



**Centre for Advanced Technologies  
in Online Safety | CATOS**

# Welcome Introduction

**Online Trust and Safety Forum | 15 May 2024**

**Dr Yang Jinping**

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)

ARES PUBLIC





# **1 Welcome to the Online Trust and Safety Forum**





# Welcomes All Guests to the OTS Forum!

Academia









NGOs







ARES PRIVATE | PUBLIC

3





## 2 Introduction to 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





## 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**

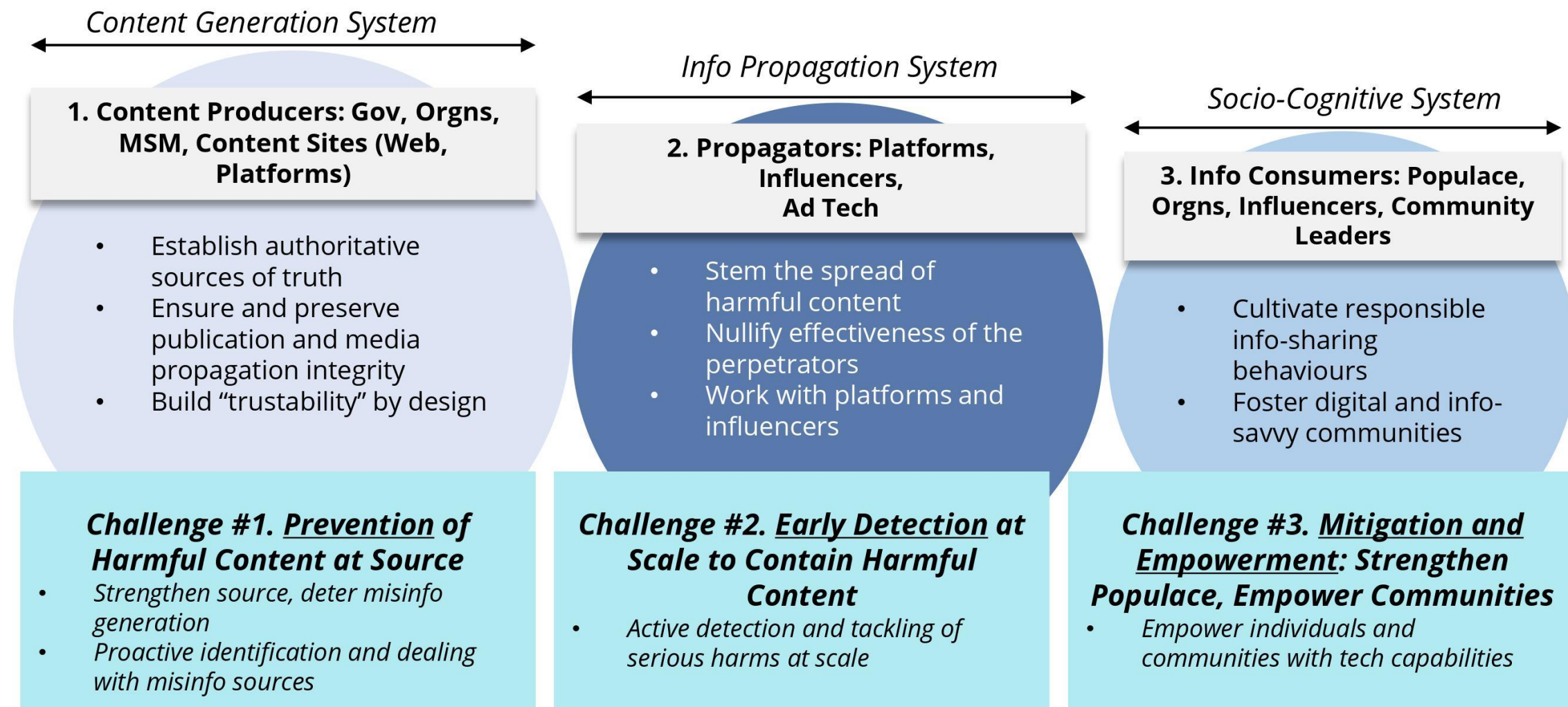
### Objectives

- 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







## 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

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

#### EARLY DETECTION

Tech 5.  
Propaganda  
Detector

Tech 6.  
Extremism  
& Hate  
Detector

Tech 7.  
Malicious  
Account/Bot  
Detector

#### MITIGATION & EMPOWERMENT

Tech 8.  
Tools to  
Improve  
Debunking  
Effectiveness

Tech 9.  
Online Trust  
& Safety  
Policy Testing  
System

Tech 10.  
Media  
Literacy  
Enhancement  
Tools





## 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





## OTS Community Engagement

### Programme Kick-off Meeting, 6 April 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

### Pre-Grant Call Networking Workshop, 25 Aug 2023



#### 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





## WEF Typology of Online Harms (2023)

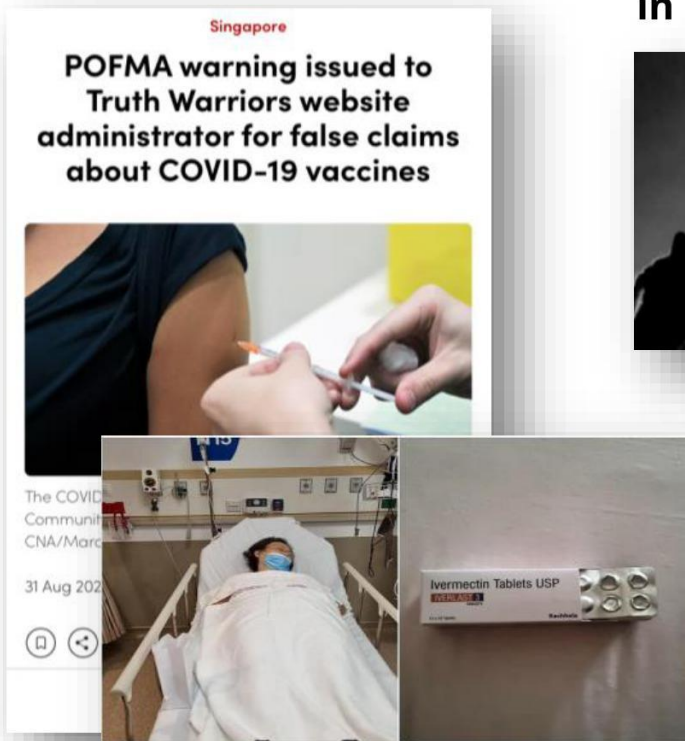
Type	Content Risks	Contact Risks	Conduct Risks
