

Trust and Truth in the Age of Generative AI



ILLINOIS TECH



ILLINOIS INSTITUTE OF TECHNOLOGY

A Global South Perspective

Dr. Dilrukshi Gamage

University of Colombo School of Computing · Microsoft Research Fellow

ex affiliate Berkman Klein Center, Harvard University

ex Research Fellow Weizenbaum Institute, Berlin · Centre for Protecting Women Online, The
Open University UK



9th Oct 2026

About who we are - University of Colombo School of Computing

← → ↺

cmb.ac.lk/tag/ranking

📄 ⭐

📥

ting Started McAfee Security Dell Upper extremity prost... OpenEdMOOC – Intro... Home HIHC-S-2026-4488_Pr... https://www.facebook... Microsoft Word - hans... retrieving - Google Se...

Alumni Staff Gallery IT Services Pay Library International UCIC 2026 Contact Us

f

📷

in

🐦

📺

🎵

🔍



කොළඹ විශ්වවිද්‍යාලය
கொழும்புப் பல்கலைக்கழகம்
UNIVERSITY OF COLOMBO

ABOUT

ACADEMIC

RESEARCH

COURSES

ADMINISTRATION

STUDENTS

NOTICE

ABOUT

ACADEMIC

RESEARCH

COURSES

ADMINISTRATION

STUDENTS

NOTICE

ABOUT

ACADEMIC

RESEARCH

COURSES

ADMINISTRATION

STUDENTS

NOTICE

ranking

HOME \

TAG "RANKING"

AUG 08

HOME NORTH AMERICA LATIN AMERICA EUROPE ASIA AFRICA ARAB WORLD OCEANIA RANKING BY A

Home » Asia » Sri Lanka

Current edition

July 2016 Edition
2016.2.1 (corrected)

About Us

• About Us

• Contact Us

About the Ranking

• Methodology

• Objectives

• FAQs

• Rules

• Previous editions

Resources

• Best Practices

• Links

Sri Lanka

Ranking	World Rank	University	Det.	Presence Rank	Impact Rank
1	1946	University of Colombo	👍	2653	3965
2	2158	University of Peradeniya	👍	4591	6091
3	2415	University of Moratuwa	👍	2116	3487
4	2559	University of Kelaniya	👍	1593	6696
5	2634	University of Ruhuna	👍	314	8064
6	3045	University of Sri Jayawardenapura	👍	1625	6582
7	4004	University of Jaffna	👍	4547	9910
8	4121	University of the Visual & Performing Arts	👍	2485	1451
9	4479	Open University of Sri Lanka	👍	1241	9492
10	5489	Wimala University of Sri Lanka	👍	5084	8050
11	6015	Sabaragamuwa University	👍	11834	12424

World University Web Ranking – July 2016

BY UNIVERSITY OF COLOMBO UNIVERSITY OF COLOMBO NEWS RANKING NEWS

DETAIL >

SEARCH

🔍

LATEST POST



Post of Project Coordinator – Centre for Nutrition, Faculty of Medicine

OCTOBER 8, 2026



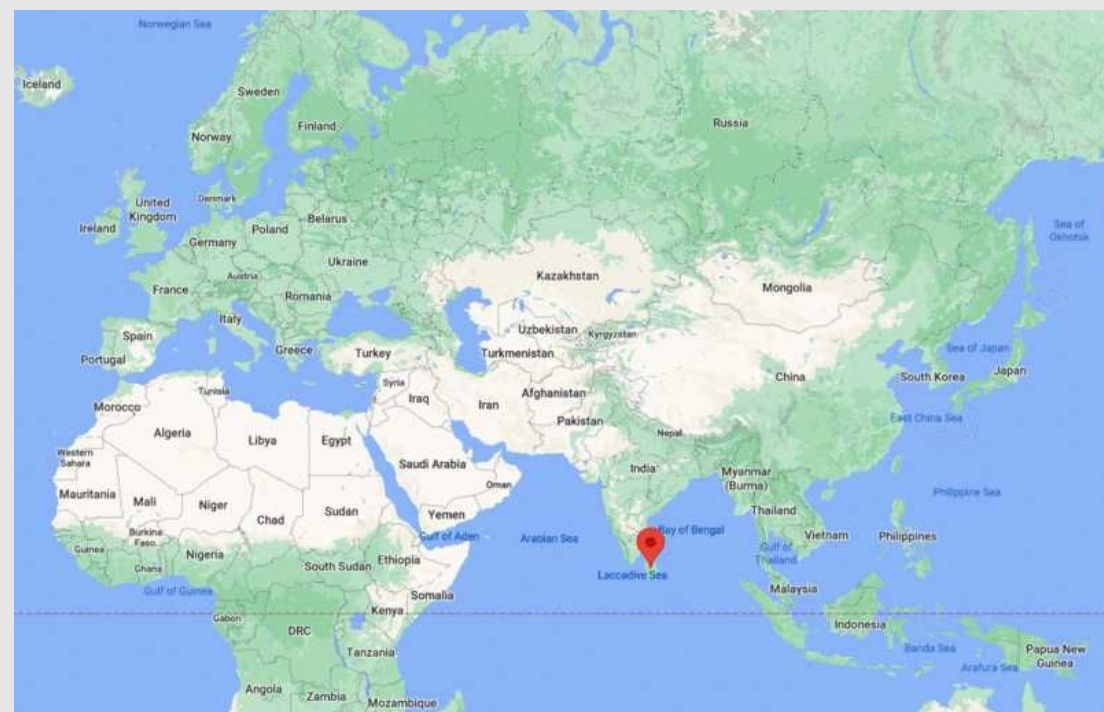
Post of Deputy Director – Centre for Quality Assurance (CQA) | Internal Applicants Only

OCTOBER 6, 2026



Notice: Renting out

Where is Colombo?



Whats my role at MSR?

<https://www.microsoft.com/en-us/research/project/project-provenance/fellowships/>



Research

Our research ▾

Programs & events ▾

News ▾

About ▾

Register: Research Forum

All Microsoft ▾

Search 🔍

Project Provenance

Empowering everyone to understand digital content provenance through human-centric design



Designing Provenance Tools That Help People Understand AI-Modified Media



[Dilrukshi Gamage](#)
Fellow
Assistant Professor, University of Colombo School of Computing
Sri Lanka

Labeling Across the Continuum of AI Involvement and the Provenance of Provenance



[David G. Rand](#)
Fellowship Advisor
Professor, Cornell University
United States



[Gabrielle Peloquin-Skulski](#)
Fellow
PhD Student, MIT
United States



Neeltje Berger
Senior Designer

[Learn more >](#)



Jingya Chen
UX Designer
Microsoft

[Learn more >](#)



Pratik Ghosh
Senior Research Designer

[Learn more >](#)



Eric Horvitz
Chief Scientific Officer

[Learn more >](#)



Viktor Kewenig
Cognitive Science Researcher



Clare Morgan
Senior Business Program Manager, MSR
Cambridge

[Learn more >](#)



Sean Rintel
Sr. Principal Research Manager

[Learn more >](#)



Abdullah Safir
Research Intern

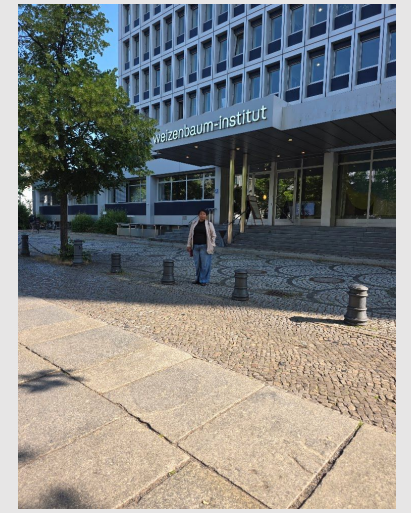
[Learn more >](#)



Abigail Sellen
VP/Distinguished Scientist



Jessica Young
Director, Science & Technology Policy



Whats my Focus?



DR. DILRUKSHI GAMAGE

[HOME](#) [PUBLICATIONS](#) [AWARDS & FELLOWSHIPS](#) [INVITED TALKS](#)



This is a AI Generated image of Dilrukshi, I may or maynot look like this;

Contact Me

Preferred email- dilrukshi.gamage@gmail.com

Institutional email - dgs@ucsc.cmb.ac.lk

dgame@cyber.harvard.edu

Twitter - [@Dilrukshi_LK](#)

Github - [@iceLearn](#)

LinkedIn - [dilrukshigamage](#)

[Google Scholar](#)

I'm in a mission to expand HCI in South Asia
Interest in joining my research group?

[Project Provenance](#) centered at [Microsoft Research Cambridge](#), with collaborators across the Office of the Chief Scientific Officer, [Microsoft Research in Redmond](#), and other teams . At the same time, I work as a Senior Lecturer at the [University of Colombo School of Computing](#) in Colombo, Sri Lanka and also a visiting research fellow at the Center for Protecting Women Online (CPWO) at the Open University UK. Previously, I worked with the Google Trust and Safety Team through Google Academic Research Award (GARA). I was also a research fellow at the [Weizenbaum Institute in Berlin, Germany](#) and also served as an [Affiliate at the Berkman Klein Center for Internet Society at Harvard University as well](#). Soon after my Ph.D., I was a [postdoctoral fellow](#) at the [Sasahara Lab, Department of Innovation Science, Tokyo Institute of Technology](#) (Tokyo Tech) in Japan.

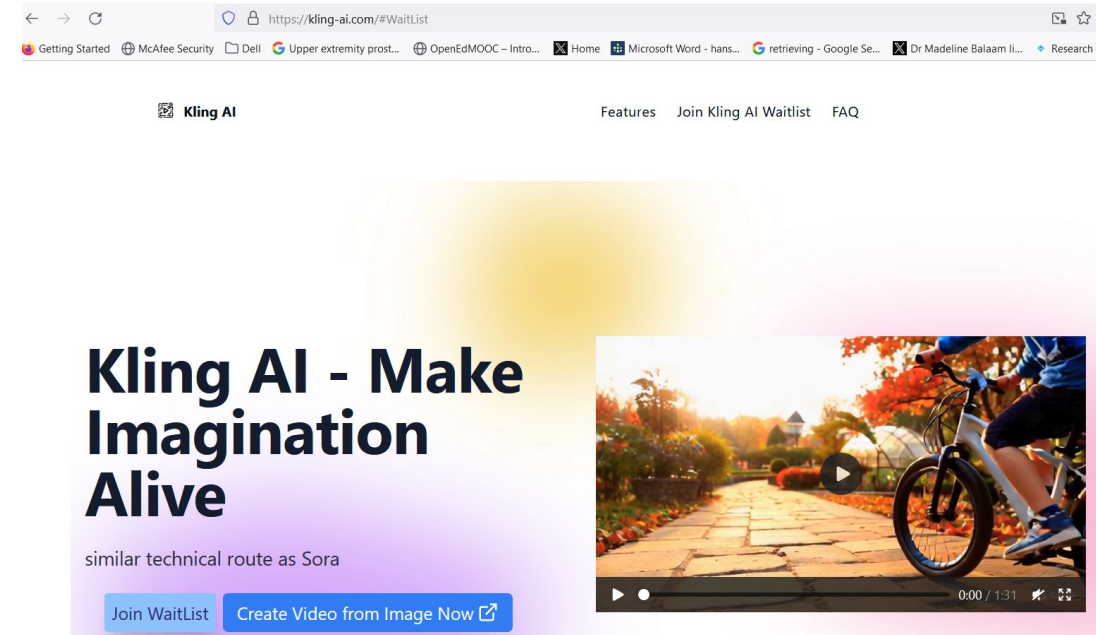
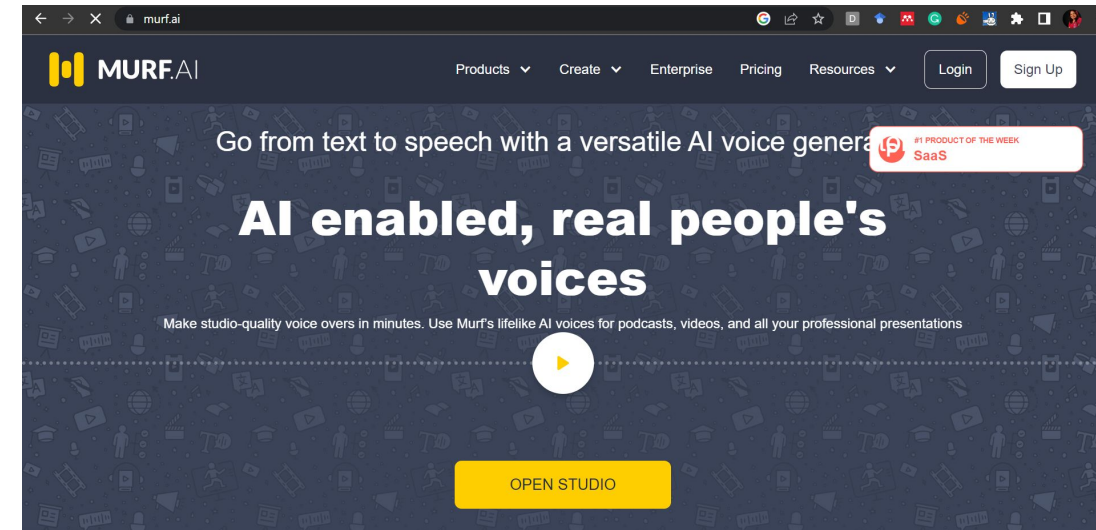
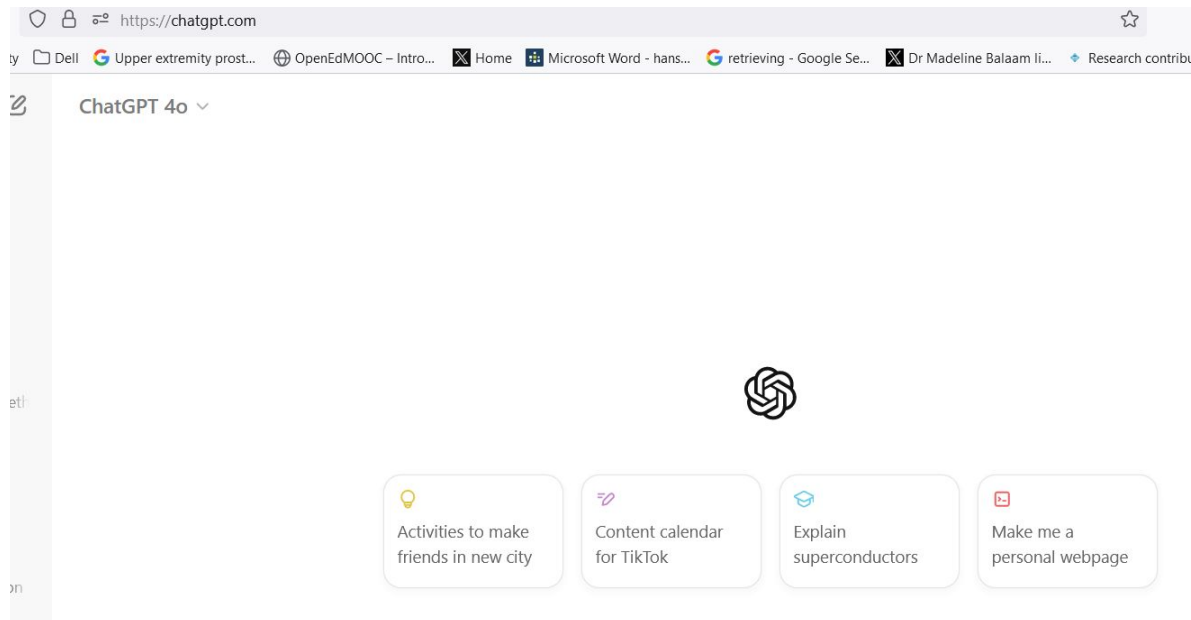
[My research is in the field of Human-Computer Interaction and Computational Social Science.](#)

