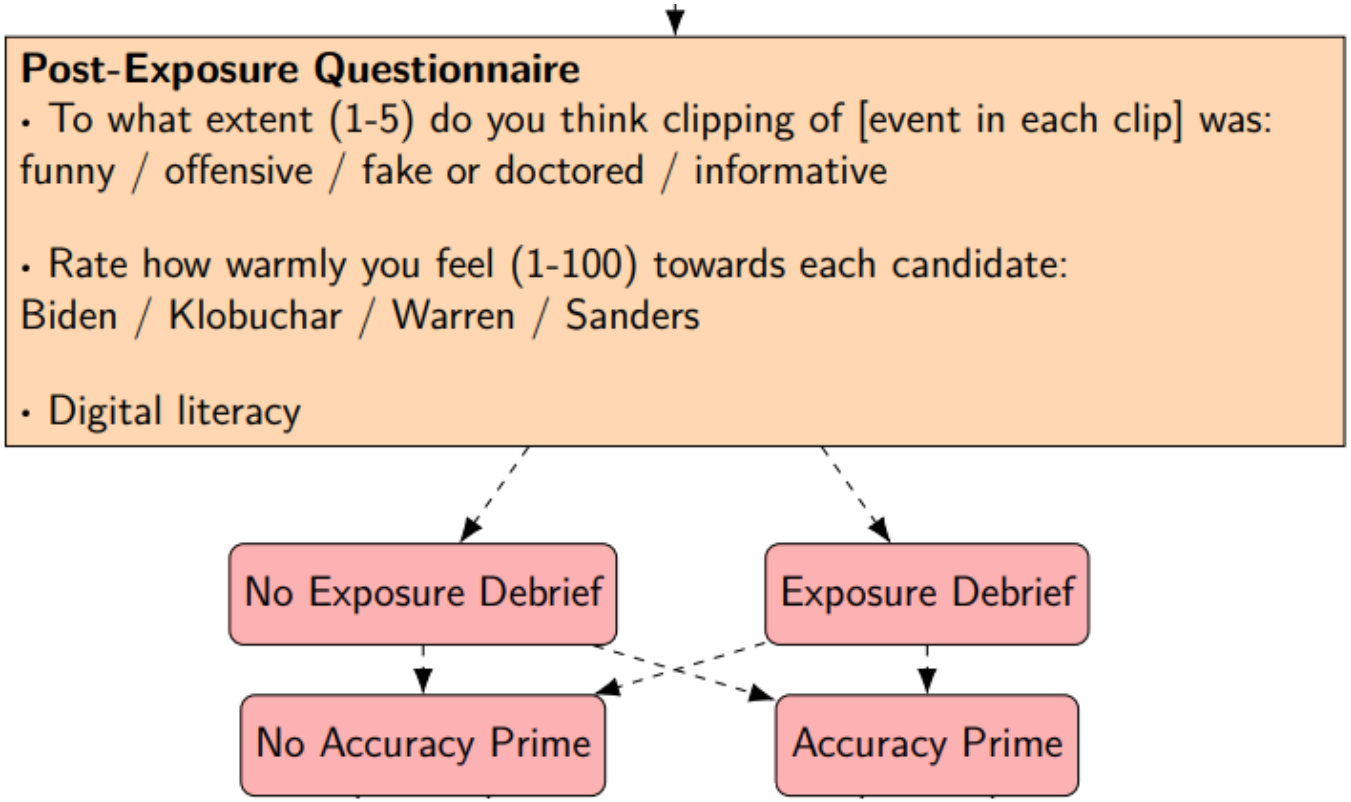
Post-Exposure Questionnaire

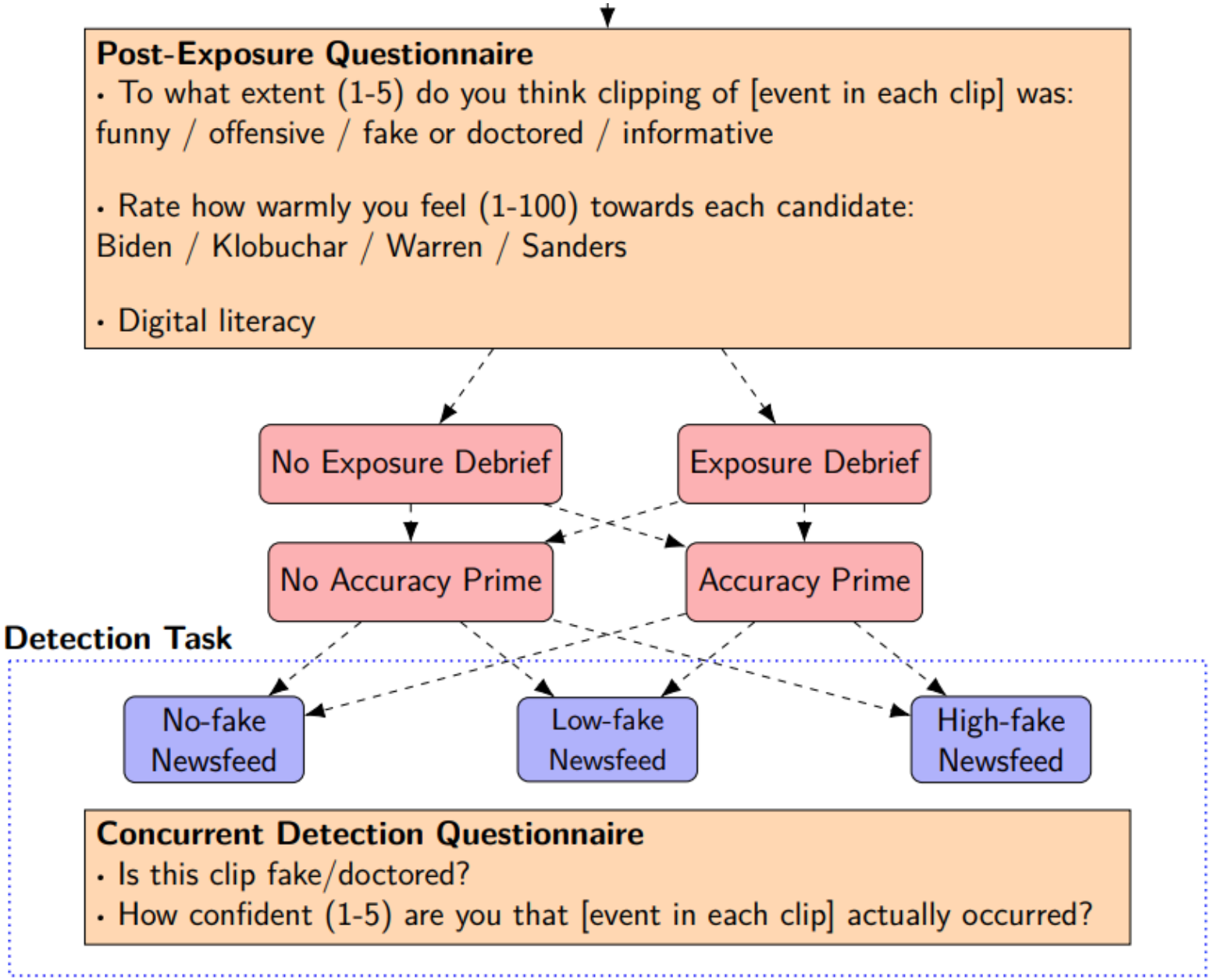
- To what extent (1-5) do you think clipping of [event in each clip] was: funny / offensive / fake or doctored / informative

- 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
No Exposure Debrief
Exposure Debrief
No Accuracy Prime
Accuracy Prime


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
No Accuracy Prime
Accuracy Prime
Detection Task
No-fake Newsfeed
Low-fake Newsfeed
High-fake Newsfeed
Concurrent Detection Questionnaire
• Is this clip fake/doctored?
• How confident (1-5) are you that [event in each clip] actually occurred?