Threats to personal and community safety	<ul style="list-style-type: none"> <li>• <b>Child sexual abuse material (CSAM)</b></li> <li>• <b>Child sexual exploitation material (CSEM)</b></li> <li>• <b>Pro-terror material</b></li> <li>• <b>Content that praises, promotes, glorifies or supports extremist organizations or individuals</b></li> <li>• <b>Violent graphic content</b></li> <li>• <b>Content that incites, promotes or facilitates violence</b></li> <li>• <b>Content that promotes, incites or instructs in dangerous physical behaviour</b></li> </ul>	<ul style="list-style-type: none"> <li>• Grooming for sexual abuse</li> <li>• Recruitment and radicalization</li> </ul>	<ul style="list-style-type: none"> <li>• Technology-facilitated abuse (TFA)</li> <li>• Technology-facilitated gender-based violence</li> </ul>
Harm to health and well-being	<ul style="list-style-type: none"> <li>• <b>Material that promotes suicide, self-harm and disordered eating</b></li> <li>• <b>Developmentally inappropriate content</b></li> </ul>		
Hate and discrimination	<ul style="list-style-type: none"> <li>• <b>Hate speech</b></li> </ul>		<ul style="list-style-type: none"> <li>• Algorithmic discrimination</li> </ul>
Violation of dignity		<ul style="list-style-type: none"> <li>• Sexual extortion</li> </ul>	<ul style="list-style-type: none"> <li>• Online bullying and harassment</li> </ul>
Invasion of privacy			<ul style="list-style-type: none"> <li>• Doxxing</li> <li>• Image-based abuse</li> </ul>
Deception and manipulation	<ul style="list-style-type: none"> <li>• <b>Disinformation and misinformation</b></li> <li>• <b>Deceptive synthetic media</b></li> </ul>		<ul style="list-style-type: none"> <li>• Impersonation</li> <li>• Scams</li> <li>• Phishing</li> <li>• Catfishing</li> </ul>





