Centre for Advanced Technologies
in Online Safety | CATOS

Welcome Introduction

Online Trust and Safety Forum | 15 May 2024

Dr Yang Yinping
Director, Centre for Advanced Technologies in Online Safety (CATOS)
Lead Principal Investigator, Online Trust and Safety (OTS) Research Programme

Senior Principal Scientist, A*STAR Institute of High Performance Computing (IHPC)

1 Welcome to the Online Trust and Safety Forum

Centre for Advanced Technologies in Online Safety | CATOS

Welcomes All Guests to the OTS Forum!

Academia

NUS National University of Singapore
Nanyang Technological University Singapore
Singapore Management University
SIT Singapore Institute of Technology
SUSS Singapore University of Social Sciences
SUTD Singapore University of Technology and Design
Stanford Cyber Policy Center — Freeman Spogli Institute and Stanford Law School
AI Singapore
Macquarie University Sydney Australia
Mohamed bin Zayed University of Artificial Intelligence
resaro

Industry

Google
Meta
Adobe
aws
sph Media
mediacorp
QAI
PRIMER
ST Engineering
Ant Group
Meltwater
KNOREX
Custodio Technologies
TrustLab
Pindrop
Logically.
truescope

NGOs

National Research Foundation Prime Minister's Office Singapore
ActiveFence
sentient.io
NLB National Library Board Singapore
SHE
Cyber Youth Singapore
CheckMate
Temasek Foundation
Roses of Peace
SCWO Singapore Council of Women's Organisations
Bower Asia Group

Public Agencies

Ministry of Communications and Information
MINDEF Singapore
Ministry of Home Affairs
DSO National Laboratories
MOHT MOH Office for Healthcare Transformation
Infocomm Media Development Authority
CSA Singapore
DSTA Defence Science & Technology Agency
SGH
Agency for Science, Technology and Research Singapore
GovTech Singapore
HTX
CSIT Centre for Strategic Infocomm Technologies
PTC Public Transport Council

2 Introduction to CATOS

Centre for Advanced Technologies in Online Safety | CATOS

Online Trust and Safety (OTS) Programme

Smart Nation & Digital Economy (SNDE) Domain of Singapore's Research Innovation and Enterprise (RIE) 2025 Plan

Ministry of Communications and Information

National Research Foundation
Prime Minister's Office
Singapore

Agency for Science, Technology and Research
Singapore

Centre for Advanced Technologies in Online Safety | CATOS

CATOS: Mission, Vision and Objectives

Mission

To develop **whole-of-nation technology capabilities and ecosystem** to monitor and tackle online harm

Vision

To be a technology leader with robust capabilities that **counter online harms** and **create a safe online space for all**

Objectives

- Develop technology leadership through research excellence and the creation of novel IPs
- Translate and integrate research outputs into needle-moving applications
- Facilitate growth of leading Singapore-based companies
- Establish Singapore's position as OTS thought leader in the region
- Strengthen international partnerships
- Develop a strong core of local talents

OTS Space Demands a "System of the Systems" Perspective

Ministry of Communications and Information

Content Generation System → **1. Content Producers: Gov, Orgns, MSM, Content Sites (Web, Platforms)**

- Establish authoritative sources of truth
- Ensure and preserve publication and media propagation integrity
- Build "trustability" by design

Challenge #1. <u>Prevention</u> of Harmful Content at Source

- *Strengthen source, deter misinfo generation*
- *Proactive identification and dealing with misinfo sources*

Info Propagation System → **2. Propagators: Platforms, Influencers, Ad Tech**

- Stem the spread of harmful content
- Nullify effectiveness of the perpetrators
- Work with platforms and influencers

Challenge #2. <u>Early Detection</u> at Scale to Contain Harmful Content

- *Active detection and tackling of serious harms at scale*

Socio-Cognitive System → **3. Info Consumers: Populace, Orgns, Influencers, Community Leaders**