[I primarily focus on social implications of AI - specifically "DeepFake/ AI" generated Misinformation and societal implications. Currently I lead few projects that bring AI/ Deepfakes literacy - explore and build technologies to assist human to make judgements.](#) My research is funded by

**A World Where Seeing
Is No Longer Believing!**

AI Generated Content

- Text
- Audio
- Video
- Images



What are they doing in Real World?

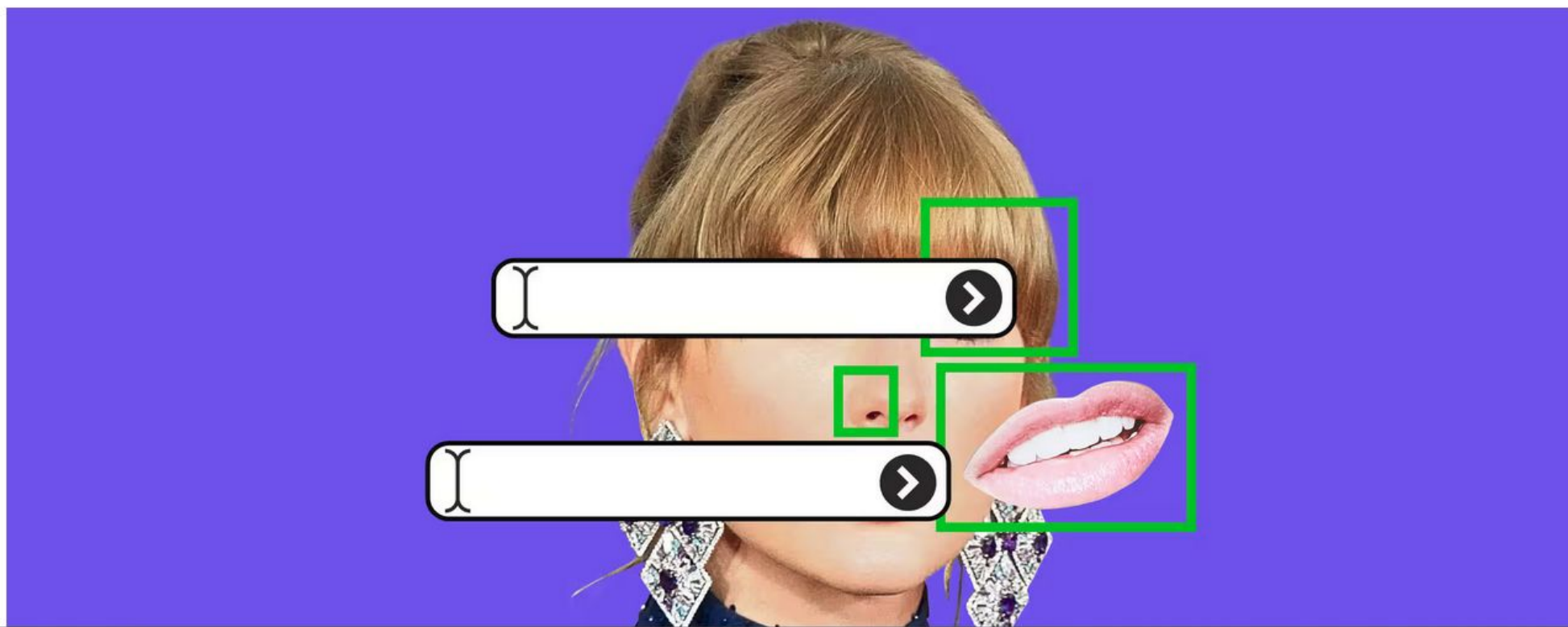
Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'



Emine Saner

Wed 31 Jan 2024 05.00 GMT

 **Share**



Ethics and Ownership

Get 12 weeks for \$29.99 \$6

THE
NEW YORKER

Newsletter

Sign In

Subscribe



FOOD AND COOKING The Joy of Blasphemous Fusion Food Learning to Love an Induction Stove The Hard-Won Pleasures of a Yeasted Cake Kitchen Notes

ANNALS OF GASTRONOMY

THE ETHICS OF A DEEPPFAKE ANTHONY BOURDAIN VOICE

The new documentary “Roadrunner” uses A.I.-generated audio without disclosing it to viewers. How should we feel about that?



By Helen Rosner

July 17, 2021



LinkedIn use of Deepfake images

TECHNOLOGY

That smiling LinkedIn profile face might be a computer-generated fake

March 27, 2022 · 7:00 AM ET

Heard on [Morning Edition](#)



SHANNON BOND



7-Minute Listen



<https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles>



Deepfake Sinan for content marketing scam

<https://twitter.com/sinanaral/status/1444433801267097605>



← Tweet



Sinan Aral (@sinanaral@mastodon.social) ✓
@sinanaral

THIS IS A SCAM. IF YOU WANT TO SEE A DEEPPFAKE OF ME USED TO SELL A STOCK TRADING ALGORITHM, CLICK HERE. BUT DO NOT GIVE YOUR PERSONAL INFORMATION OR FINANCIAL DETAILS AS THIS IS A SCAM

manifestgrowth.space/3p1gqS2p

4:17 AM · Oct 3, 2021

9 Retweets 14 Quote Tweets 62 Likes





Tweet your reply

Reply

 Search Twitter

Relevant people



Sinan Aral (@si... ✓
@sinanaral

Following

David Austin Professor @MIT;
Director @MIT_IDE; Founding Partner,
Manifest Capital; Author, The Hype
Machine, ex-Chief Scientist:
[@SocialAmp](#) [@Humin](#); In Stealth

Trends for you

Technology · Trending ...

#ArtificialIntelligence

12.4K Tweets

Entertainment · Trending ...

#AK62



World

Watch

World / Asia

Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'



By Heather Chen and [Kathleen Magramo](#), CNN

🕒 2 minute read · Published 2:31 AM EST, Sun February 4, 2024



Using deepfakes for unethical conduct

<https://www.whiskeyriff.com/2021/03/15/pennsylvania-mom-arrested-for-creating-deepfake-pictures-of-daughters-cheerleading-rivals-to-get-them-kicked-off-team/>

Pennsylvania Mom Arrested For Creating “Deepfake” Pictures Of Daughter’s Cheerleading Rivals To Get Them Kicked Off Team



Matt O'Leary ·

LIFE

VIDEOS

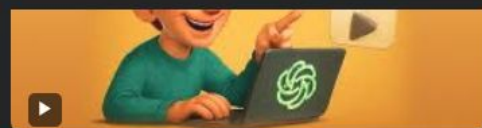
· March 15, 2021



WHISKEY RIFF SHOP



It is much *easy* to create



YouTube
Sora AI to create Stunning 3D Image ...



YouTube
Create Stunning AI Images with Real ...



CNET
How to Use Sora to Create an AI Video ...



The Knowledge Academy
How to use Sora OpenAI? A Step-by-Ste...



OpenAI
Sora | OpenAI



Art News
OpenAI's Sora Is Already Creating N...



YouTube
Create Cinematic AI Video in Sora (Full ...



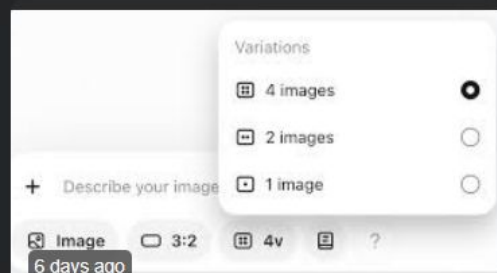
YouTube
Sora AI Tutorial — How to Create ...



OpenAI Help Center
Creating images on Sora | Open...



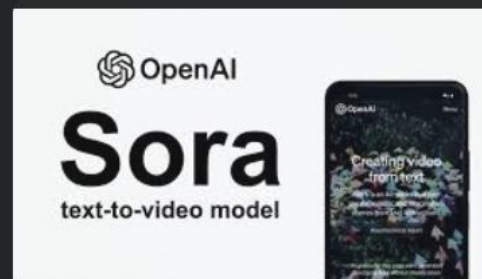
Tom's Guide
Create AI videos in Sora with these 5 ...



OpenAI Help Center
Creating images on Sora | OpenAI Help ...



upGrad
What is Sora AI? A Simple Guide for 2...



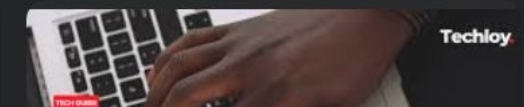
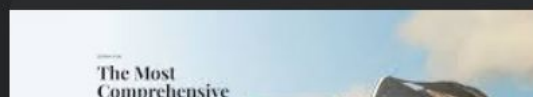
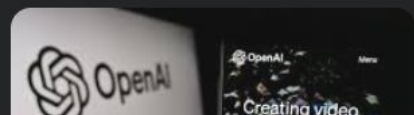
Great Learning
What is OpenAI's Sora? and How to Use ...



The Knowledge Academy
How to use Sora OpenAI? A Step-by-Ste...



Marketing-Interactive
text to video AI model ...





Sora 2 AI Video App is Terrifying



SAMTIME ✓
520K subscribers

Subscribe

7.6K



Share

Thanks



Online Deepfake Maker

Deepfake App to swap faces using AI.

Create a Deepfake Video

\$4 / hour

Average Cost and Time

Basic Deepfake (10,000 Iterations) - \$20

High-Quality Deepfake (50,000 Iterations) - \$80

Create a Deepfake Video

* The quality of your deepfake depends on the quality and length of your videos. The more well trained your AI model is, the better the result.

* Please use our technology responsibly and review [Our content policy](#) and our [terms of use](#).

Responsible Deepfake Technology



Visible Watermarks

Every deepfake generated on our tool has a clear and visible watermark indicated that the video is a deepfake. We also leave clear traces of manipulation in the video data so it's easy to identify as fake. We believe deepfake technology should be clearly labelled.