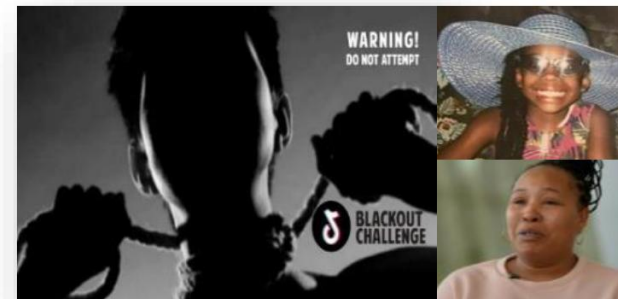
## Case examples: False claims, dangerous challenges, and racially/ religiously divisive content

### 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>



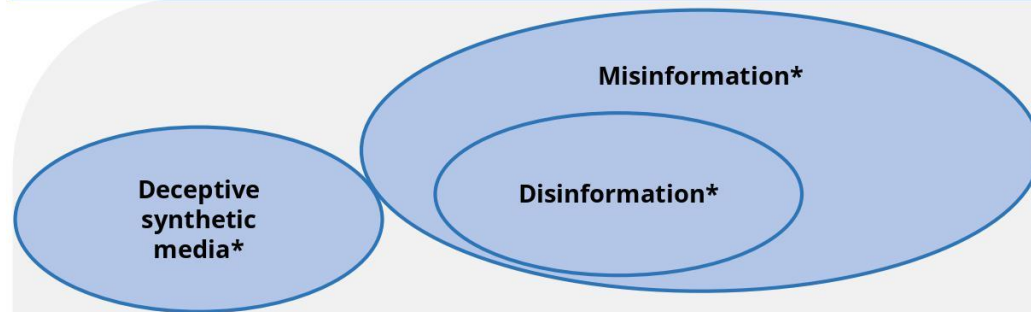
Ishak S. (2021, May 27). S'pore Muslim religious teachers condemn offensive online poll asking which female asatizah 'should be gang-banged'. Mothership. <https://mothership.sg/2021/05/ustazah-offensive-poll>



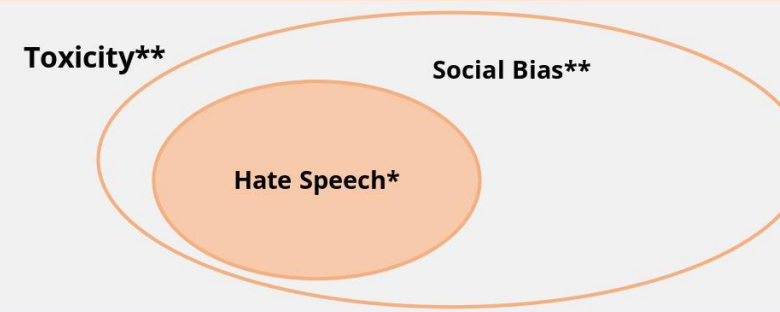


### Five Focal Areas of Harmful Online Content

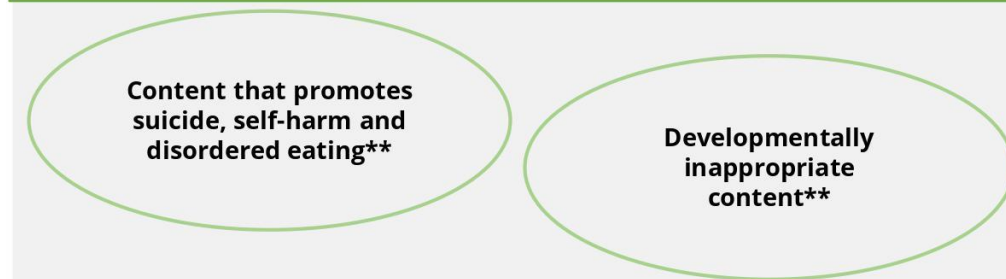
#### Area A. Deception and manipulation



#### 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](https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf)



# Online Trust and Safety Forum

15 May 2024



## Centre for Advanced Technologies in Online Safety | CATOS

### Catos.sg



Home Deep Tech Research Technologies What's On About

### Online Safety is a Foundation to a Thriving Digital Society

The Internet and digital platforms provide instant ways for people to connect, socialise and share information and their thoughts and feelings. At the same time, the spread of harmful online content gets easier, cheaper, and faster.