- Cultivate responsible info-sharing behaviours
- Foster digital and info-savvy communities

Challenge #3. <u>Mitigation and Empowerment</u>: Strengthen Populace, Empower Communities

- *Empower individuals and communities with tech capabilities*

Centre for Advanced Technologies in Online Safety | CATOS

Three Strategic Pillars: Deep Tech, Systems Engineering, and Data

Data Pillar

X/Twitter, Facebook, Reddit, Instagram, YouTube, Tiktok, Weibo, Xiao Hongshu, Kuaishou, Pinterest, Sina Weibo, Douyin, Clubhouse, Discord, MeWe, Rumble etc.

Pull live data pipeline

Systems Engineering Pillar

MULTI-LANGUAGE SOCIAL LISTENING

TESTING & SANDBOXING

USER APPLICATIONS

Test and integrate research outputs

Deep Tech Research Pillar

PREVENTION			EARLY DETECTION			MITIGATION & EMPOWERMENT			
Tech 1. Content Provenance & Authenticity for Trust by Design	Tech 2. Mis/Dis-Information Source Attribution	Tech 3. Multimodal Deepfakes Detector	Tech 4. Non-Factual Claims Detector	Tech 5. Propaganda Detector	Tech 6. Extremism & Hate Detector	Tech 7. Malicious Account/Bot Detector	Tech 8. Tools to Improve Debunking Effectiveness	Tech 9. Online Trust & Safety Policy Testing System	Tech 10. Media Literacy Enhancement Tools

Centre for Advanced Technologies in Online Safety | CATOS

OTS Toolkit Version 1

Built from A*STAR Tech BIP, outputs from diverse research departments *(details on tech features in Annex A)*:

Audio-Visual Deepfake Detector (ALETHEA - Bastion of Truth)

Multimodal Emotion Analysis Engines (CrystalFeel, Crystalace, Digital Emotions)

Machine Translation Engines (Chinese, Malay, BI)

Virality Analysis (Network; Content)

Advanced Social Listening System (Resonance Social)

CATOS introduces Version 1 of "OTS Deep Tech Toolkit" based on strong IPs and incorporating ecosystem inputs

Multimodal Analysis Engines with Intense Emotions, Hate and Toxicity Detection

- ✓ Intense emotions, e.g., fear, anger, sadness, in text and video
- ✓ Hate speech detection
- ✓ Toxicity detection (beta)

Multimodal Automatic Fact Checker & Audio-Visual Deepfakes Detector

- ✓ Fact-checking system from text, video, and multi-language inputs (beta)
- ✓ Configurable source database
- ✓ Deepfake w face, frame by frame

Integrative OTS Social Monitoring "Master App"

- ✓ Data API connection (FB, Reddit, Youtube etc.)
- ✓ Enriched insights using OTS engines (intense emotions, toxicity, Chinese/Malay/BI-to-English translation)
- ✓ Web-based dashboard

Neural Machine Translation Engines

- ✓ Chinese-to-English
- ✓ Malay-to-English
- ✓ Indonesia-to-English

Virality Predictor

- ✓ SG Twitter network analysis
- ✓ SG TikTok network analysis

Seamless Interface for Content Publication with Credentials

- ✓ CATOS Engineering's initial effort
- ✓ To bridge the C2PA tech gap for the real-world environment, e.g., publishing an image with content credentials with a web content management system

3 Our Progress and Activities

Centre for Advanced Technologies in Online Safety | CATOS

OTS Community Engagement

Programme Kick-off Meeting, 6 April 2023

Pre-Grant Call Networking Workshop, 25 Aug 2023

Key outcomes

- **Introduced OTS programme** and announced the start of the programme
- Engaged **100+ attendees** of working-level stakeholders from public agencies, companies and academia
- Generated **keen follow-up interest** from several organisations