Imperfect by Design

Deepfake technology is incredibly advanced and can easily fool humans. We intentionally don't push the limits, so deepfake videos can be enjoyed while still being able to identify it's fake through the imperfections.



Accessibility

Deepfake technology has great potential in entertainment, gaming, satire and culture when used responsibly. Our system allows anyone to create a high quality deepfake at extremely low cost.

Deepfakes and Society

The Emergence of Deepfakes and its Societal Implications: A Systematic Review

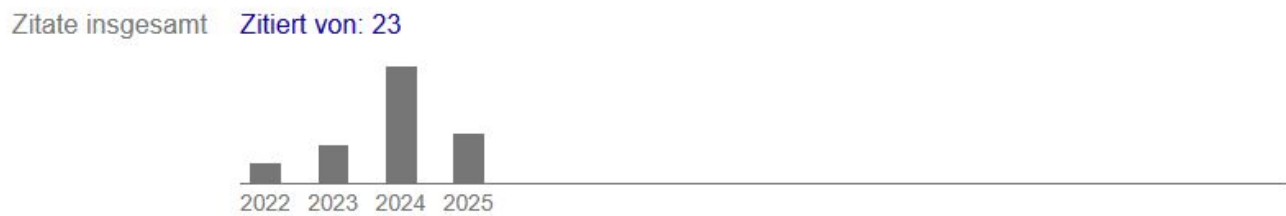
AutorenDilrukshi Gamage, Jiayu Chen, Kazutoshi Sasahara

Publikationsdatum2021

KonferenzConference for Truth and Trust (TTO) 2021 <https://truthandtrustonline.com/>

Beschreibung

The appearance of Deepfake tools and technologies in the public is proliferating. Scholarly research is very centered on technology of deepfake but sparse in understanding how the emergence of deepfakes impacts society. In this systematic review, we explored deepfake scholarly works that discuss societal implications than the technology-centered focus. We extracted studies from major publication databases- Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Springer Digital Library and Google Scholar. The corpus reflects patterns based on their research methodologies, area of focus, and the distribution of such research. Out of 787 works, 88 were highly relevant, with the majority of the studies being reviews of the literature. While research focus is generally drawn upon exploring security related harms, less focus is put on issues such as ethical implications and legal regularities for areas other than pornography, psychological safety, cybercrimes, terrorism, and more. The field research for Deepfake social impact research is emerging and this paper brings more insights drawn from a methodical, subject focused and distribution point of view.



[PDF] von truthandtrustonline.com

SPRINGER NATURE Link

Log in

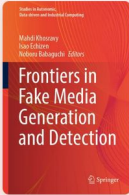
Find a journal Publish with us Track your research Search

Cart

Home > [Frontiers in Fake Media Generation and Detection](#) > Chapter

Deepfakes and Society: What Lies Ahead?

Chapter | First Online: 29 May 2022
pp 3–43 | [Cite this chapter](#)



Frontiers in Fake Media Generation and Detection

Dilrukshi Gamage , Jiayu Chen, Piyush Ghasiya & Kazutoshi Sasahara

Access this chapter

AI Safety Report

www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

ell Upper extremity prost... OpenEdMOOC – Intro... Home Microsoft Word - hans... retrieving - Google Se... Dr Madeline Balaam li... Rese



Menu

[Home](#) > [Business and industry](#) > [Science and innovation](#) > [Artificial intelligence](#)

Research and analysis

International Scientific Report on the Safety of Advanced AI: interim report

An interim report on the science of advanced Artificial Intelligence (AI) safety.

From: [Department for Science, Innovation and Technology](#) and [AI Safety Institute](#)

Published 17 May 2024



Yoshua Bengio • Following

Full professor at Université de Montréal, President and Scientific Director of Law...
5mo •

Today, we are publishing the first-ever International AI Safety Report, backed by 30 countries and the OECD, UN, and EU. It summarises the state of the science on AI capabilities and risks, and how to mitigate those risks.

This unprecedented, large-scale effort involved 100 independent AI experts from around the world including Nobel laureates and Turing Award winners. These contributors represent different perspectives within the AI community and disagree on several questions. But, for this Report, they all worked together to produce an authoritative document on the state of the science. As the Chair of the Report, I am deeply impressed by their work.

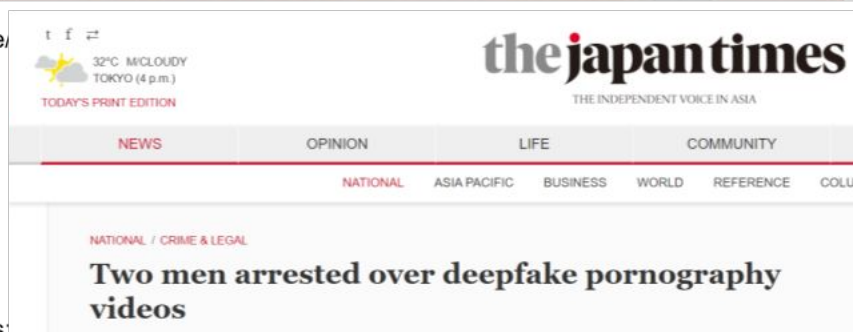
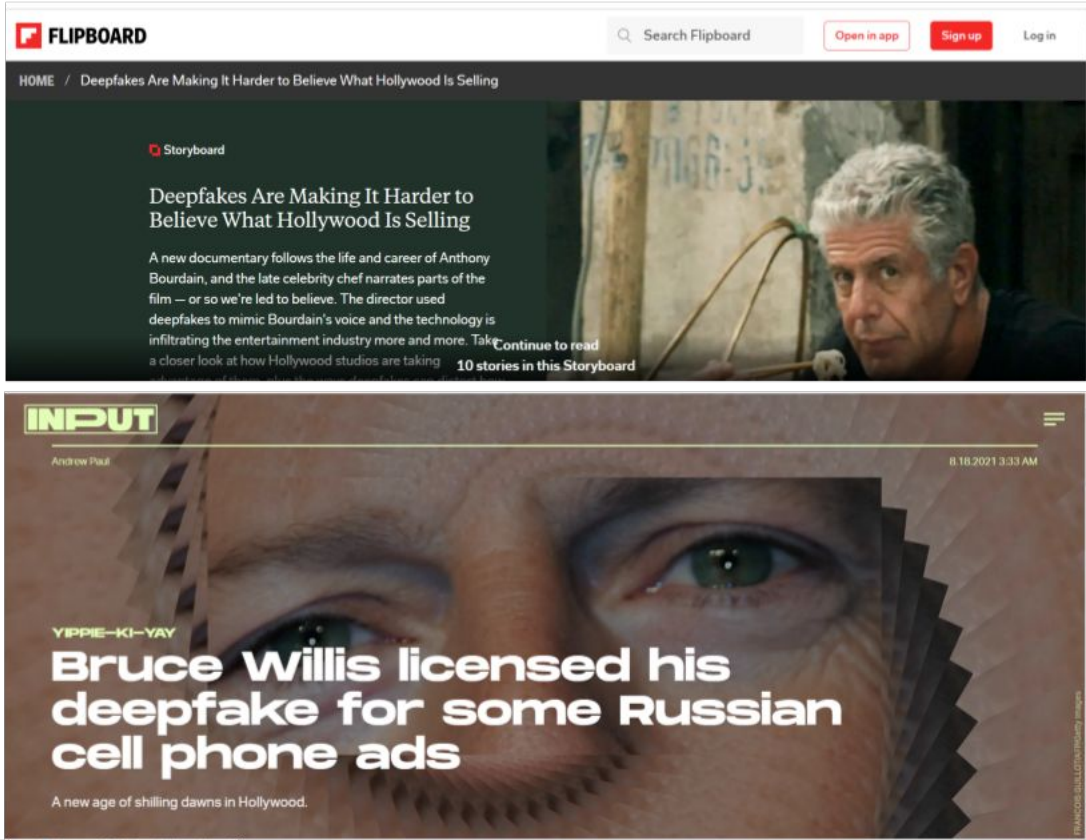
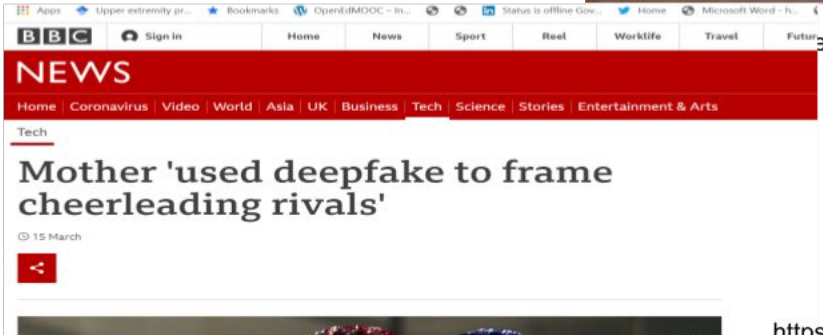
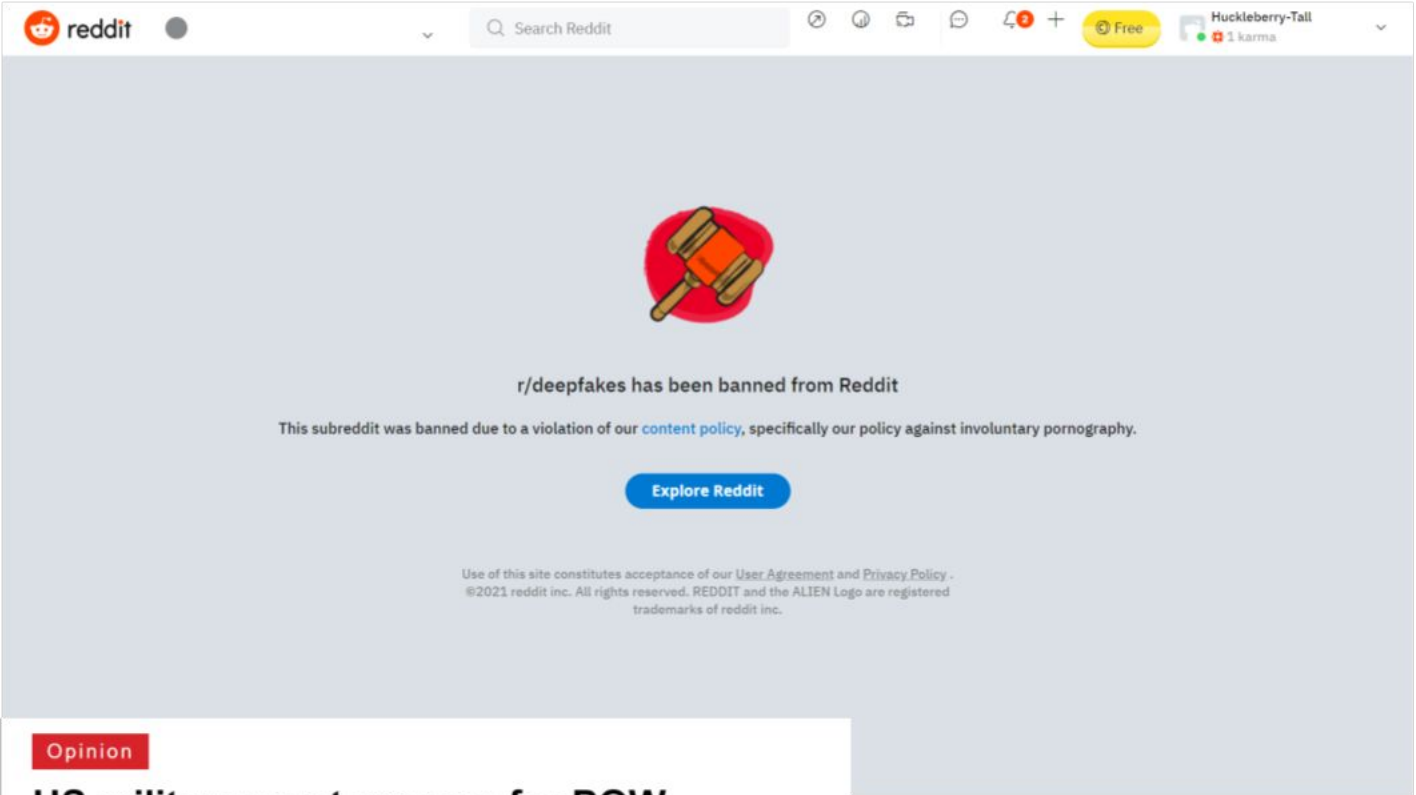
A central conclusion of the Report is that even the short-term future of general-purpose AI is remarkably uncertain. Both very positive and very negative outcomes are possible. Therefore, much depends on how societies and governments act.

With all the noise around AI, this Report aims to provide policymakers with an evidence-based, balanced overview of AI risks and mitigations. The remarkably collaborative spirit of all contributors gives me hope. It shows how scientific disagreement can be highly productive, and I want to thank each one of them for their work. I am also grateful to the industry and civil society organisations who contributed invaluable feedback.