The Centre for Advanced Technologies in Online Safety (CATOS) focuses on building robust technological capabilities that combat online harms, including misinformation and content that undermines people's well-being.

[Learn more about CATOS](#)

Technologies

#### Technologies

To facilitate research translation, we consolidate a collection of technologies which has mid-to-high readiness for evaluation, adoption and various potential use cases.

Considering the landscape and ecosystem needs, our portfolio of technologies focuses on addressing two broad areas of harmful online content: a) Misinformation, and b) Content that undermines people's well-being. The focus of the technologies is facilitated by prioritizing content that may present a severe and large reach of the population, which can be characterized by high emotion intensity and virality. Enquiries and interest to try out any of these tools can be sent to [connect@catos.sg](mailto:connect@catos.sg).



#### Multimodal Emotion Analysis Engines

Early sensing of content and communities with high intensity fear, anger, sad and joy emotions.



#### Multimodal Context-aware Automatic Fact Checker (mCAFC)

Discerns false from authentic news despite human or machine generated.



#### Neural Machine Translation Engines

Automatic translations of Mandarin, Malay, and Bahasa Indonesian content to English for multi-language analysis.



#### Virality Predictor

Predictive scoring versus monitoring of digital content popularity and virality using network-based and content-based features.



#### Facial & Audio Deepfakes Detector

Automatic detection of manipulations in audio and video deep fakes that involve human speech.



#### Integrative OTS Platform

Enable access to OTS high readiness technologies and continuous social listening for online emotions, potential falsehood, misinformation, disinformation and disinformation, especially those that are likely to go viral.

What's On / [Key Announcements](#)

#### Key Announcements

In partnership with leading organisations, CATOS organises activities and events to galvanise broad research and practitioner communities and advance the development of the OTS programme. These activities include Open Grant Call, Technology Challenges, Workshops, Forums, etc. Here are the key activities announced to date.

18 May 2024

#### OTS Forum

Organised by CATOS, the inaugural OTS Forum brings together Singapore-based and international thought leaders and professionals to exchange insights on the latest developments and trends in policy, education, and technology in creating a safer internet for all. The Forum is designed to foster interaction between speakers, facilitators and participants from academia, industry, public sector agencies, and non-governmental organisations (NGOs). About 200 participants are expected on a by-invite basis.

31 January 2024

#### Online Safety Prize Challenge

CATOS is pleased to be the Singapore's Supporting Partner in the organisation of the Online Safety Prize Challenge (OSPC). Launched on 31 January 2024, this 10-week competition aims to advance AI research in developing models to detect between benign and harmful content relevant to Singapore's diverse societal background and digital landscape. Click here for more details via the DSPC website.

1 November 2023

#### Launch of OTS Open Grant Call

CATOS launched the OTS Open Grant Call for Deep Tech Research Proposals and Solutions on 1 November 2023. The objective is to invite research proposals to enhance technological capabilities and pipeline in three themes: Prevention, Early Detection, and Mitigation and Empowerment. Click here for more details.

#### OTS Open Grant Call

Online Trust And Safety Research Programme  
Open Grant Call For Deep Tech Research Proposals And Solutions  
1 Nov 2023 – 10 Jan 2024

##### Background

Harmful online content causes severe threats and damage to people's well-being, costs lives, and undermines societal stability. Being a digitalised and connected country, Singapore is highly susceptible to harms from the online space. Local mainstream media and public agencies are intensively dealing with false claims targeted at Singapore-specific named entities or their identities. While leading international social media platforms are taking steps to address harmful content or misinformation spread, and the private sector offers commercial tools, the effectiveness and timeliness of existing solutions very significantly when applied in the Singapore context. While we shall tap into global tech advances in the Online Trust and Safety (OTS) space to grow in tandem, Singapore needs content aware tools to ensure the technological means to defend ourselves from such internet harms.

The OTS Research Programme is a funding initiative from the Ministry of Communications and Information (MCI). It is a part of the Smart Nation and Digital Economy (SNDE) domain of the Research, Innovation and Enterprise 2025 programme (RIE2025) administered by the National Research Foundation (NRF). The overarching aim of the OTS Research Programme is to develop technological capabilities to combat online falsehood and harms and assist translation efforts with high-impact use cases.