Key outcomes

- **Over 120 participants from academia, industry, and public agencies**
- Served to **inspire and catalyse new research collaborations** in addressing significant problems in the online safety space for the OTS Open Grant Call

Centre for Advanced Technologies in Online Safety | CATOS

WEF Typology of Online Harms (2023)

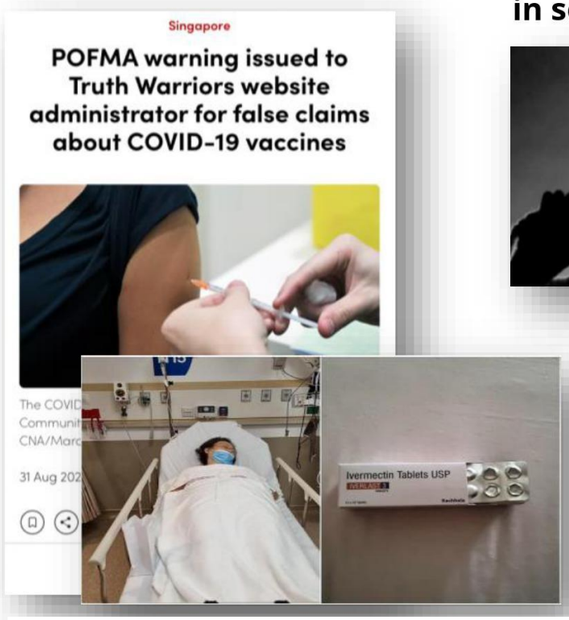
Type	Content Risks	Contact Risks	Conduct Risks
Threats to personal and community safety	• **Child sexual abuse material (CSAM)** • **Child sexual exploitation material (CSEM)** • **Pro-terror material** • **Content that praises, promotes, glorifies or supports extremist organizations or individuals** • **Violent graphic content** • **Content that incites, promotes or facilitates violence** • **Content that promotes, incites or instructs in dangerous physical behaviour**	• Grooming for sexual abuse • Recruitment and radicalization	• Technology-facilitated abuse (TFA) • Technology-facilitated gender-based violence
Harm to health and well-being	• **Material that promotes suicide, self-harm and disordered eating** • **Developmentally inappropriate content**		
Hate and discrimination	• **Hate speech**		• Algorithmic discrimination
Violation of dignity		• Sexual extortion	• Online bullying and harassment
Invasion of privacy			• Doxxing • Image-based abuse
Deception and manipulation	• **Disinformation and misinformation** • **Deceptive synthetic media**		• Impersonation • Scams • Phishing • Catfishing

https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

Centre for Advanced Technologies in Online Safety | CATOS

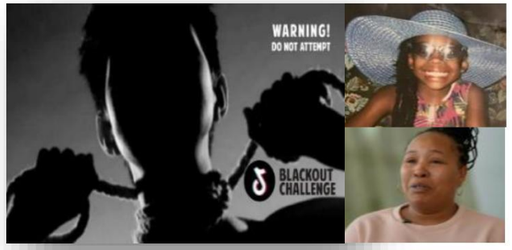
Case examples: False claims, dangerous challenges, and racially/ religiously divisive content

Case #1. COVID-19/ vaccine-related falsehoods



Chia O. (2021, October 4) Grandmother hospitalised after taking ivermectin to 'protect herself' against Covid-19. The Straits Times. https://www.straitstimes.com/singapore/grandmother-hospitalised-after-taking-ivermectin-to-protect-herself-against-covid-19

Case #2. "Unintended" viral content to vulnerable groups in society, such as children



Hahn, J.D., (2021, December). 10-Year-Old Girl Dies Trying 'Blackout Challenge' from Social Media, Mom Says, People, https://people.com/human-interest/10-year-old-girl-dies-trying-blackout-challenge-from-tiktok/; Picture credit: https://cybersafett.com/tiktok-blackout-challenge-what-you-need-to-know/