I look forward to the important discussions that will unfold during the AI Action Summit in Paris on February 10-11 and in various other forums in the coming months.

Link to full Report: <https://lnkd.in/eJijzBsS>

Real incidents - CHI2022



Home > Conferences > CHI > Proceedings > CHI '22 > Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications

RESEARCH-ARTICLE OPEN ACCESS

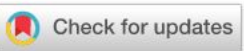


Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications

Authors: [Dilrukshi Gamage](#), [Piyush Ghasiya](#), [Vamshi Bonagiri](#), [Mark E. Whiting](#), [Kazutoshi Sasahara](#)
[Authors Info & Claims](#)

CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems • April 2022 • Article No.: 103 • Pages 1–19 • <https://doi.org/10.1145/3491102.3517446>

Published: 28 April 2022 [Publication History](#)



1 1,645



Exploring social implications of deepfake conversations in Reddit

RQ1, What type of conversations lead in Reddit communities about deepfakes and how do these conversations change over time?

RQ2: What are the societal implications visible in these conversations?

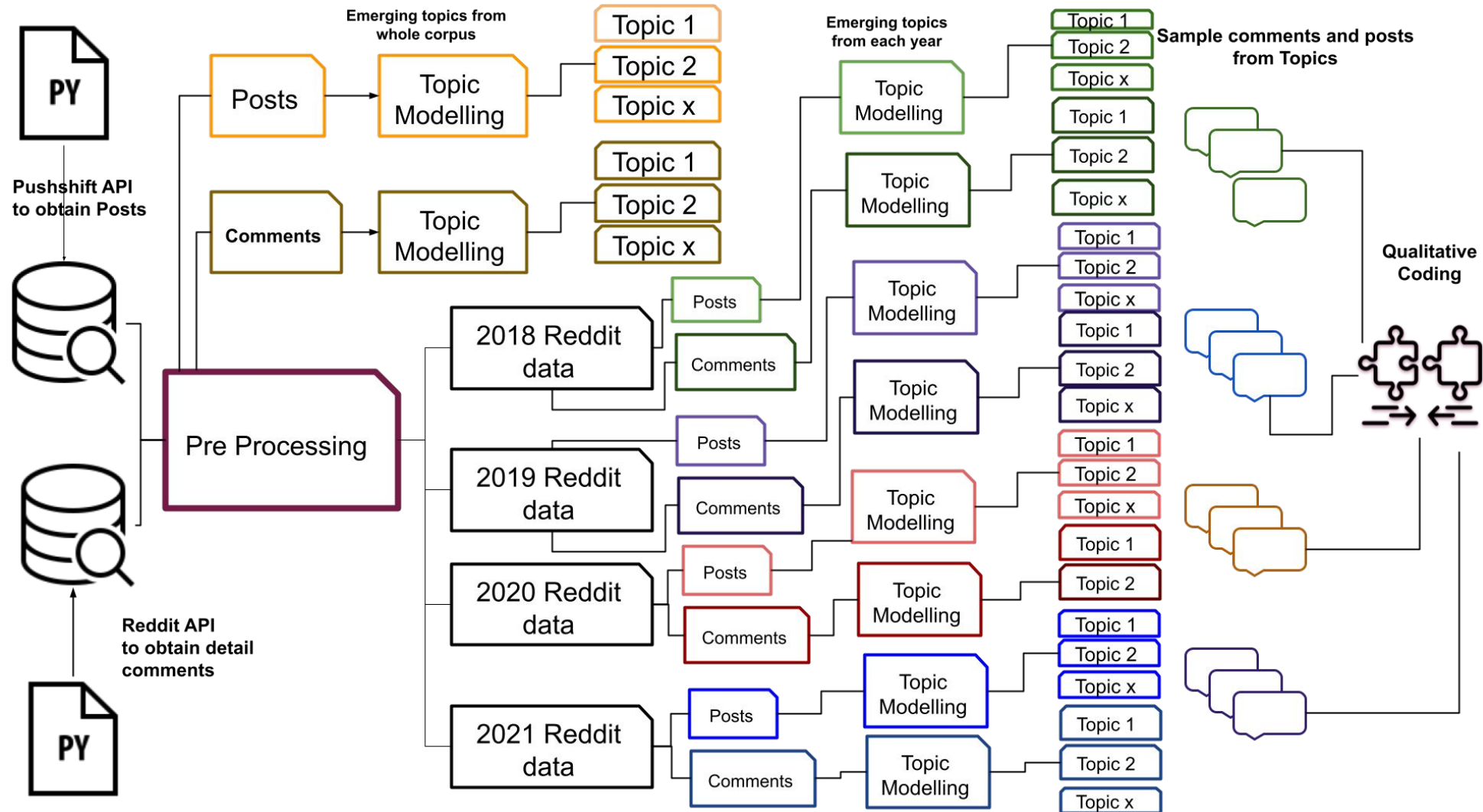
The screenshot shows a web browser at reddit.com/search/?q=deepfake. The search results are filtered by 'Best results'. Under the 'COMMUNITIES AND USERS' section, three communities are listed: u/DeepFake (0 Karma), r/SFWdeepfakes (15.8k Members), and r/deepfakesfw (2.2k Members). To the right of these listings is a text box with the following content: 'When most people hear the term "deepfake" they automatically think of something NSFW. Let's adjust that stigma. This subreddit is dedicated to SFW machine learning and other DF resources, including audio deepfakes.' Below this text is another paragraph: 'A secondary deep fakes subreddit that will be 100% Safe for Work. This is for people who find the technology interesting, and would like to experiment with it, but without the NSFW stuff'. There are 'Follow', 'Joined', and 'Join' buttons next to the respective community listings. On the right side of the page, there is a 'Create Community' button and a 'Have an idea for a new community?' prompt. At the bottom, there are links for 'Help', 'Reddit Coins', 'About', and 'Careers'.

12MB Data

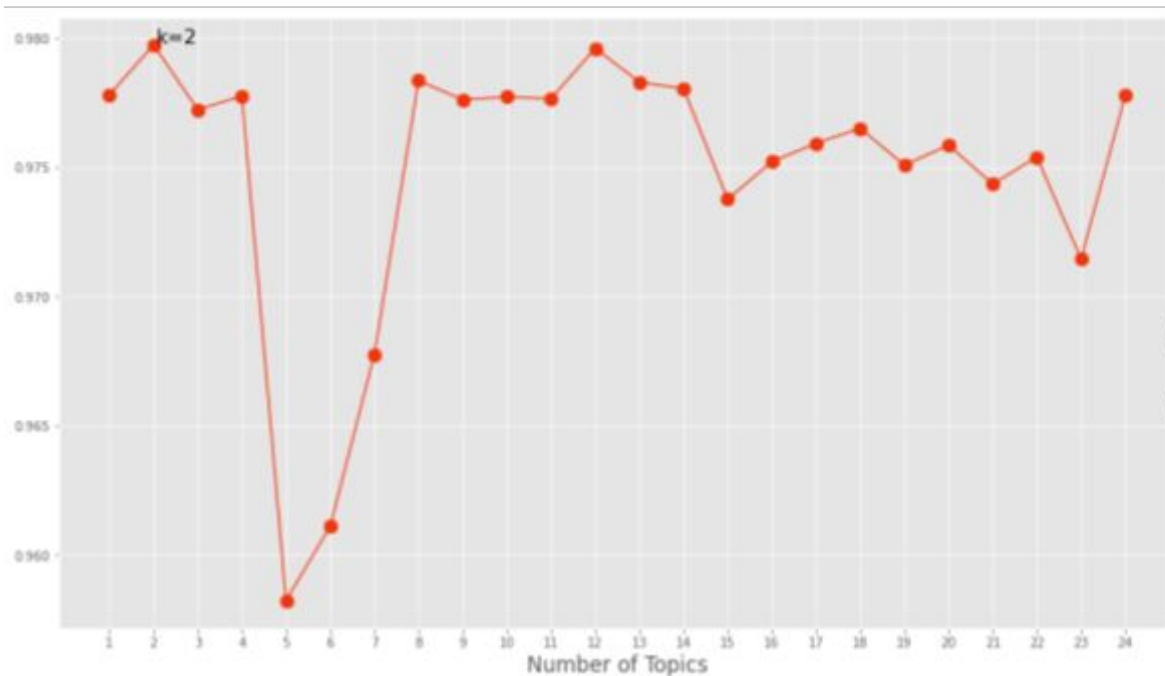
86,425 Comments

between 2018-2021

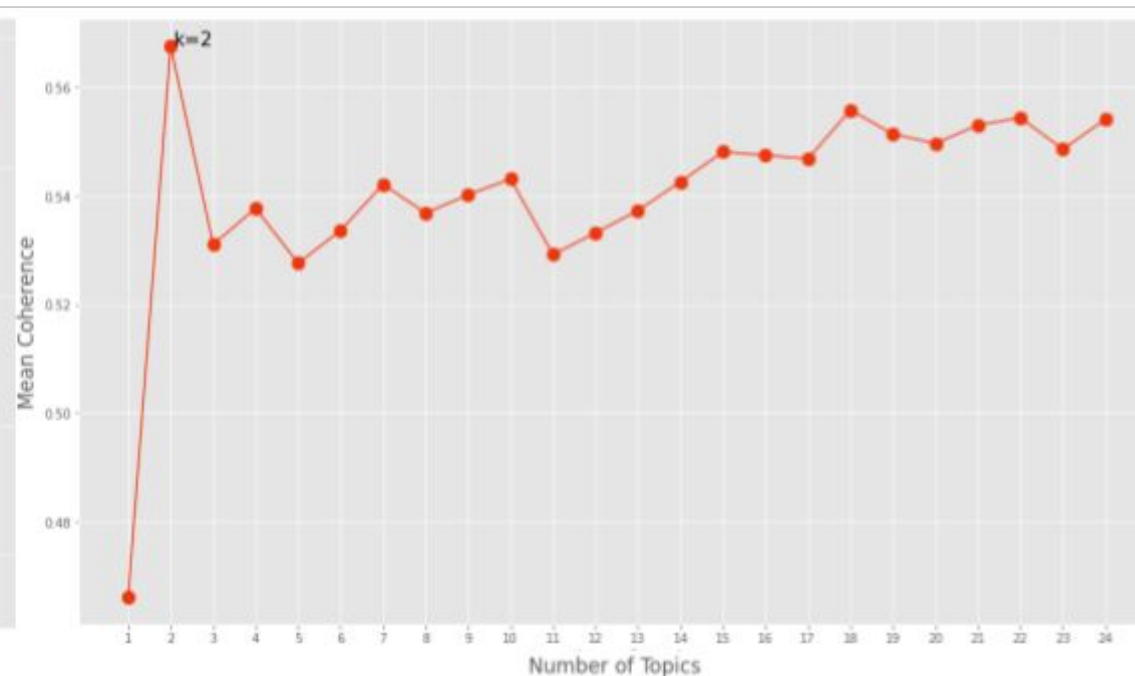
Research Design



Topic modeling to understand the discussed topics



Mean coherence of Posts



Mean coherence of Commentes

Topic Modeling for the entire corpus (Data from 2018 to 2021)

A sample topics generated from comments in 2021

Topic 01: people, want, real, deepfakes, believe, porn, need, mean, lot, things
Topic 02: https, check, link, channel, article, want, source, github, free, discord
Topic 03: fake, deep, real, tell, trump, news, detect, videos, porn, technology
Topic 04: please, subreddit, automatically, question, moderators, bot, contact, concern, action, perform
Topic 05: look, real, exactly, luke, bad, weird, kinda, hair, kind, something
Topic 06: deepfake, porn, someone, technology, tech, videos, image, need, software, de
Topic 07: think, thing, something, guy, saw, yeah, mean, right, sure, way
Topic 08: see, best, love, want, thing, internet, wait, sub, something, great
Topic 09: know, yeah, mean, tell, guy, talk, everyone, someone, old, right
Topic 10: good, pretty, job, thing, yeah, man, point, bad, love, damn
Topic 11: video, original, watch, real, edit, trump, youtube, link, game, frame
Topic 12: get, ta, guy, try, probably, help, away, ban, little, old
Topic 13: post, comment, remove, sub, link, read, rule, twitter, thank, reddit
Topic 14: face, voice, body, put, actor, someone, head, ai, original, actors
Topic 15: really, cool, great, hope, sound, scary, wow, love, hard, man
Topic 16: go, time, back, take, need, way, lol, happen, watch, long
Topic 17: fuck, na, shit, gon, holy, man, wan, oh, sick, dude
Topic 18: better, work, great, train, luke, way, deepfakes, job, mean, technology