With support from the MCI, NRF and partner organisations, the Agency for Science, Technology and Research (A\*STAR) established the Centre for Advanced Technologies in Online Safety (CATOS) in April 2023 to host the OTS Research Programme as its flagship programme. Based on A\*STAR's motto of high performance technology, CATOS comprises functions responsible for key strategic pillars for the OTS Research Programme, including Deep Tech Research (focusing on selecting and funding low TRL research with high competitive advantages and potential to move to mid TRL), Systems Engineering (focusing on mid-high TRL technology evaluation, integration and translation of technology outputs into ready-moving use cases), and Programme Coordination activities (focusing on engaging research and practitioner communities through workshops, forums, and collaborative networking).

CATOS works in close consultation with MCI to prioritise technologies that address use cases affecting Singapore at the national level. Based in Singapore, we differentiate by building content aware OTS tech capabilities that can take into account the local user behaviour, local context, and local languages. We envisage that upon successful execution, technological capabilities developed from the OTS Research Programme will significantly complement existing legislation and education efforts to detect and deal with the rapidly evolving threat of online harms to Singapore and our society.

##### Objectives

This OTS Research Programme - Open Grant Call for Deep Tech Research Proposals and Solutions (hereafter referred to as 'OTS Open Grant Call') is launched to select high-quality projects that can deliver a strong suite of robust technological solutions that address significant unmet problems in the OTS space.

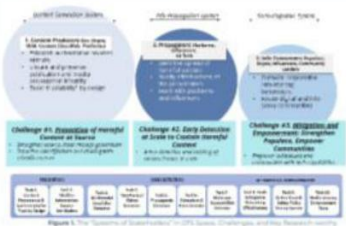
We seek projects that can potentially yield technology R&D outcomes for:

- Help Singapore develop technology leadership in the OTS space through research excellence such as notable publications and presentations at top scientific and industry conferences.
- Foster strong pipeline capabilities to equip Singapore-based companies and public agencies with effective tools to combat online harms.

The projects supported by the OTS Open Grant Call should deliver technologies at mid-TRL, in the form of code, software development kits, Docker images or similar, with user documentation that describes the training method and validation results. Deliverables should be ready for industry-strength evaluation to test the delivered technology's robustness in out-of-training-sample test cases that come from real-world or simulated scenarios.

##### Research Themes and Technology Topics of Interest

The Online Trust and Safety (OTS) space is exceptionally complex and rapidly evolving. To effectively define a technology research agenda and prepare for high-impact translational use cases, CATOS has consulted academia, industry and public agencies and developed a framework to define the key research-worthy technology topics of interest based on MCI's Systems perspective of the online 'harm space' and the associated challenges identified in each of the systems (Figure 1).



Specifically, the OTS Open Grant Call seeks for proposals of technologies tackling online harms under the following three research themes. Ten Technology Topics have been identified under the three Research Themes.

- Research Theme 1: Prevention
- Research Theme 2: Early Detection
- Research Theme 3: Mitigation & Empowerment



#### Seminar Series

Online Safety Thematic Seminar Series aim to bring together researchers to learn from each other's findings and progress. The Seminar is open to all research institutions and Online Trust and Safety (OTS) community via a mailing list.

##### Challenging Disinformation and Misinformation



#### The State of Disinformation Research

Professor Nicholas Fordy  
Senior Lecturer, School of Law  
National Police, Forensic Speech Analysis for International Studies  
The University of Queensland, Australia



#### Studies on Evaluating, Detecting and Challenging Misinformation in Social Media

Professor Priscilla Kelsey  
Regional Centre of Expertise in Language Technology  
Professor of Natural Language Processing  
National Institute of Education, Singapore

##### Fighting Misinformation in Public Health Emergencies



#### Public Health and Social Media: Multimodal Vaccine Misinformation Detection

Dr. Steven Kwan  
Associate Professor  
College of the Holy Spirit, Research and Engineering, Chinese Culture University, Taiwan



#### Digital Technology and Global Health Resilience in the Post-Pandemic Landscape

Professor Ming Q. Lu  
CATOS Adjunct Senior Principal Scientist  
Chair, New York University School of Communication and Information  
Residential Chair in Communication Studies  
New York University School of Communication and Information



#### Detecting and Addressing Multimodal Harms Online

Professor Goh Kah Seng  
Professor of Communication  
University of California, Davis