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?**
  $\rightsquigarrow$ prime target for deepfake video: controversies, salience, gender, supply of impersonators

Many Trade-Offs Considered in Our Design

- **Why Warren in the incidental exposure?**
  $\rightsquigarrow$ 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?**
  $\rightsquigarrow$ prime target for deepfake video: controversies, salience, gender, supply of impersonators
- **Why those clips in detection task?**
  $\rightsquigarrow$ 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?**
  $\rightsquigarrow$ prime target for deepfake video: controversies, salience, gender, supply of impersonators
- **Why those clips in detection task?**
  $\rightsquigarrow$ 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?**
  $\rightsquigarrow$ prime target for deepfake video: controversies, salience, gender, supply of impersonators
- **Why those clips in detection task?**
  $\rightsquigarrow$ 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"?)**
  $\rightsquigarrow$ 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?**
  $\rightsquigarrow$ prime target for deepfake video: controversies, salience, gender, supply of impersonators
- **Why those clips in detection task?**
  $\rightsquigarrow$ 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"?)**
  $\rightsquigarrow$ 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?**
  $\rightsquigarrow$ no, *but*, if deepfakes are now indistinguishable from real videos, our findings hint that's still a problem

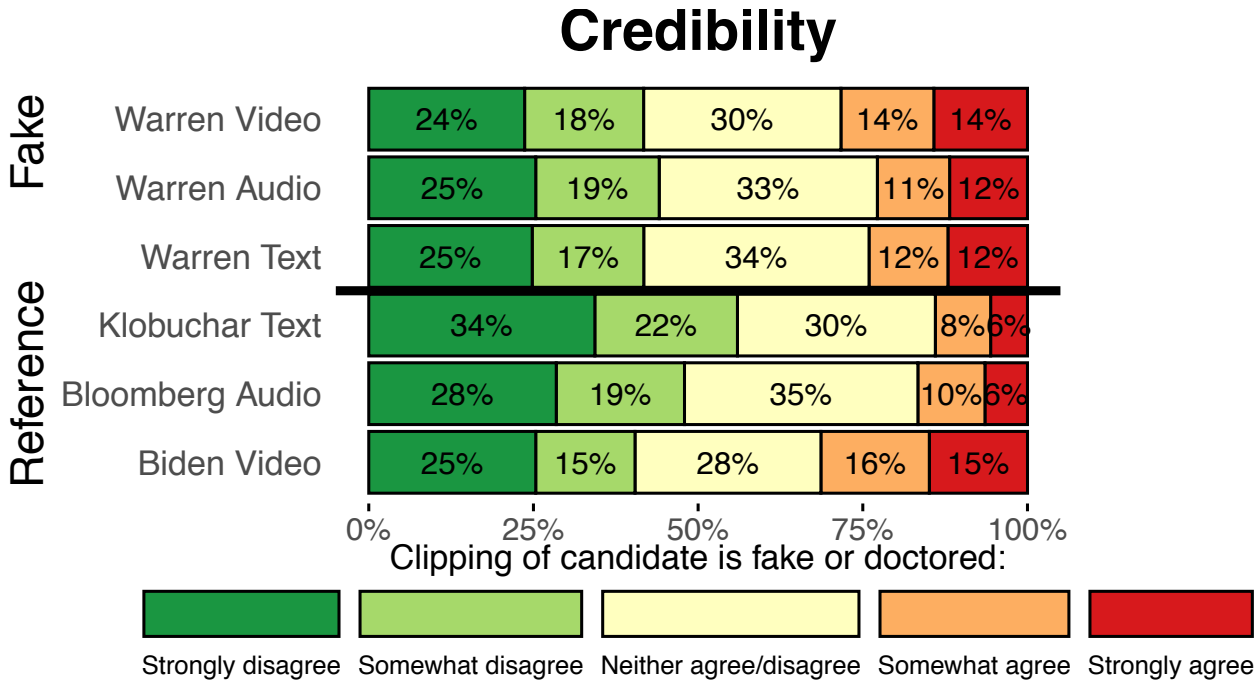
Findings

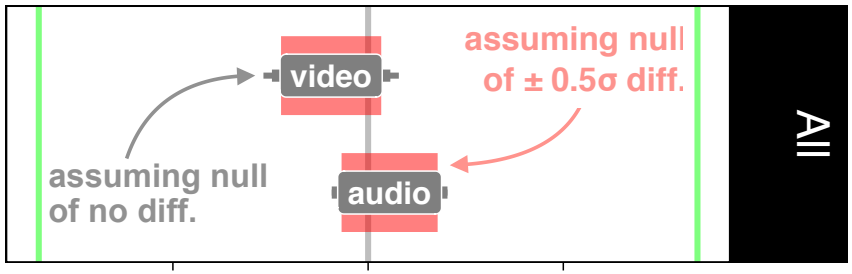
RQ1/2:

RQ1/2: No

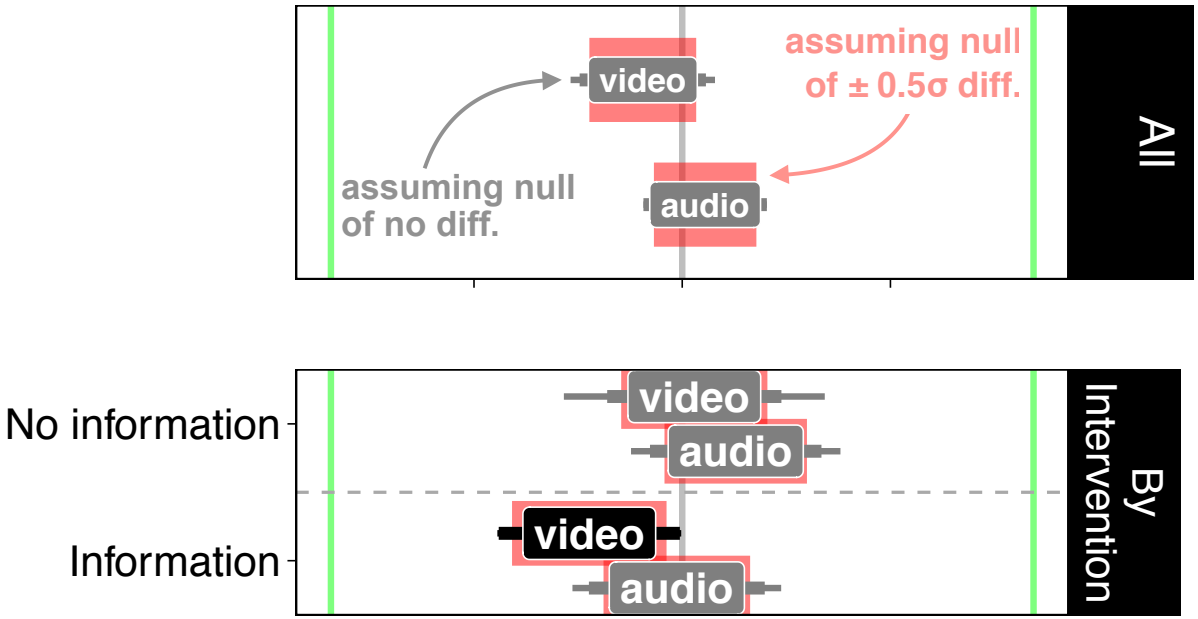
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)