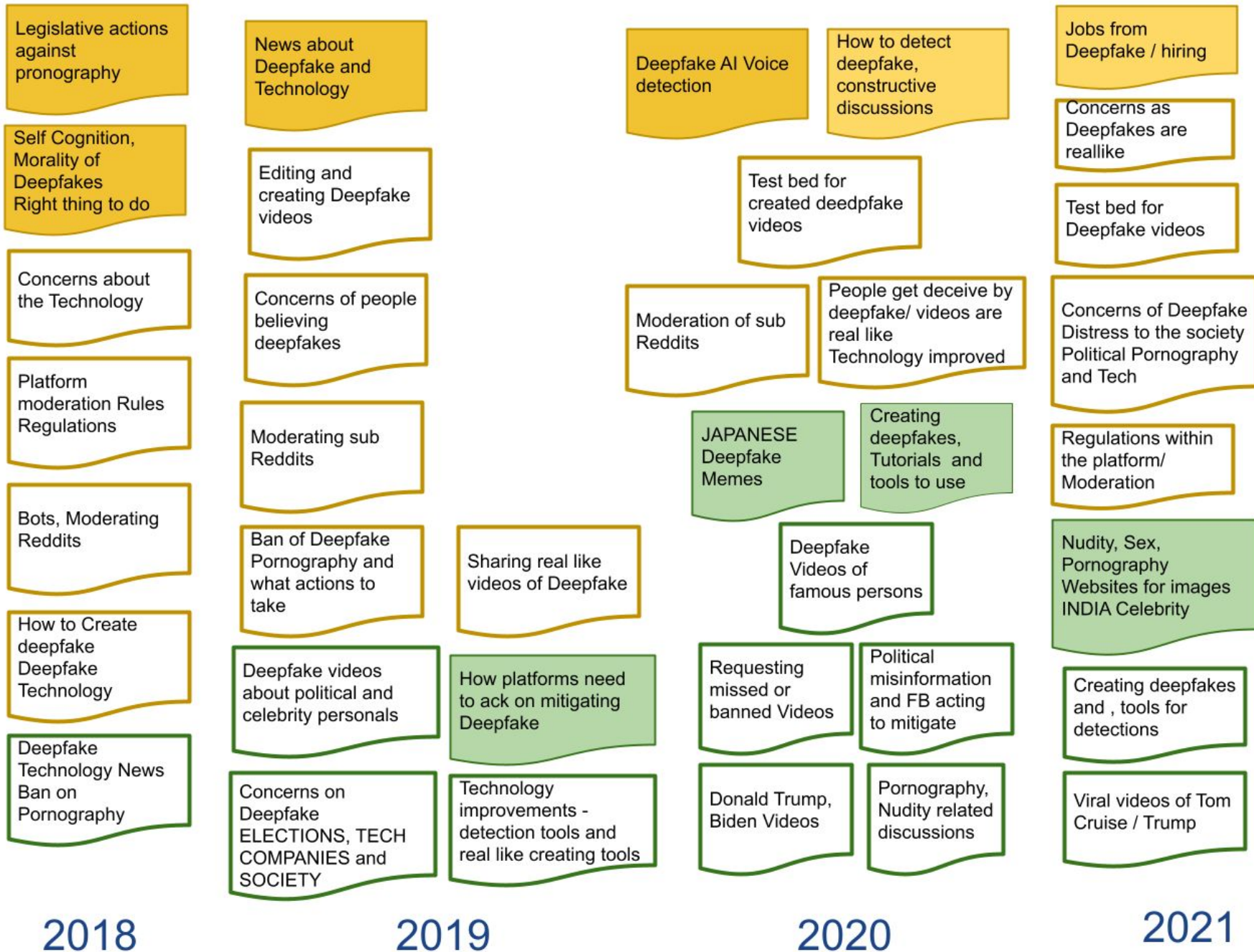
RQ1:

What type of conversations lead in Reddit communities about deepfakes and how do these conversations change over time?

Summary of Topics

Comments

Posts



Implications of deepfake discussions in Reddit

1. Reddit space becomes an on demand archive for banned deepfake videos
2. Reddit space incubating deepfake creators
3. Reddit becomes a collective space to question and raise concerns about deepfakes
4. Reddit expert users support anyone to create customized pornography
5. Reddit users provide inspirations for deepfake movies and knowledge on surrounding legal clearances

Implications of deepfake discussions in Reddit

6. Reddit users ingenuity to mix variety of sources with a variety of targets in creating deepfakes
7. Reddit platform and other social media platforms disproportionately control their deepfake policies
8. Reddit community is more concerned with deepfakes videos but less concerned with the deceiving power of AI-generated text
9. Reddit and other social media platforms influence elections with deepfakes
10. Reddit users create a deepfake marketplace

Implications to Human Computer Interactions

Theoretical Implications	Design Implications
Comparing Morality on Creating / Sharing Deepfakes Morality on deepfake marketplace Reddit space as incubator for deepfakes - ethical considerations	Ethical reviews to deepfake tools Deepfakes detections and education about harmful implications Deepfake reporting mechanisms to the platform

Regulatory acts and interventions to safeguard the potential harm



Labeling requirements for artificial or manipulated content (deepfakes).

Whenever deepfakes are disseminated, platforms must now indicate as such.
Those building deepfake creation and generative AI tools must assist also in their labelling.

But

What if they did not label? What technologies to detect?

Future work

- Understanding interventions that can reduce deepfake sharing behavior
- Understanding the impact of deepfakes - from victims perspectives
- Effect of Labeling deepfake content - Types of labels for Deepfakes

“Bringing global perspectives to AI research and practice”

< 15%

of the world's population is WEIRD — Western, Educated, Industrialized, Rich, Democratic — yet today's AI systems are rooted in exactly that narrow set of data, languages, and institutional contexts.

This talk looks at

AI & Learning

AI & Democracy

AI & Environment

AI & Fairness

AI & Humanities

AI & Governance

Synthetic media is where democracy, fairness, and governance collide — and where the WEIRD defaults of today's defenses fail first.

Everyone can fabricate. The defenses were built for someone else.

01

Trivially easy to fake

Convincing images, audio, and video can now be generated on a low-cost smartphone — production has been democratized faster than any defense.

02

Defenses face West

Warning labels, detection tools, and literacy curricula are designed for Western platforms, languages, and formal learners — and exported as-is.

03

Harms fall unevenly

Where fact-checking infrastructure is thin, harms concentrate on women, minorities, and marginalized communities — with the fewest routes to redress.

The Global South is home to the majority of the world's internet users — and largely invisible in synthetic-media research and policy.

South Asia is already living with it

Sri Lanka

AI-generated content impersonating the President, national cricketers, and the Central Bank Governor — deployed for financial fraud and reputational destruction, in a post-conflict environment where one video can ignite communal violence.

India

The 2024 Lok Sabha cycle: voice-cloned candidate speeches by linguistic community, fabricated celebrity endorsements, synthetic 'doctors' pushing dangerous medical advice through WhatsApp forward chains.

Bangladesh

A deepfake of ousted PM Sheikh Hasina 'confessing' to the aynaghor detention centers circulated widely — verified as AI-manipulated by fact-checkers, credited with real political turmoil.

Across all three: synthetic pornography systematically weaponized against women in public life.

One quarter of the world, underserved

103

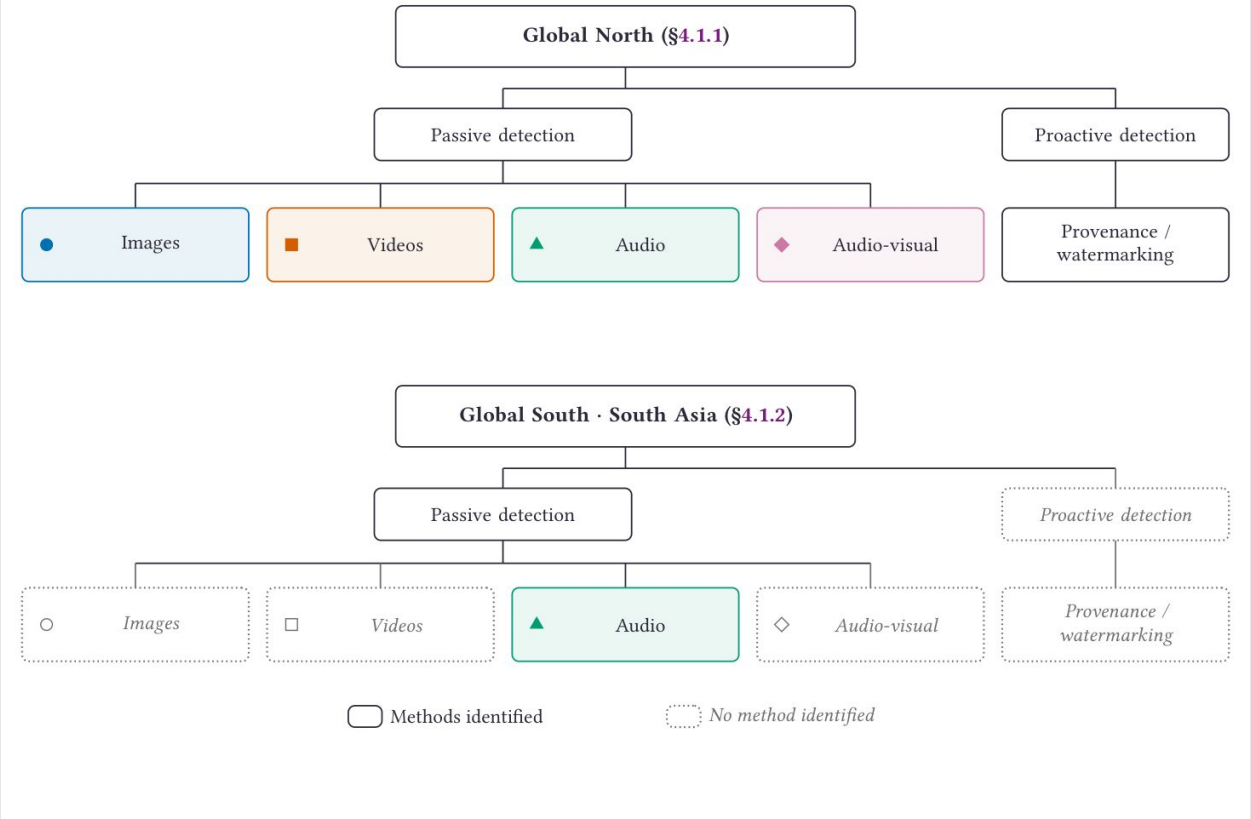
studies reviewed, 2018–2026 — detection methods, datasets & benchmarks, surveys, and empirical studies

4

detection models built specifically for South Asian content — all for synthetic speech, published only in 2024–25

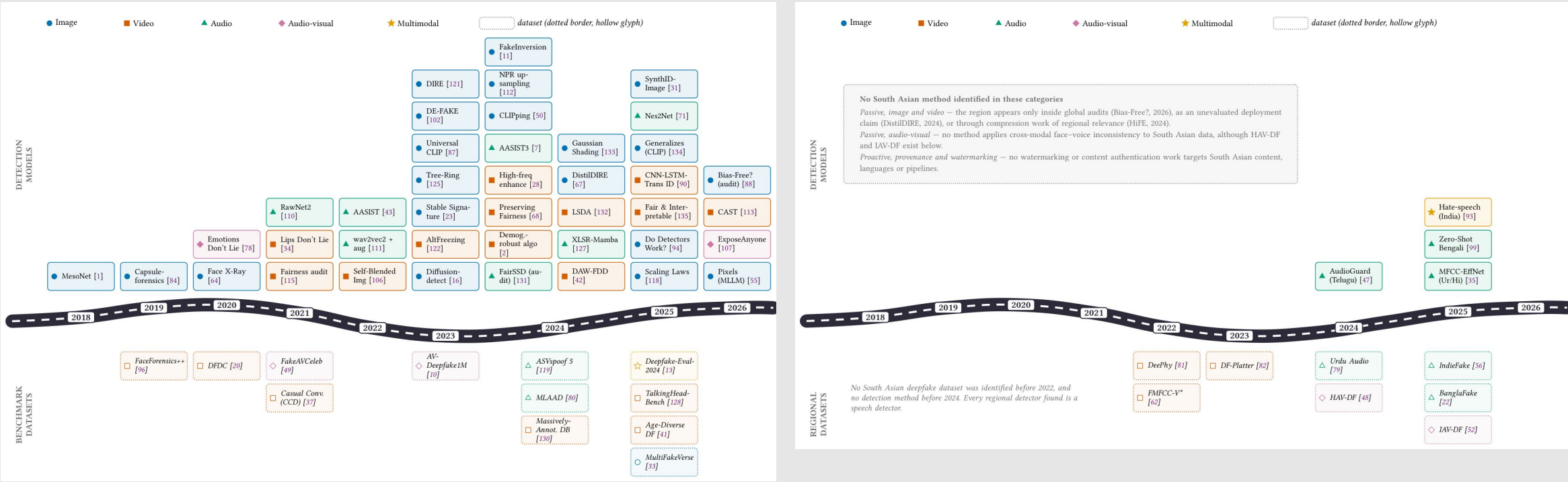
0

South Asian methods for image, video, or audio-visual deepfakes — and zero provenance or watermarking work targeting the region's content, languages, or pipelines



Our systematic review mapped the same paradigm × modality space for the Global North (top) and South Asia (bottom). Solid boxes: methods exist. Dotted grey: no method identified — for South Asia, everything except audio.