#### Guardian of the Digital Space: AI's Role in Combating Online Harms

Professor Ray Lee  
CATOS Adjunct Senior Principal Scientist  
National Institute of Education, Singapore  
Design Fellow (2021)  
Institute of Systems Technology and Design, Singapore  
University of Technology and Design (2022)



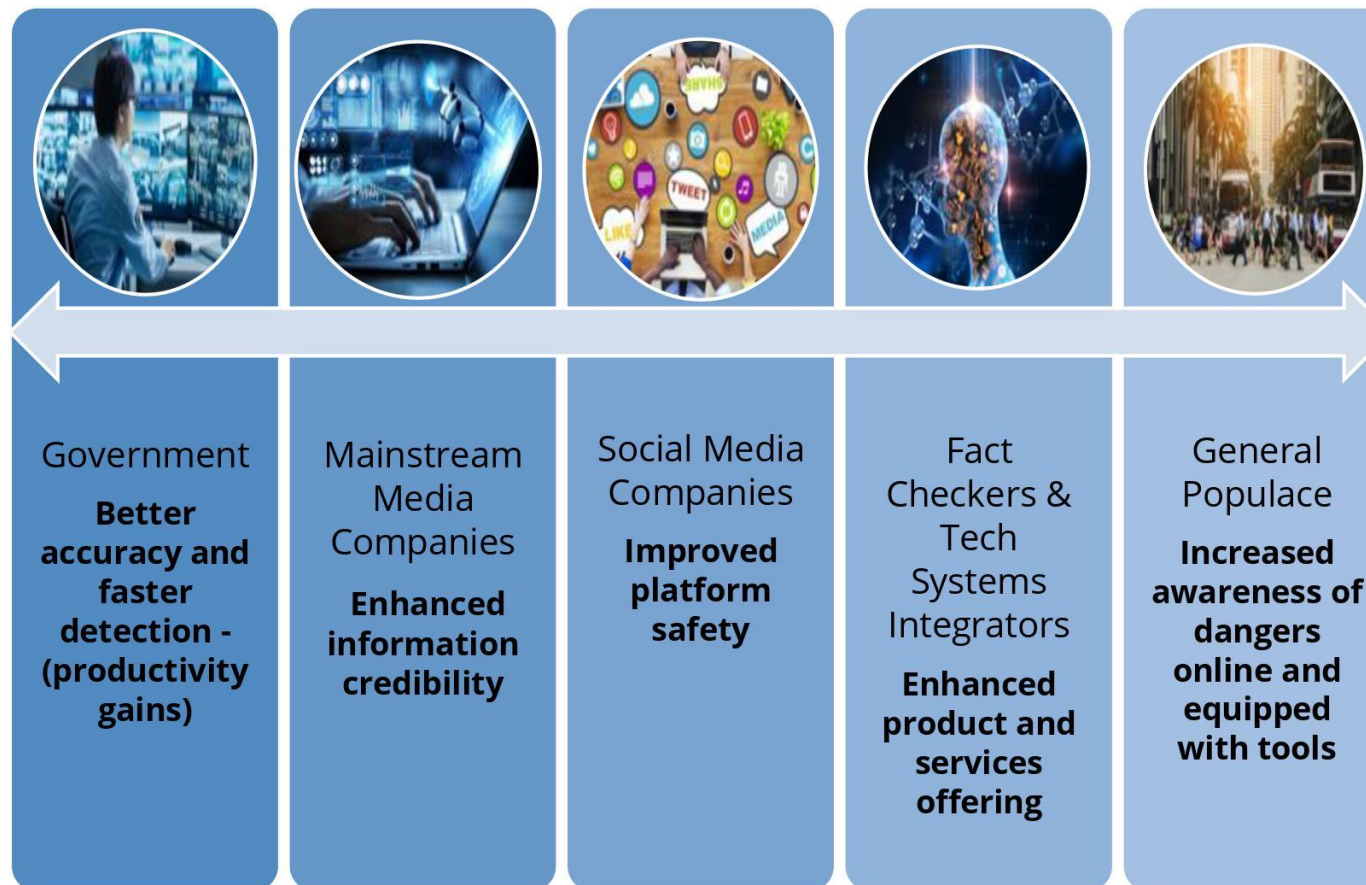


## 4 Conclusion





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



### 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.





# THANK YOU

For more information, please visit [www.catos.sg](http://www.catos.sg)