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

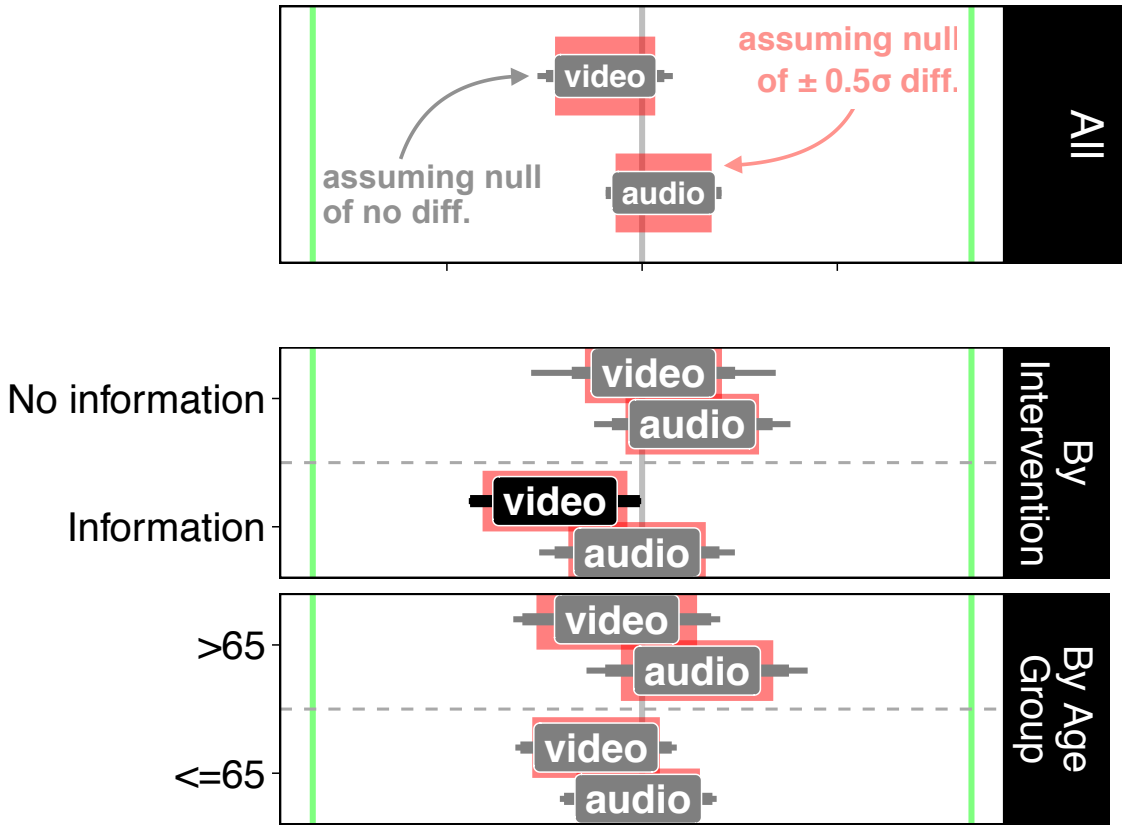




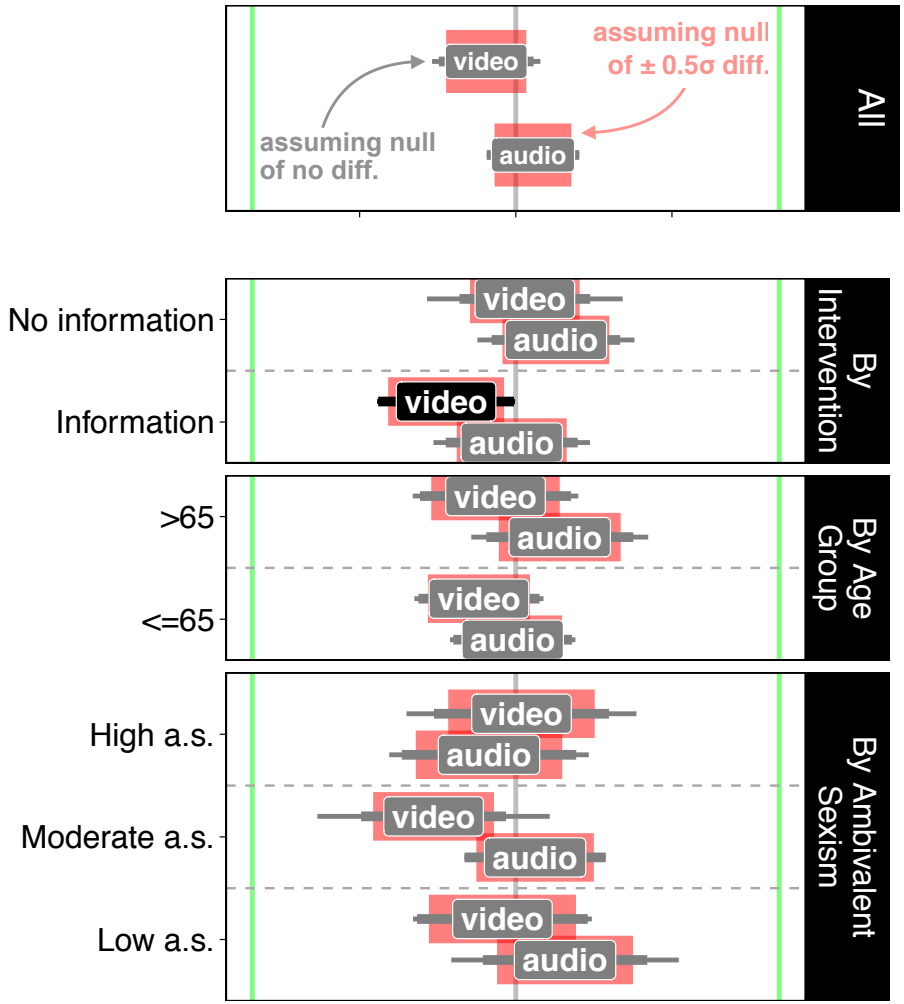
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



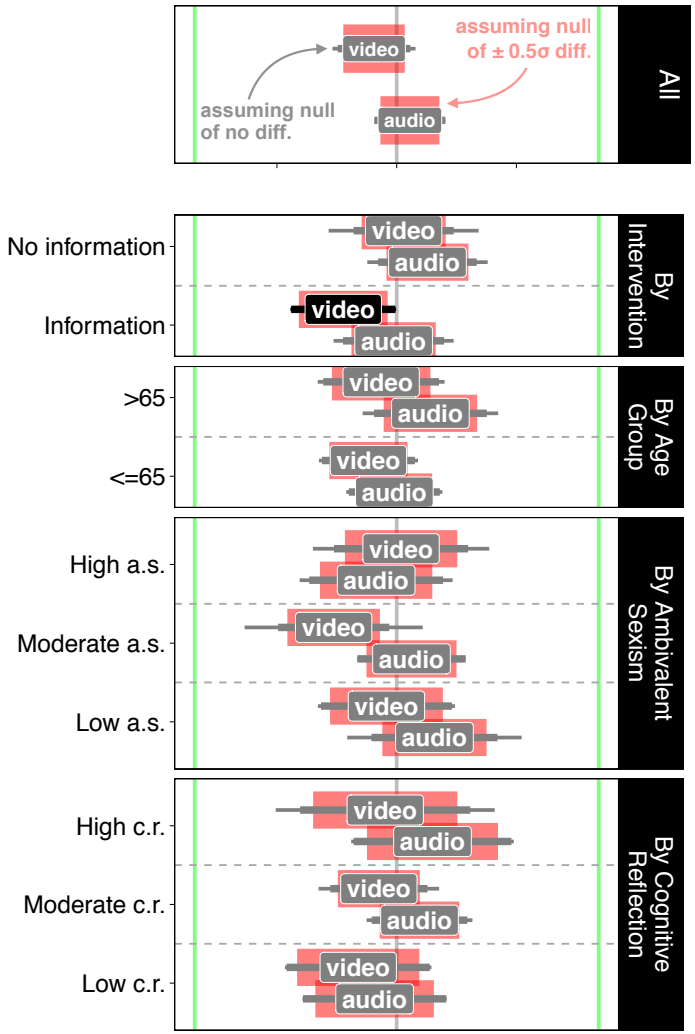
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