Eight years of research — side by side

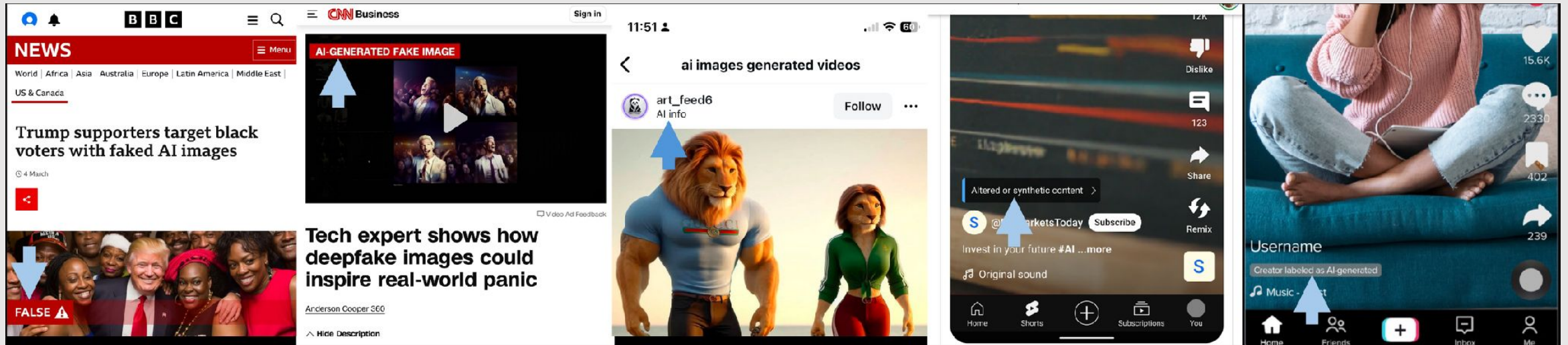


The gap is not a lag. It is a structural absence.

Five threads of evidence, one argument

01	The detection gap	A systematic review of deepfake detection for South Asia	103 studies · 2018–2026
02	Labels	How users perceive warning-label designs for AI-generated content	CHI 2025 · N = 911
03	Literacy	Deepfake.help — an inoculation-based literacy tool, field-tested in Sri Lanka	Field experiment · N = 100
04	Communities	How stakeholders and online communities experience, discuss, and contest deepfakes	3 countries · 49 stakeholders
05	Provenance & policy	Do C2PA, SynthID, and the EU AI Act translate to low-resource contexts?	Synthesis

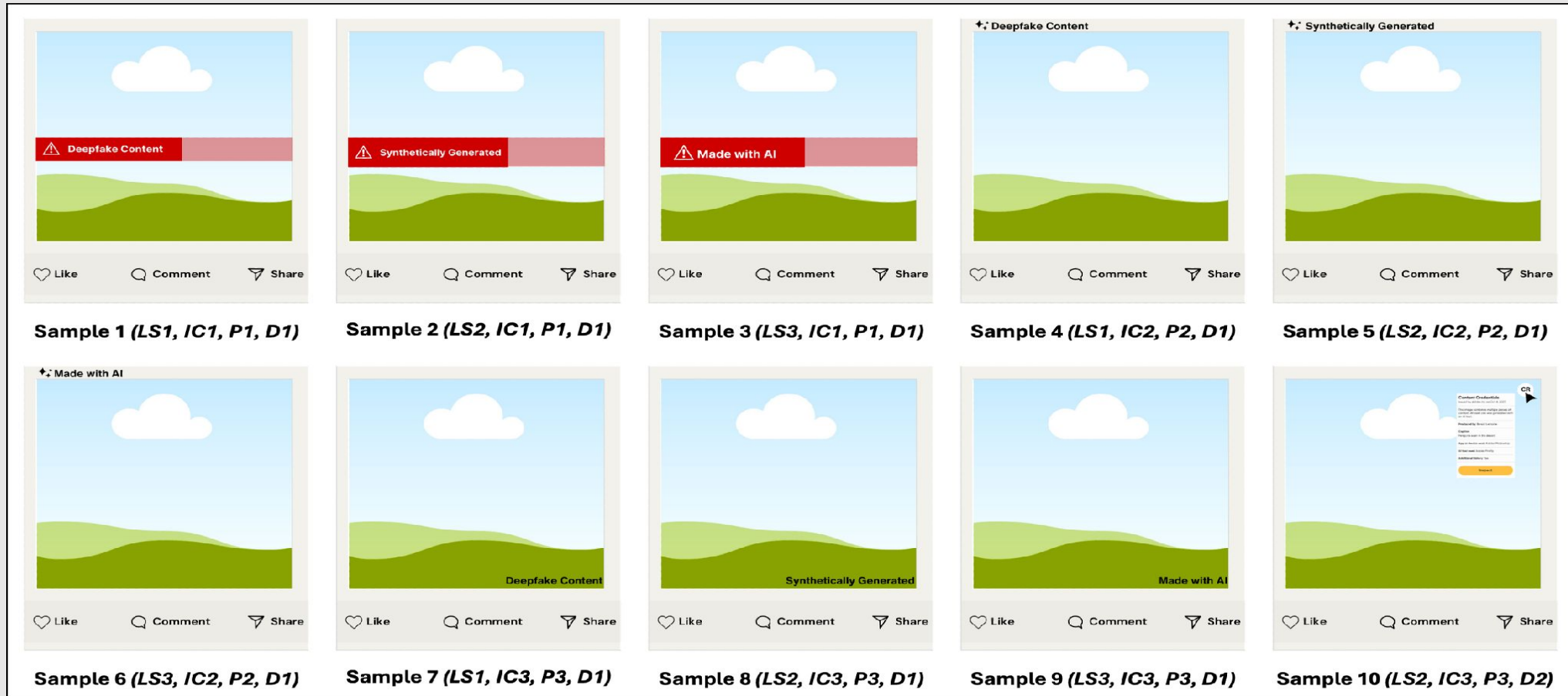
Platforms are already labeling. Does anyone benefit?



Label designs in the wild, from our CHI 2025 review — BBC News ('FALSE'), CNN ('AI-GENERATED FAKE IMAGE'), Meta ('AI info'), YouTube ('Altered or synthetic content'), TikTok ('Creator labeled as AI-generated'). Same goal, wildly different tone, iconography, placement, and detail.

We mapped this design space along four dimensions — sentiment, icon/color, position, level of detail — and derived ten prototype labels.

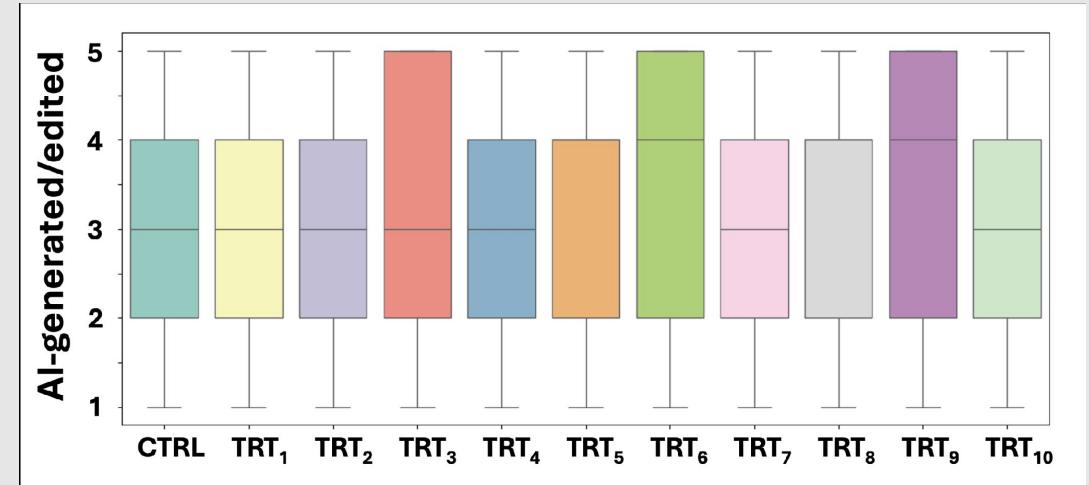
Ten label designs, one experiment — 911 users



The ten samples systematically vary label sentiment (hazardous 'Deepfake Content' → neutral 'Made with AI'), icon and color (red warning → monochrome AI sparkle → none), position (obscuring, above, or non-obscuring), and detail (one-line text → Sample 10's Content Credentials panel).

Labels shift belief — not behavior

- **Belief moves** — Every label design increased users' belief that content was AI-generated or edited. 'Made with AI' phrasing communicated best; 'deepfake' wording was ambiguous to many.
- **Trust depends on design** — Trust in the label varied significantly by design. The Content Credentials-style label (Sample 10) earned the highest trust — and label trust correlates with platform trust ($r = .73$).
- **Engagement unchanged** — Liking, commenting, and sharing did not significantly change with a label present. Labels inform; they do not restrain.

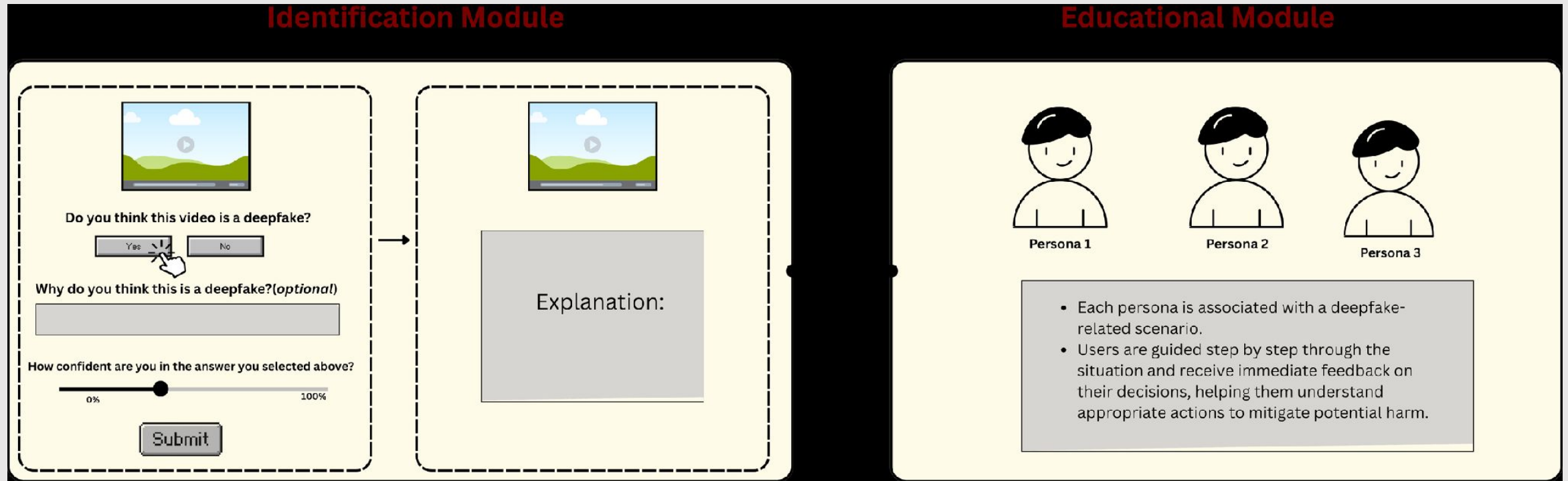


Belief that content is AI-generated (1–5), control vs. the ten label designs: every labeled condition sits above the unlabeled control.

Design implication: neutral, AI-mentioning labels that don't obscure content are trusted and understood — but disclosure alone will not stop the spread.