Case #3. Racially/religiously divisive claims

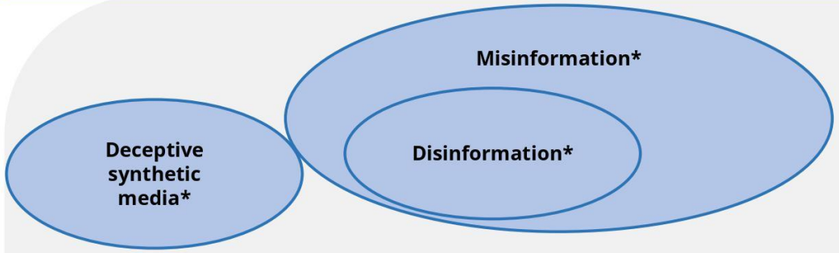
Zhuo, T. (2020, March 20). Police investigating Facebook post deemed offensive to Christians, Muslims: Shanmugam. The Straits Times. https://www.straitstimes.com/singapore/courts-crime/police-investigating-facebook-post-deemed-offensive-to-christians-muslims

Ishak S. (2021, May 27). S'pore Muslim religious teachers condemn offensive online poll asking which female asatizah 'should be gang-banged'. Mothership. https://mothership.sg/2021/05/ustazah-offensive-poll

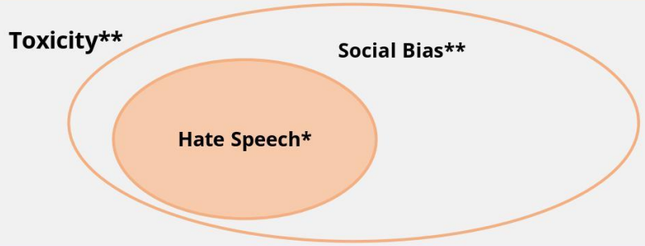
Centre for Advanced Technologies in Online Safety | CATOS