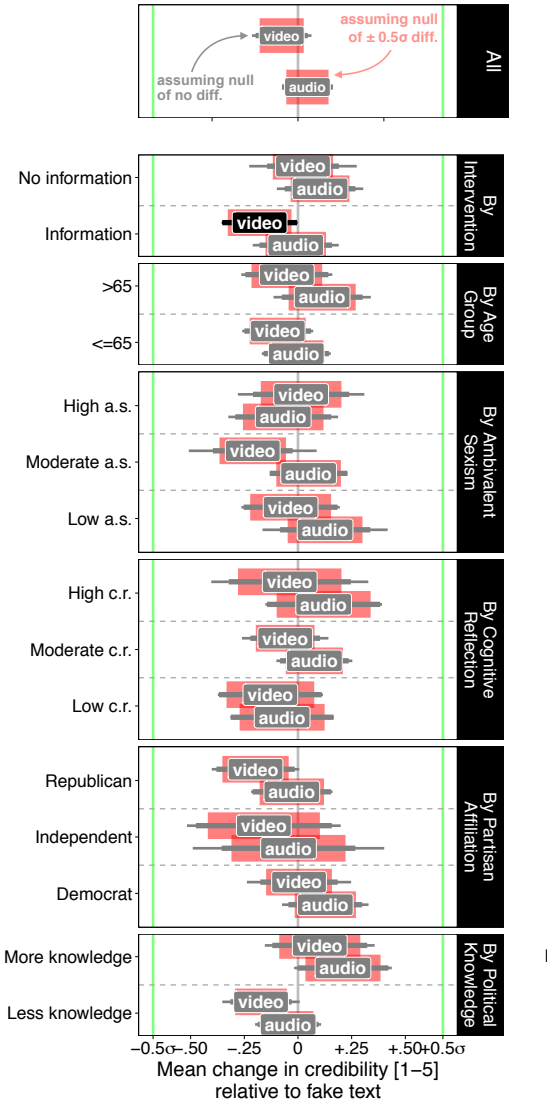
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



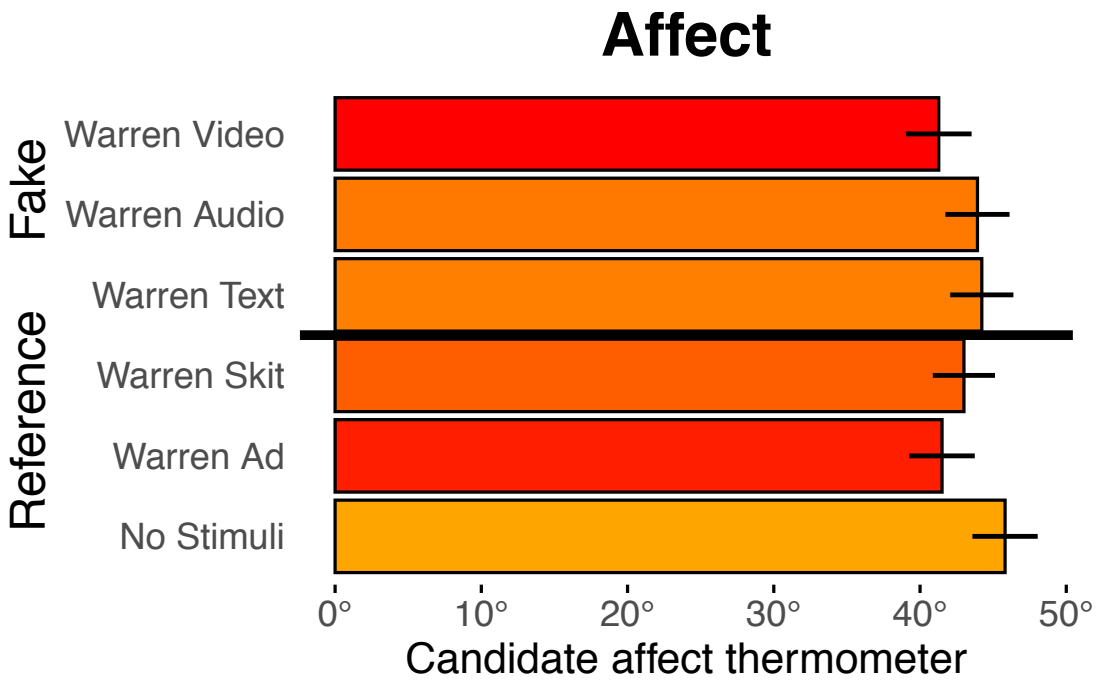
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)



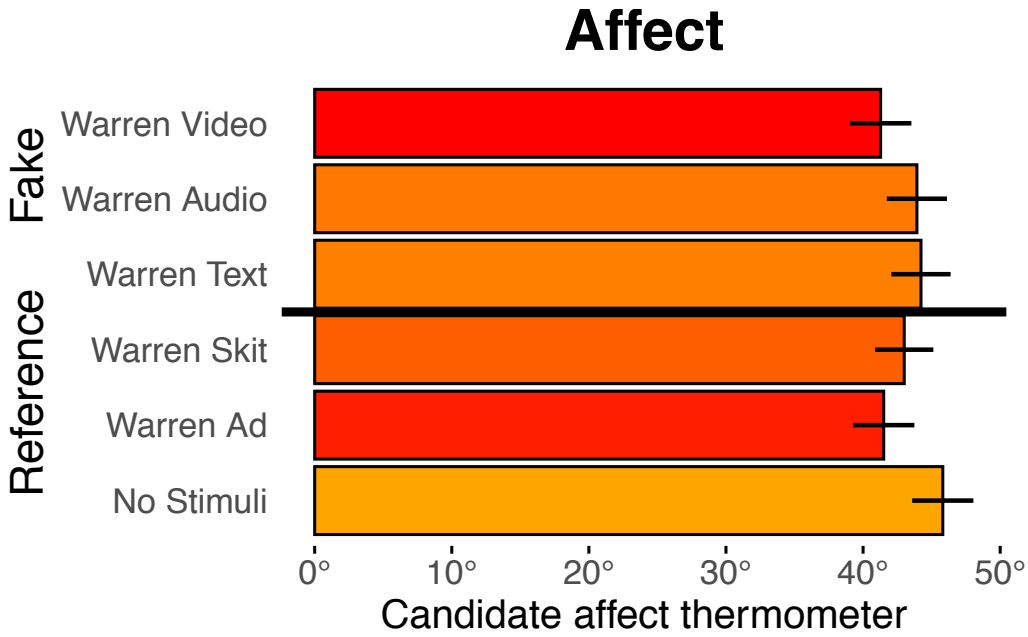
RQ1/2: No (deepfakes no more *credible* or affectively appealing than comparable fake media)