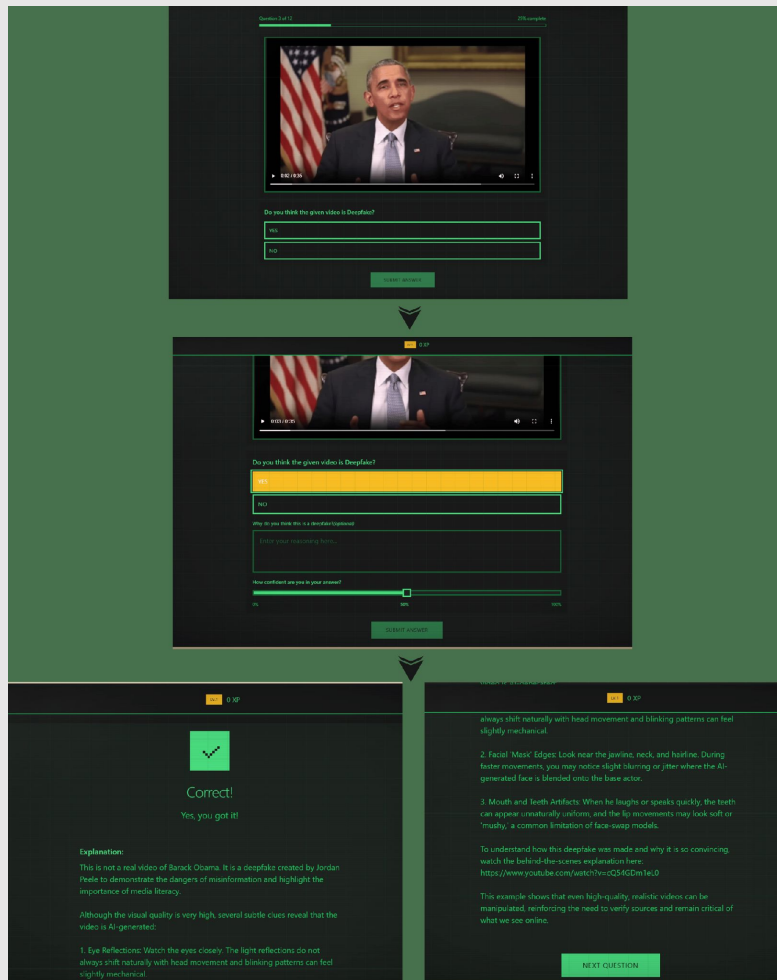
Deepfake.help — literacy as inoculation, not lecture

Inoculation theory — exposure to weakened manipulative content, with refutation · **Reinforcement learning as pedagogy** — iterative judgment with immediate corrective feedback · **Critical-thinking scaffolding** — from detection to socially situated action

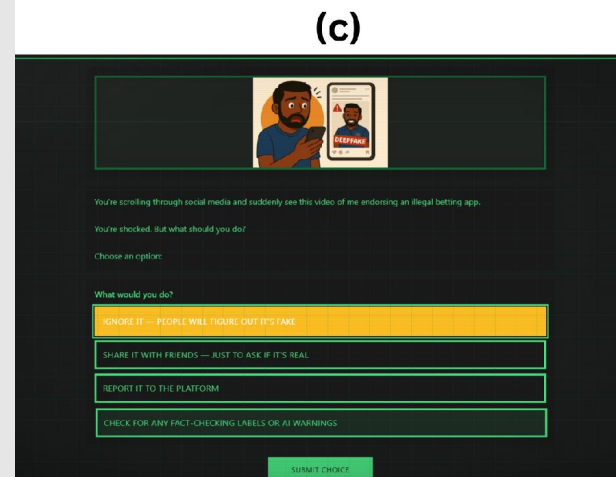
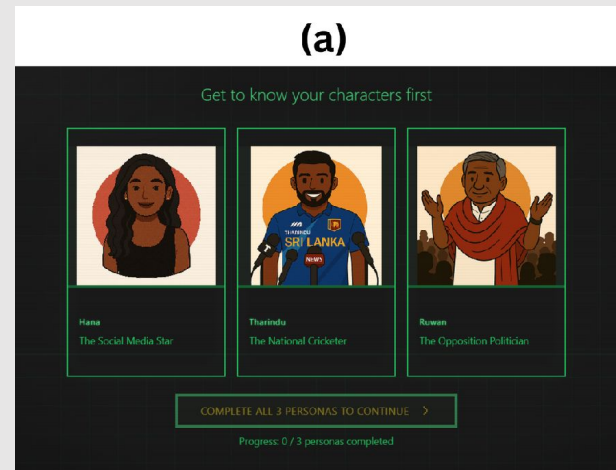


The two modules of Deepfake.help: an Identification module (watch a video, judge it, justify the call, calibrate confidence, get cue-level feedback) feeding an Educational module (persona-based harm scenarios with guided, step-by-step response choices).

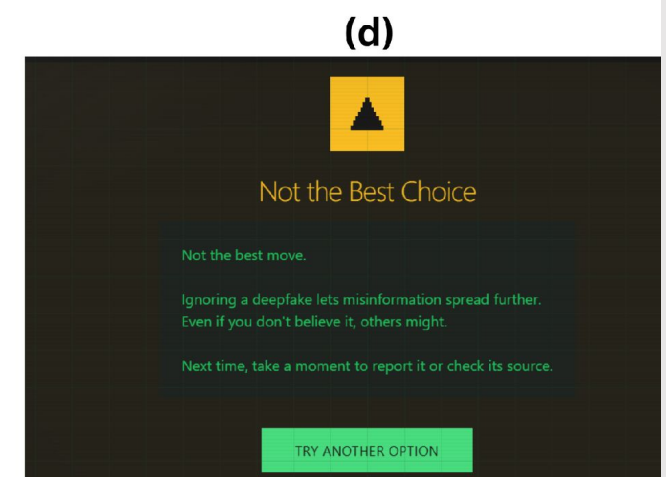
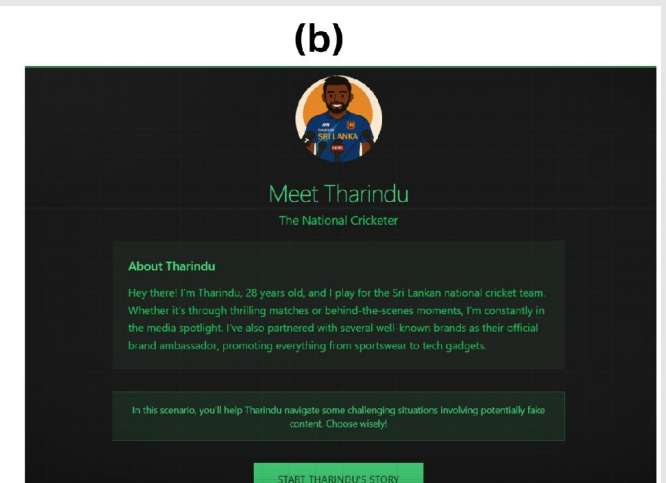
Judge, get corrected, then act — inside culturally familiar scenarios



Identification module: classify → justify → calibrate confidence → immediate feedback naming the cues (eye reflections, facial 'mask' edges, mouth artifacts).



Educational module: three personas mirroring real South Asian incidents — a social-media star facing non-consensual imagery, a national cricketer exploited for a betting app, a politician's voice cloned in an election — each with step-by-step guidance on responsible action (report, warn, don't forward).

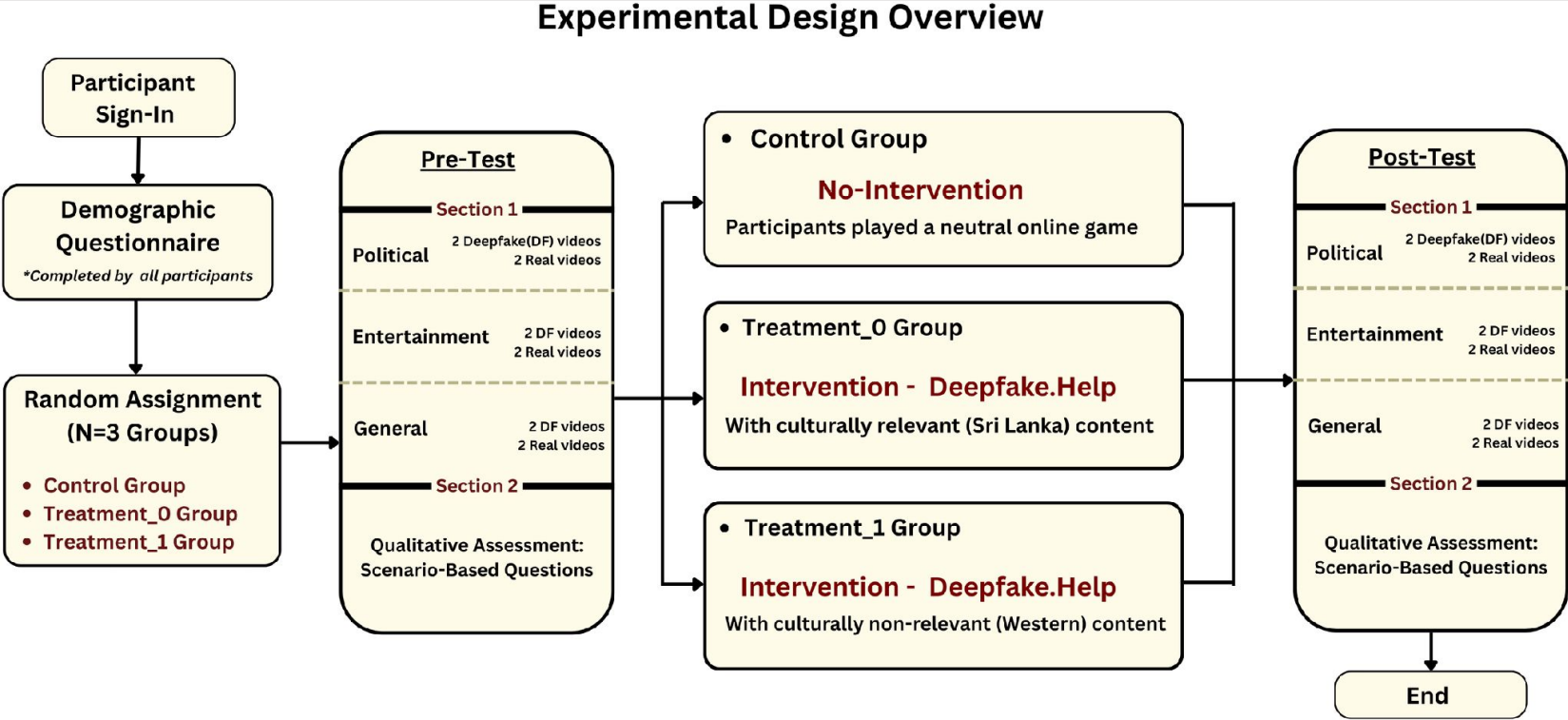


A 30-day field experiment in Sri Lanka

Control no intervention (Tetris)

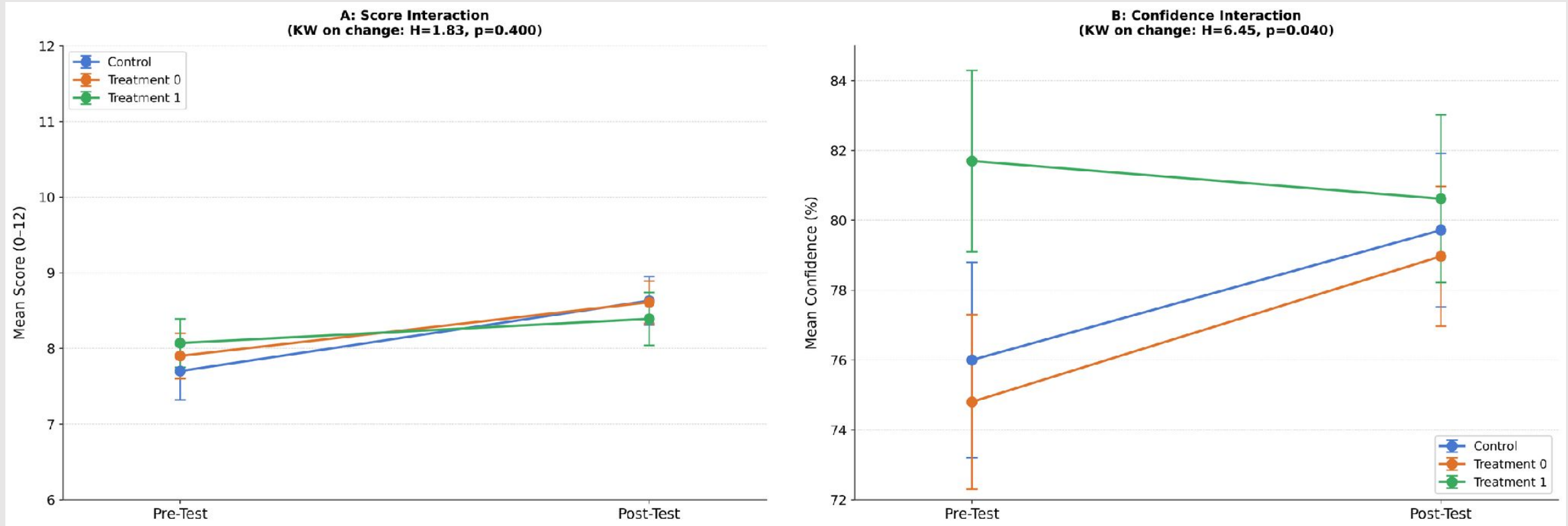
Treatment 0 Deepfake.help, culturally localized (Sri Lankan) content