Five Focal Areas of Harmful Online Content

Area A. Deception and manipulation

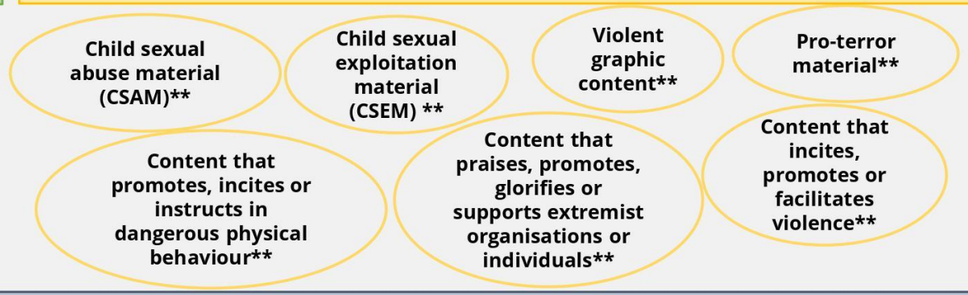


Area B. Hate and discrimination

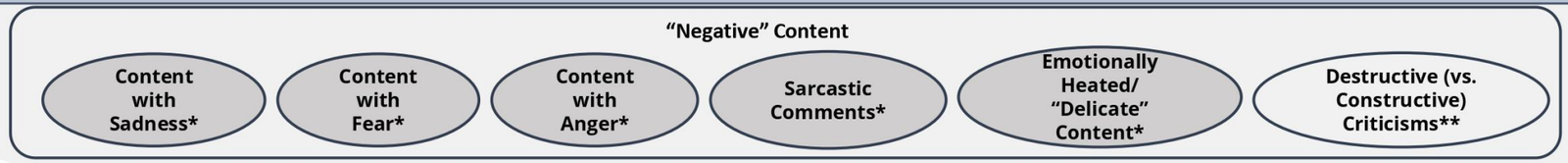


Area C. Harm to health and well-being

Area D. Threats to personal and community safety

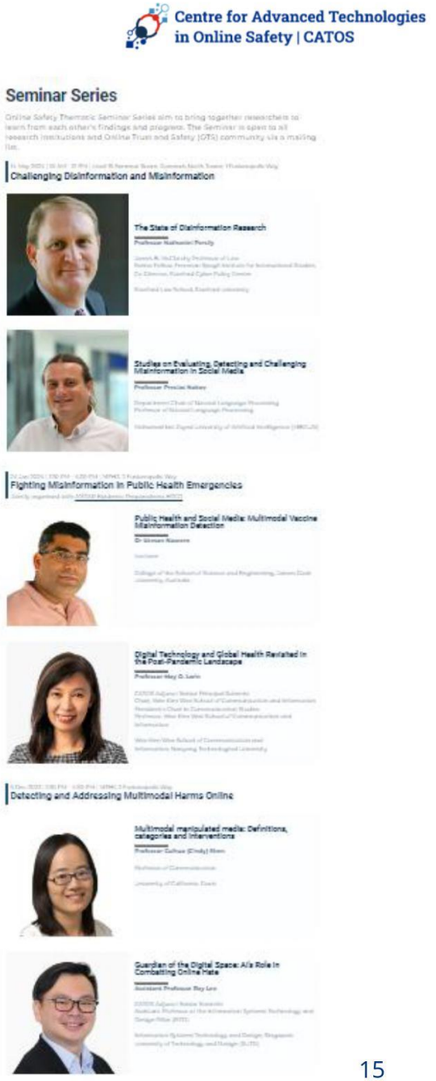


Area E. Harm to organizations and brands*



Refs for Areas A-D: https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

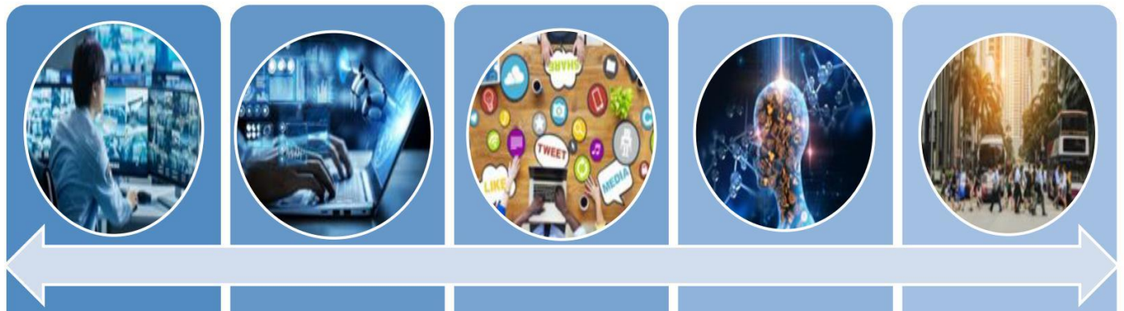
Catos.sg



4 Conclusion

Centre for Advanced Technologies in Online Safety | CATOS

We seek to collaborate and deliver advanced tech solutions that benefit our local & global OTS stakeholders at scale



Government	Mainstream Media Companies	Social Media Companies	Fact Checkers & Tech Systems Integrators	General Populace
Better accuracy and faster detection - (productivity gains)	**Enhanced information credibility**	**Improved platform safety**	**Enhanced product and services offering**	**Increased awareness of dangers online and equipped with tools**

Systems Engineering:

- **High-readiness technology / tools** are ready for adoption/use
- **Develop mid-readiness technology** with industry evaluation and validation
- Build **end-user applications**

Deep Tech Research:

- **Open Grant Call for proposals**, where the topic is at the basic and early research stage

Coordinating Office:

- **Workshops, seminars, events, outreach activities** etc.

Centre for Advanced Technologies
in Online Safety | CATOS

THANK YOU

For more information, please visit www.catos.sg