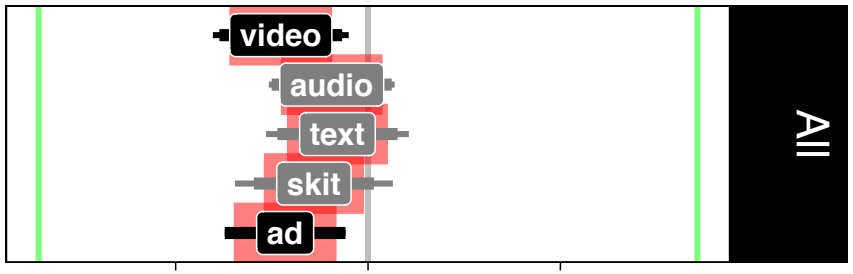


RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)

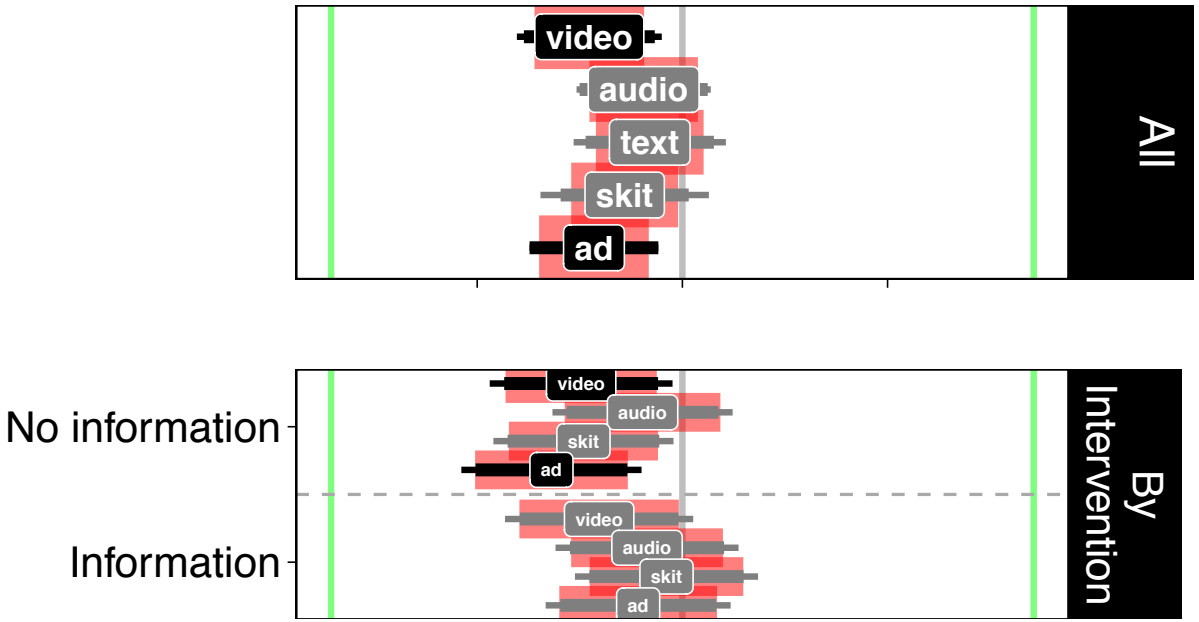


RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)