Treatment 1 Deepfake.help, Western content



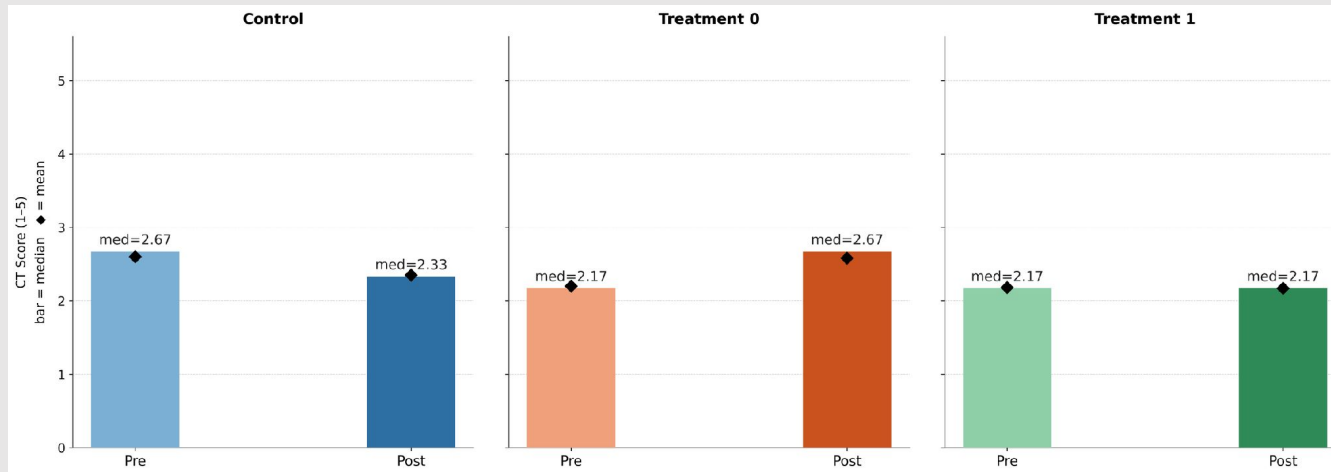
N = 100 · pre/post: 12-video detection test with confidence ratings, plus three open-ended harm scenarios scored on a SIFT-based

Localized content produced the real gains — in calibration, not raw detection

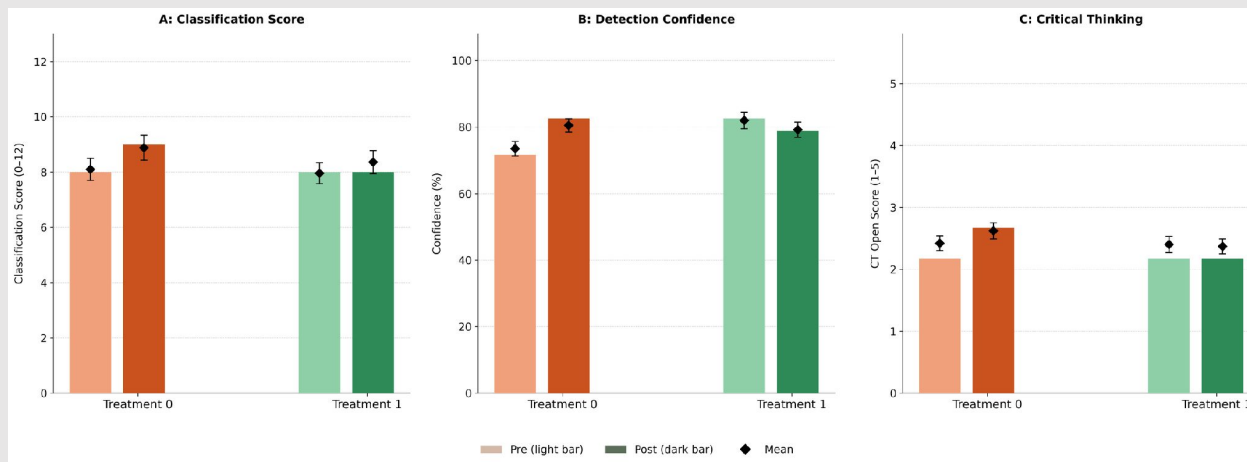


Pre→post change across the three groups. Left: classification accuracy — all groups improve modestly, no significant between-group difference (human perceptual detection hovers near chance; ceilings are real). Right: detection confidence — the 'crossing' pattern. Localized content: 71.7% → 82.6% ($p < .001$); Western content: flat-to-declining from a high, poorly calibrated baseline.

The intervention protected critical reasoning that otherwise decayed



SIFT-scored critical-thinking behavior, pre vs. post, by condition.



Localized vs. Western content, post test: same accuracy, different minds.
 Sri Lanka vs. Western content, post test: same accuracy, different minds.
 DP Literacy / DeepPrakr.HELP / Gamage et al., PACM HC (USENIX), 2020

Control declined

Without scaffolding, open-ended reasoning scores dropped over 30 days (2.67 → 2.33, $p = .036$).

Treatments held

Both intervention groups sustained — and the localized group improved — their reasoning (change across groups $p = .007$).

Culture moderates

Localized vs. Western content: same detection accuracy, significantly stronger confidence ($p = .016$) and critical-thinking gains ($p = .021$).

Asking those who live it: 49 stakeholders, three countries

Participatory workshops with journalists, fact-checkers, lawyers, feminist advocates, educators, and civil society — communities as co-producers of knowledge, not populations to be studied.

Table 1: Overview of Participatory Workshops			
Country	Participants (n)	Stakeholder Groups Represented	Date
Sri Lanka	15	Journalists, NGOs, educators, technologists	April 2025
India	22	Fact-checkers, lawyers, academics, activists	August 2025
Bangladesh	12	Lawyers, media activists, students, feminists	September 2025

Three online workshops (April–September 2025), facilitated in Sinhala, Tamil, Hindi, Bengali, and English; every group combined the people who detect harm, litigate it, are targeted by it, and educate against it.

Spot the Fake → Experience Sharing → Scenario Probing → Co-Design → Collective Prioritization

The harm landscape is gendered, linguistic, and institutional

Table 3: Cross-Country Comparison of Key Harm Themes Identified in Workshops

Theme	Sri Lanka	India	Bangladesh
Gendered harms / non-consensual synthetic media	✓	✓	✓✓
Political misinformation & electoral manipulation	✓✓	✓✓	✓
Health misinformation via AI-generated content	—	✓✓	✓
Financial scams / celebrity endorsement fraud	✓	✓✓	✓
Low digital literacy among vulnerable populations	✓	✓✓	✓✓
Weak / absent national legal frameworks	✓	✓	✓✓
Platform under-moderation in local languages	✓✓	✓✓	✓
Psychological harm to victims (trauma, anxiety)	✓	✓	✓✓
Ethnic / communal tensions amplified by deepfakes	✓✓	✓	—
Distrust in AI tools & data privacy concerns	—	✓	✓✓

✓✓ = primary theme; ✓ = present; — = absent or marginal

Gendered

Non-consensual synthetic media is routine, not exceptional — and intersectional: queer, indigenous, and rural victims have the fewest routes to redress.

Linguistic

Verified synthetic content in Sinhala, Tamil, Hindi, Bengali stays up for weeks; comparable English content comes down in hours.

Institutional

Victims face absent laws, unreachable platforms, and courts that offer no practical recourse.

Cross-country comparison of harm themes surfaced in the workshops (✓✓ primary, ✓ present, — marginal): gendered harms, weak legal frameworks, and platform under-moderation of local languages cut across all three countries.

Communities did not ask for better detectors

Legal frameworks that treat synthetic abuse as a rights violation

Deepfake-specific electoral law (Sri Lanka), enforceable IT Rules (India), a new cyber law modeled on the EU AI Act with gender-based-abuse provisions (Bangladesh).

Community-governed verification in local languages

Locally built detection, volunteer-sourced databases, fact-checker coalitions — technology serving human judgment, not replacing it.

Literacy embedded in community practice

Gamified, narrative tools built on culturally resonant scenarios — for rural users, feminist networks, and students without devices.

Survivor-centered support infrastructure

Safe encrypted reporting, peer counseling, legal-aid networks — victim support is a design challenge as urgent as detection.

In every domain: solutions developed with affected communities, not for them.

C2PA, SynthID, the EU AI Act — lost in translation?

Provenance standards

C2PA / Content Credentials won the highest trust of any label we tested — yet users told us they didn't understand what it meant. Metadata is stripped by the messaging apps where Global South content actually travels.

Detection & watermarking

Our review found zero provenance or watermarking work targeting South Asian content, languages, or pipelines — and detectors trained on Western benchmarks lose up to 10.7 points across racial groups, more than 30 for some architectures.

Regulation

The EU AI Act's transparency mandates set a global reference point — but South Asian stakeholders face absent or unenforceable national law, and no seat where governance frameworks are designed.

Standards built for the open web fail quietly inside encrypted, forward-chain information ecologies.

Trust & safety must be built where the users are

Design

Human-centered, participatory design with existing community networks — journalist coalitions, feminist peer-support systems — as genuine partners and distribution channels.

Literacy

Not an individual skill deficit to fix, but collective resilience: localized, inoculation-based tools embedded in the platforms and social practices where manipulated media actually spreads.

Policy

Linguistic localization as a safety-critical requirement; representative benchmarks and survivor protection over definitional debates; Global South civil society at the table where AI governance is written.

The communities most harmed by deepfakes are not passive victims — their knowledge is the infrastructure. Build the conditions for it to shape the solutions.

Thank you

Dr. Dilrukshi Gamage · dgs@ucsc.cmb.ac.lk · dilrukshigamage.org

With gratitude to co-authors and collaborators:

Dilki Sewwandi, Min Zhang, Arosha Bandara, Shreeti Shubham, Sumaia Arefin, Dilaxshini Dunstan, Kartik Joshi, Rashmi Gunawardana — and 49 workshop participants and 100 field-study users across South Asia.

Labeling Synthetic Content (CHI 2025) · Deepfake.help / DF.Literacy (PACM HCI, CSCW) · Deepfakes in the Global South (J. Online Trust & Safety, 2026) · One Quarter of the World, Underserved (ACM Computing Surveys, in submission)

PhD Seminar for PhD Program in Information Technology

ORGANIZER

Organizer: Dr. Yong Zheng
Office: IIT Galvin Tower #15C3-1, 10 West 35th Street, Chicago, IL 60616, USA
Telephone: 312.567.3575
Email: yzheng66 [at] illinoistech.edu

2026 FALL

	Date	Invited Speaker	Title	Slide
+	08/28/2026 12:45 - 1:45 PM	Dr. Apurv Shukla University of Michigan-Dearborn, USA	Leaderboards with Receipts: Certifying LLM Rankings and Directing Preference Collection (in-person)	
+	09/25/2026 12:45 - 1:45 PM	Dr. Tapadhir Das University of the Pacific, USA	Resilient AI against Adversarial Interactions in Autonomous Vehicular Perception (Remote)	PDF
-	10/09/2026 12:45 - 1:45 PM	Dr. Dilrukshi Gamage University of Colombo School of Computing, Sri Lanka	Trust and Truth in the Age of Generative AI: A Global South Perspective (Remote)	PDF

Bio: Dr. Dilrukshi Gamage is a Senior Lecturer at the University of Colombo School of Computing, Sri Lanka, where she leads the Colombo HCI Lab, and a Microsoft Research Fellow. Her research sits at the intersection of human-computer interaction and computational social science, focusing on the societal implications of AI-generated content — deepfakes, synthetic media, and misinformation and on building tools, labels, and literacy interventions that help people evaluate such content critically. She is a Visiting Research Fellow at the Centre for Protecting Women Online at The Open University, UK, and has held research affiliations with the Berkman Klein Center at Harvard and the Weizenbaum Institute in Berlin. Her work has been funded by competitive grants such as Google Academic Research Grant and MSR. She pioneered HCI as a research area in Sri Lanka and founded the Colombo ACM SIGCHI Chapter in 2019. Her work appears at ACM CHI, CSCW, ICWSM, IC2S2, Journal of Online Trust and Safety; she serves as a regular Associate Chair for CHI, Papers Chair for ACM COMPASS 2026, and on the editorial board of several key HCI journals.

Abstract: Generative AI has made it trivially easy to fabricate convincing images, audio, and yet the labels, detection tools, and literacy resources meant to help people navigate synthetic media are designed largely for Western platforms, languages, and users. In the Global South, where fact-checking infrastructure is thin and harms fall hardest on women and marginalized communities, this gap has serious consequences. This talk examines the problem through a series of empirical studies: user perceptions of warning-label designs for AI-generated content on social media; a field experiment in Sri Lanka evaluating Deepfake.help, a literacy intervention grounded in inoculation theory; and analyses of how online communities discuss, trade, and make sense of deepfakes. I will also consider how provenance standards such as C2PA and SynthID, and emerging regulation such as the EU AI Act, translate — or fail to translate into low-resource contexts. Drawing these threads together, I will argue for trust and safety infrastructure that centers Global South users, and outline pathways for building it through human-centered design, literacy, and policy.

