THE IMPACT OF NEW TECHNOLOGY ON SOCIETY

DO TRANSPARENCY LABELS ON POLITICAL DEEPFAKES PROTECT DEMOCRATIC TRUST? EVIDENCE FROM A SURVEY EXPERIMENT AMONG DUTCH TIKTOK USERS

Natalia Vliches Poperecinaia, Marie-Julie Cantraine, Luisa Homola, Thomas Dubuisson, Sergio Jiménez Puspito, and Leon Schwabe

INTRODUCTION

AI-generated synthetic media — "deepfakes" — spread rapidly on platforms like TikTok, threatening the ability of citizens to distinguish authentic political content from manipulation. Beyond deceiving individuals, deepfakes generate systemic informational uncertainty: even authentic videos may be doubted.

➔ This creates a liar's dividend (Chesney & Citron, 2019): politicians can discredit real footage by calling it fake. The EU AI Act (2024, Art. 50) responds with transparency labels, requiring AI-generated content to be disclosed. But does labelling actually help users?

Research question

Does exposure to a AI-labelled political deepfake affect:

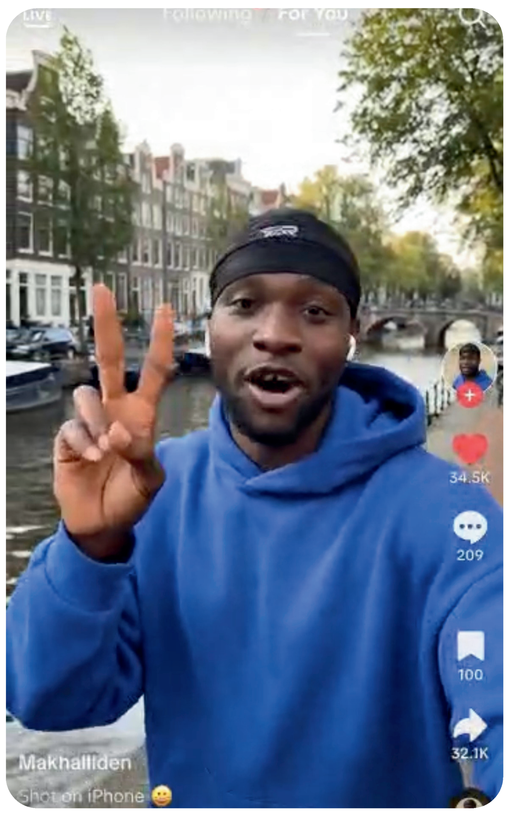
- Perceived credibility?
- Trust in political information on TikTok?

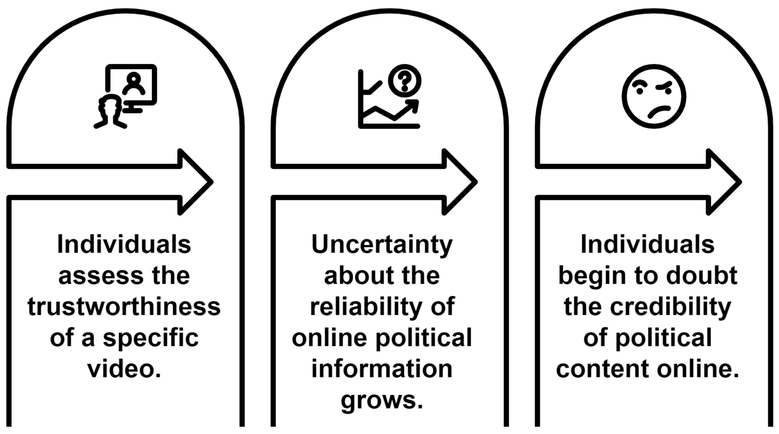
THEORETICAL FRAMEWORK

The study proposes a three-level cascade:

① Perceived credibility of the specific video
↓ may generalise to
② Epistemic beliefs about social media information
↓ may ultimately erode
③ Broader trust in political communication

→ Pre-existing political trust acts as an anchor: citizens interpret new information through prior beliefs (Akerlof & Shiller, 2009).