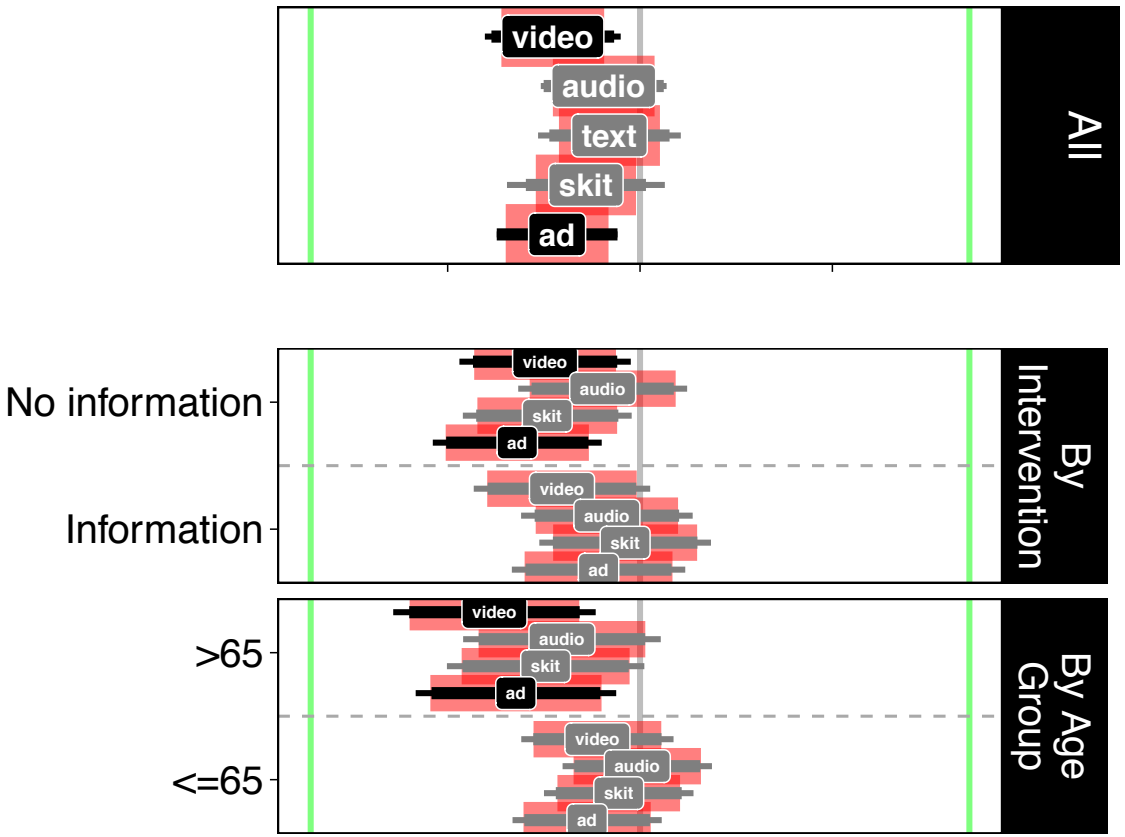




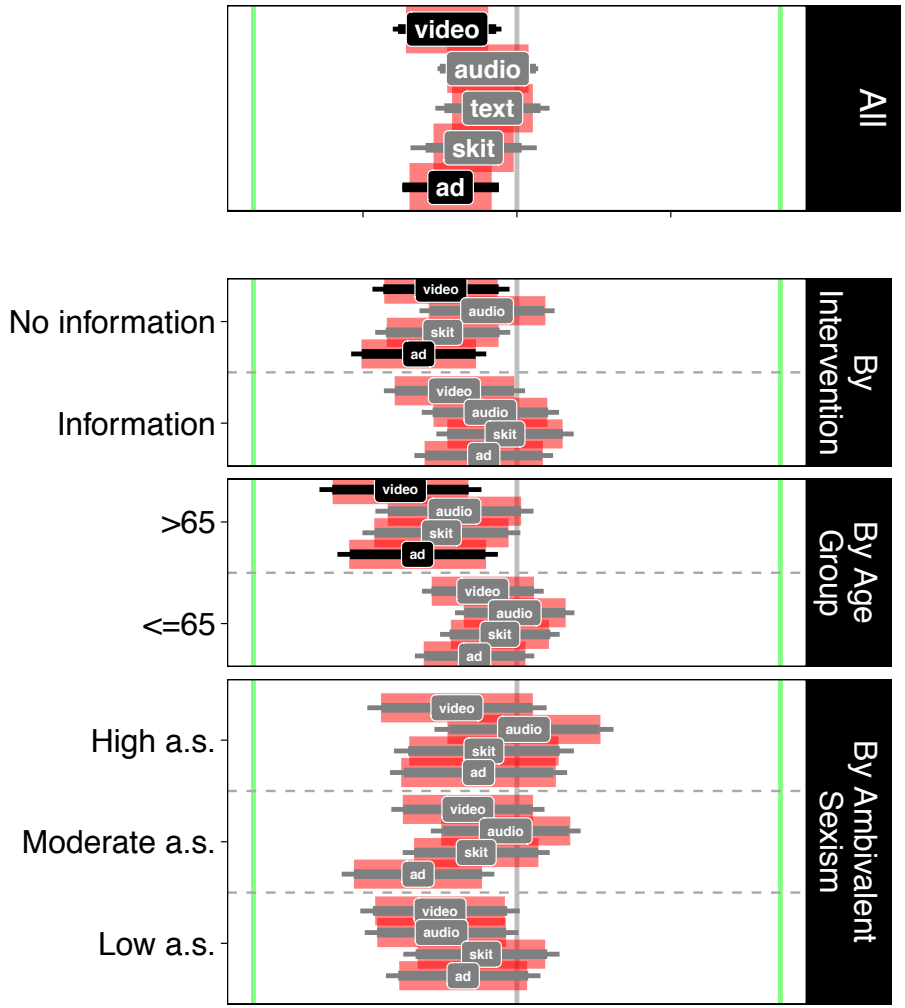
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



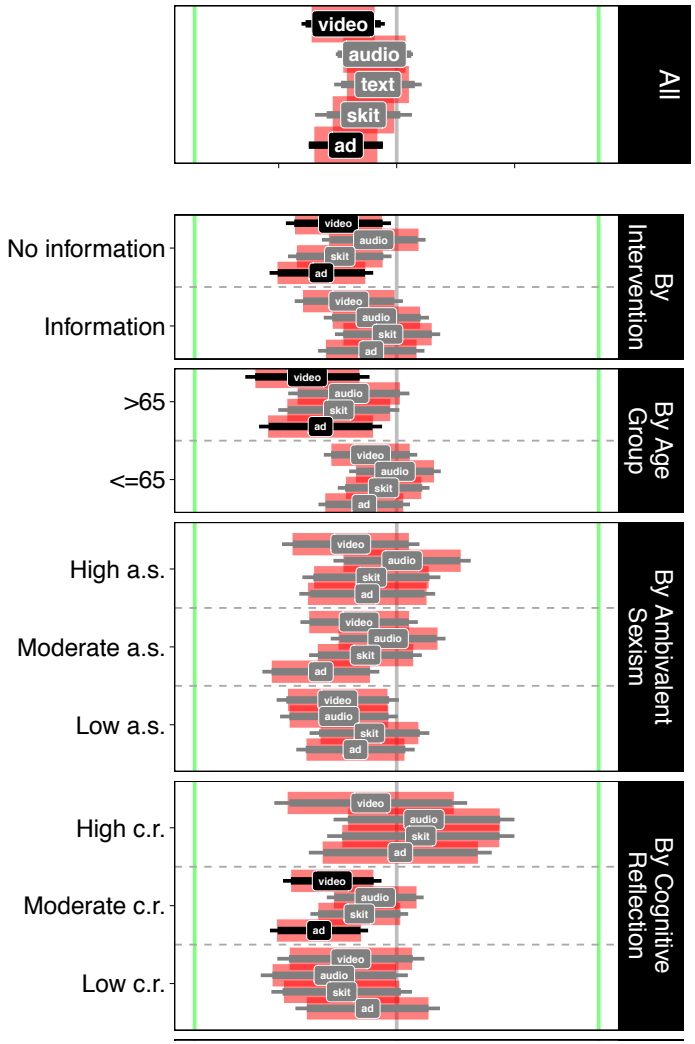
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



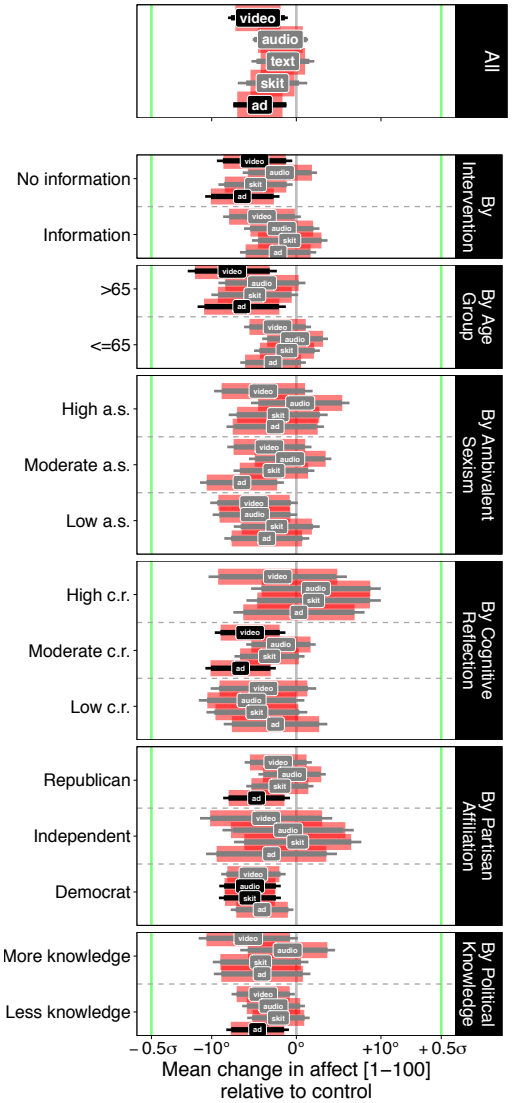
RQ1/2: No (deepfakes no more credible or *affectively* appealing than comparable fake media)



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)

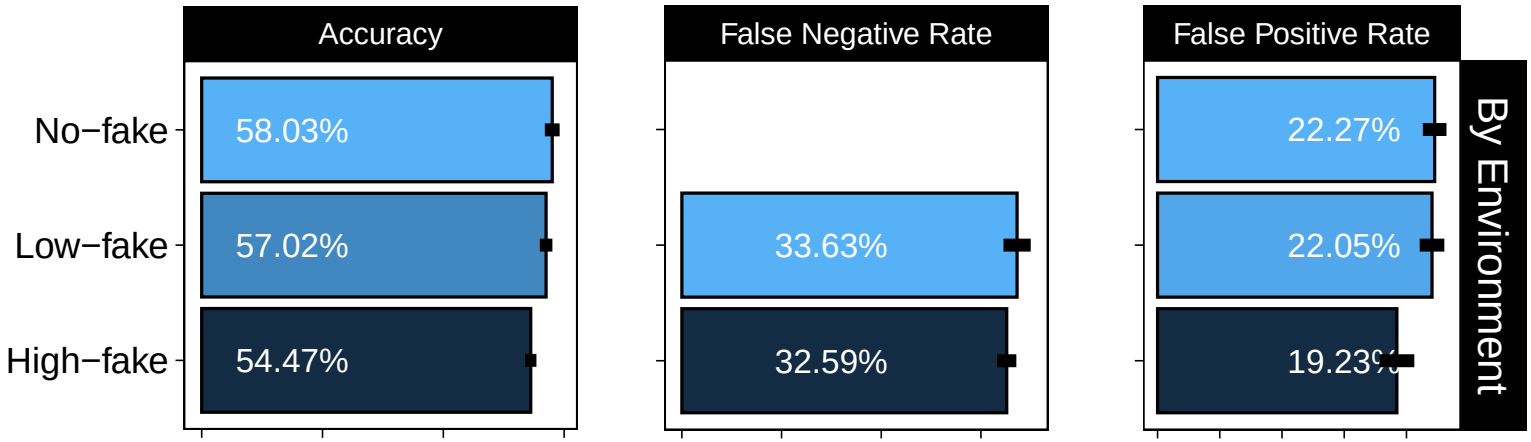


RQ3:

RQ3: Sorta

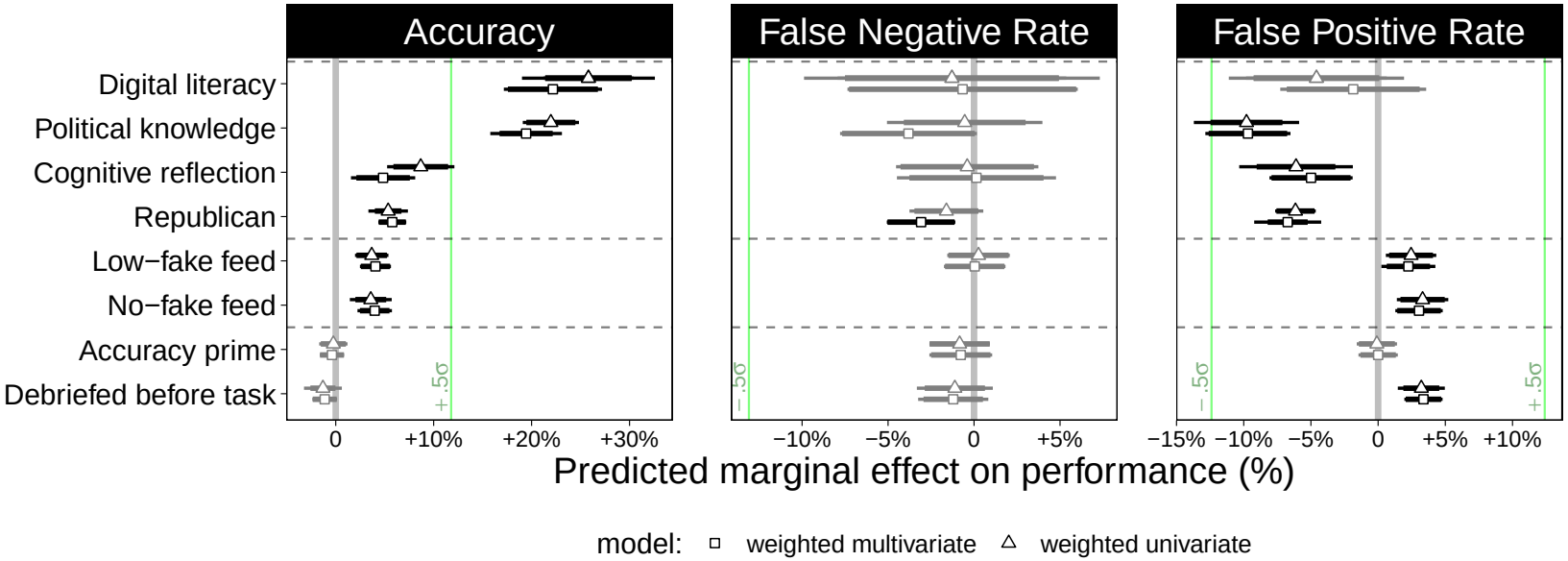
RQ3: Sorta (FNR higher than FPR)

RQ3: Sorta (FNR higher than FPR)



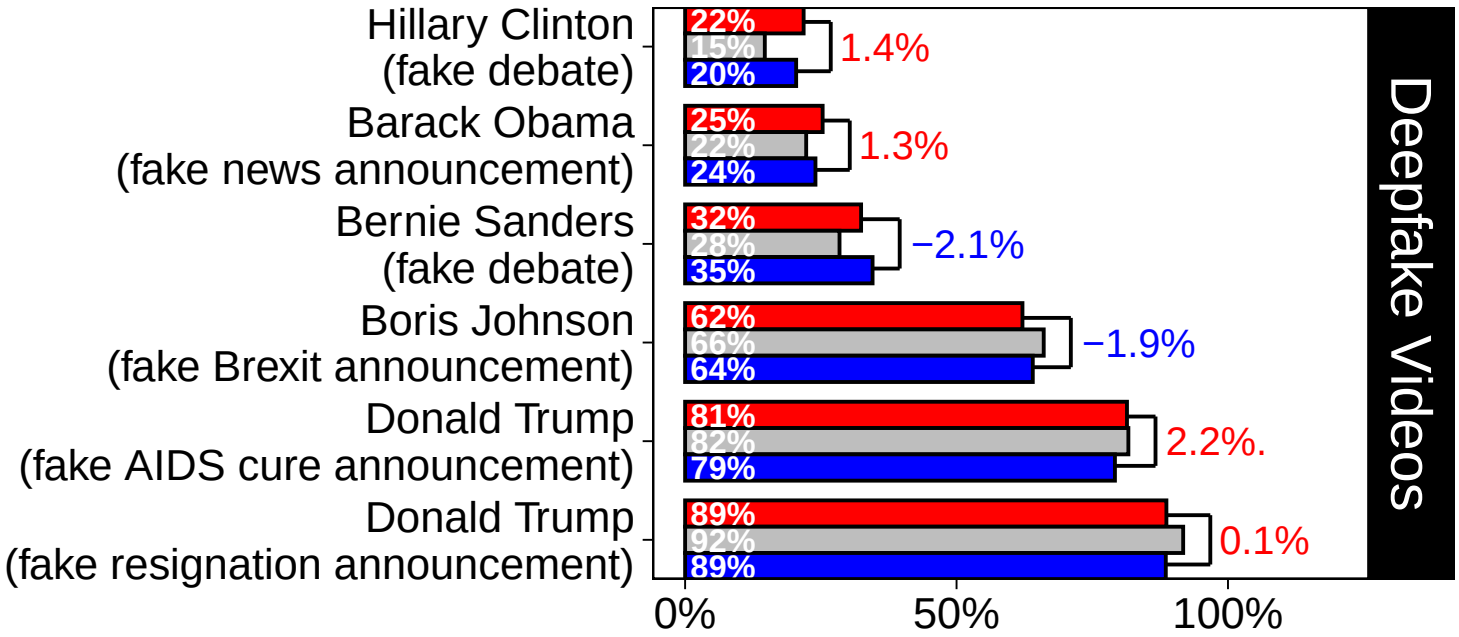
RQ3: Sorta (but, digital literacy and pol. knowledge improve FPR)