METHODOLOGY

Design: *Three-arm between-subjects online survey experiment (Qualtrics).*

→ Participants randomly assigned to one of three video conditions:

- Authentic video
- Unlabelled AI-generated
- Labelled AI-generated

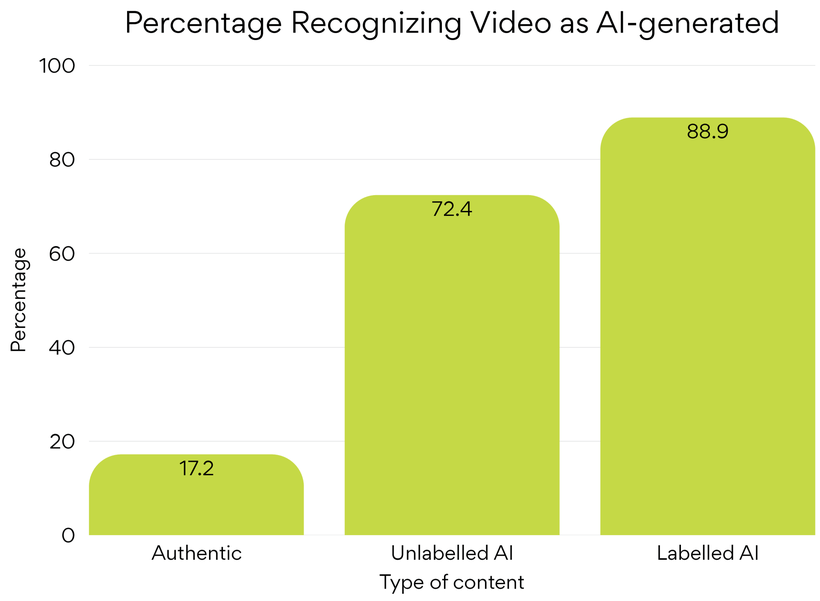
Sample

- N = 99 Dutch adult TikTok users
- Recruited via LinkedIn, WhatsApp university groups, personal networks
- Non-TikTok users screened out

KEY FINDINGS

Condition	Authentic	Unlabelled AI	Labelled AI
Mean	3.50	3.25	3.29

p = 0.699
=> no significant difference across conditions



LIMITATIONS

- Convenience sample — overrepresents students and highly educated users
- Small N (99) — likely underpowered for small effects
- Single video stimulus — limits external validity
- Controlled setting lacks TikTok's algorithmic and social context
- Label-recognition not directly measured
- Cross-sectional design — cannot capture cumulative exposure effects

DISCUSSION

Passive disclosure ≠ critical processing.
Users noticed the label and recognised the content as AI-generated, but this awareness did not translate into scepticism.
➔Recognition is necessary but insufficient for effective labelling.

Pre-existing trust dominates.
Political trust was a stronger predictor of credibility than the experimental treatment.
➔Users filter new information through their prior beliefs. A finding consistent with the liar's dividend logic (Chesney & Citron, 2019).

Economic framing.
Deepfakes function as a market for lemons (Akerlof, 1970): if synthetic and authentic videos are visually indistinguishable, users lose confidence in the entire information market.
➔ Labels address part of the asymmetry but do not resolve the underlying quality-uncertainty problem.

Legal implications.
The AI Act's transparency obligation rests on the assumption that labelling mitigates harm. This study provides empirical evidence that labels alone are insufficient.
➔ Under the DSA, TikTok may need to go further: algorithmic interventions reducing the amplification of unverified synthetic content may be required as proportionate mitigation measures.

CONCLUSION

- Transparency labelling obligations increased recognition of AI-generated content
- However, pur study demonstrated that transparency labels did not reduce perceived credibility or significantly or alter broader political trust
- These findings challenge the governance logic of the EU AI Act and call for a regulatory recalibration.
- Future policy should move beyond passive disclosure towards complementary interventions, including:
  - Algorithmic demotion of unverified, synthetic content
  - Provenance standards,
  - Efficient systemic risk mitigation under the DSA.

Future research

- Repeated exposure to deepfake
- Alternative label designs (probabilistic vs. definitive)
- Experiments in native platform environments to increase ecological validity

REFERENCES (SELECTED)

AI ACT (2024) · DSA (2022) · AKERLOF (1970) · CHESNEY & CITRON (2019) · KOHRING & MATTHES (2007) · NORRIS (2014) · PENNYCOOK ET AL. (2020) · SEYD (2016) · VACCARI & CHADWICK (2020) · LI & YANG (2024) · WITTENBERG ET AL. (2025) · ŁABUZ (2025)

Maastricht University