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

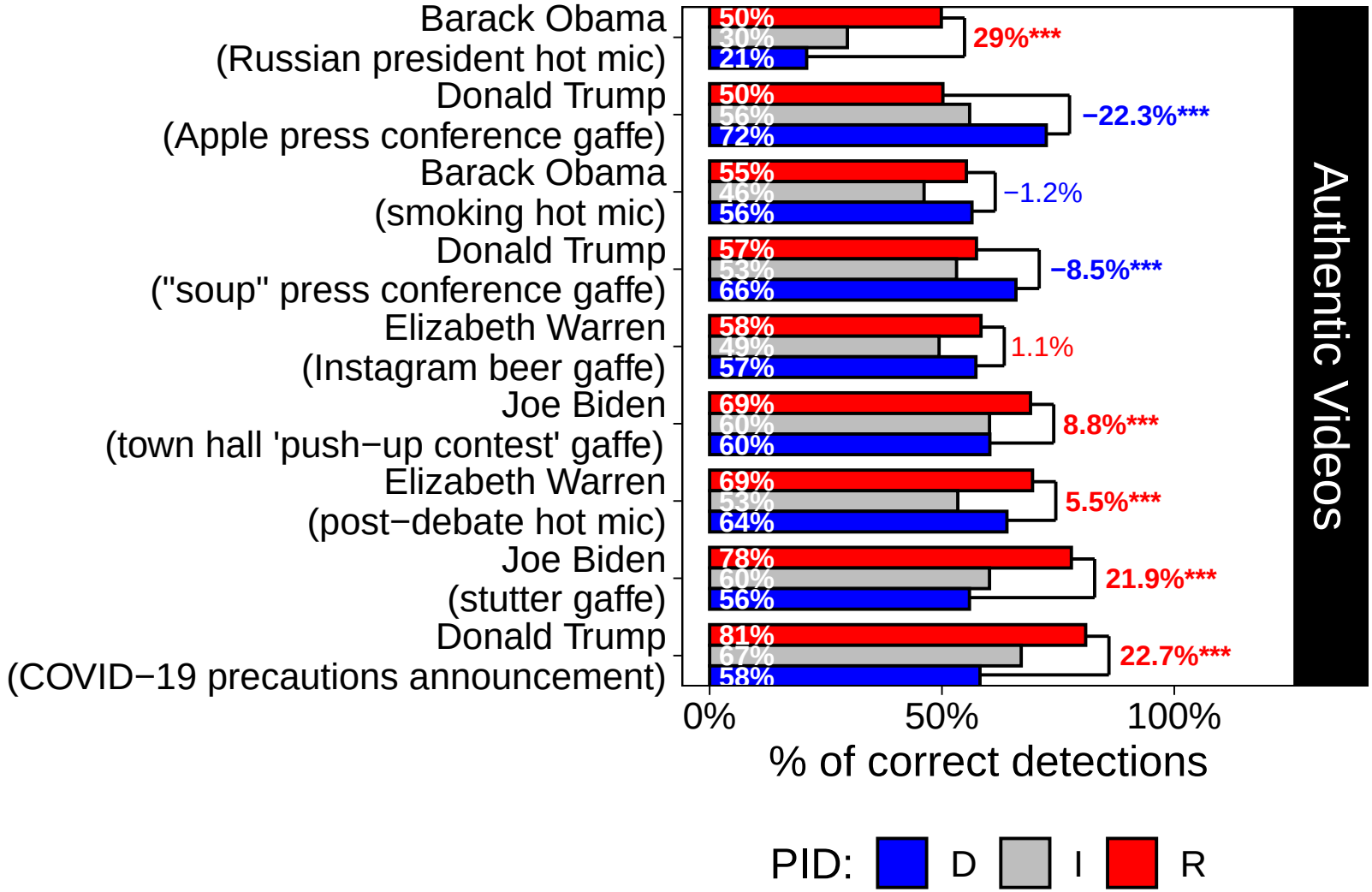


RQ3: Sorta (however, *significant* gap in FPR between Democrats and Republicans)

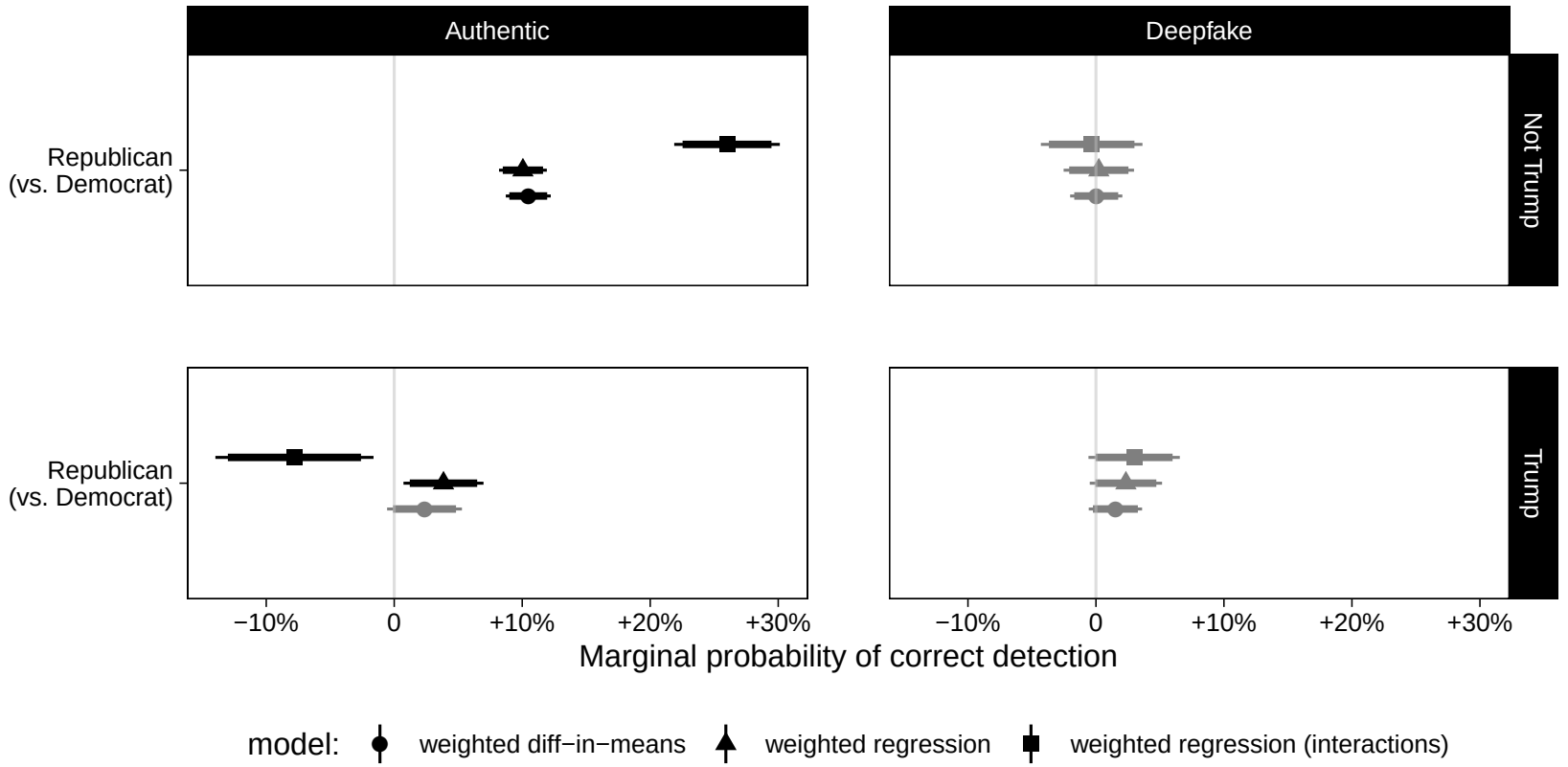
RQ3: Sorta (however, *significant* gap in FPR between Democrats and Republicans)



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



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



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 $\rightsquigarrow$ 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 $\rightsquigarrow$ 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