ONLINE TRUST AND SAFETY RESEARCH PROGRAMME – OPEN GRANT CALL FOR DEEP TECH RESEARCH PROPOSALS AND SOLUTIONS

Call for Proposals – Info Sheet

Online Trust and Safety (OTS) Research Programme Coordinating Office:

Centre for Advanced Technologies in Online Safety
(CATOS) Email: CATOS-OTS@hq.a-star.edu.sg

Hosted by
Institute of High Performance Computing (IHPC)
Agency for Science, Technology and Research (A*STAR)
1 Fusionopolis Way, #16-16 Connexis North Tower, Singapore 138632

1. BACKGROUND

1.1 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 vary 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 context-aware tools to ensure the technological means to defend ourselves from such internet harms.

1.2 The OTS Research Programme is a funding initiative from the Ministry of Communications and Information (MCI). It is 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 align translation efforts with high-impact use cases.

1.3 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 in A*STAR's Institute of High Performance Computing (IHPC), CATOS comprises functions responsible for key strategic pillars for the OTS Research Programme, including Deep Tech Research Pillar (focusing on selecting and funding low-TRL research with high competitive advantages and potential to move to mid-TRL), Systems Engineering Pillar (focusing on mid-high TRL technology evaluation, integration and translation of technology outputs into needle-moving use cases), and Programme Coordination activities (focusing on engaging research and practitioner communities through workshops, forums, and collaborative networks).

1.4 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 context-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.

2. OBJECTIVES OF THE OTS OPEN GRANT CALL

2.1 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 solicit high-quality projects that can deliver a strong suite of robust technological solutions that address significant unmet problems in the OTS space.

2.2 We seek projects that can potentially yield technology R&D outcomes to

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

2.3 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.

3. RESEARCH THEMES AND TECH TOPICS OF INTEREST

3.1 The 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 harms space" and the associated challenges identified in each of the systems (Figure 1).

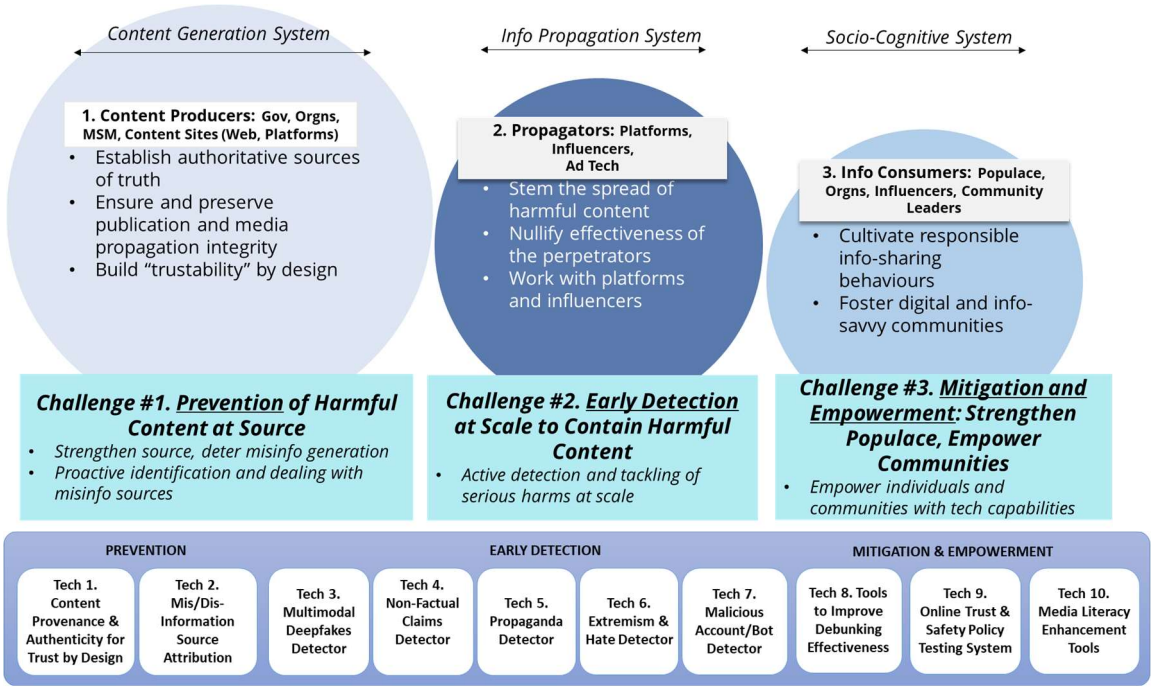

Figure 1. The "Systems of Stakeholders" in OTS Space, Challenges, and Key Research-worthy Technology Topics of Interests over Three Themes

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

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

3.3 **Research Theme 1: Prevention** engines and tools that enhance trust and/or reduce mistrust towards mainstream media and authorities and identify the common sources of mis/disinformation for actions to stem the spread from the root.

Tech Topic 1. Content Provenance & Authenticity for Trustability by Design

A content signature generator system (based on asymmetric encryption and distributed ledger technology) that publishers and content generators can use to allow info consumers to verify the provenance and uniqueness of content. Such a solution can alert the info consumer or propagator if the content has been flagged or tampered with.

Tech Topic 2. Mis/DisInformation Source Attribution

Technology that could help relevant users to identify and attribute sources and key spreaders of identified or suspected mis/disinformation. This technology shall present features to help users decode digital narratives and achieve a better understanding of unfolding events in the world.

For the Prevention thematic tools, we will prioritise methods that can leverage an advanced understanding of the notion of trust (e.g., [1-3]), and can demonstrate a substantial advancement in existing approaches (e.g., [4-10]) in real-world settings.

3.4 **Research Theme 2: Early Detection** engines that help assess the existence and intensity of harmful online content, including deepfakes, non-factual claims, propaganda, extremism and hate speech. Recognising existing work in deepfake detection (e.g., [11-13]), non-factual claim detection (e.g., [14-16]), propaganda detection (e.g., [17-18]), hate speech and hate network detection (e.g., [19-20]), and bot detection (e.g., [21-24]), the proposed solutions should be able to present features with a significant advance of user values in real-world settings.

Tech Topic 3. Multimodal Deepfakes Detector

Technology that detects inauthentic media with audio-visual content manipulation, especially the manipulation at the messaging and narrative level through cross-modal manipulation. The technology shall facilitate users with research-grounded explainable features to discern manipulation surrounding facial/non-facial regions, speaker/voice manipulation, lip-synching, lighting, language use, and even the malicious vs. non-malicious intent behind the author who posts the content.

Tech Topic 4. Non-Factual Claims Detector

Technology that assists a human user and/or is deployable in systems throughout the fact-checking process to verify claims as early as possible and be sensitive to Singapore and the region's multilingual context.

Tech Topic 5. Propaganda Detector

Technology that assists a human user and/or is deployable in systems to automatically detect the presence of propaganda at both the overall level (i.e., propaganda vs. non-propaganda) and the propaganda technique level (e.g., loaded language, name-calling and labelling, flag-waving, doubt). The content can be in news segments, messages or social media content. This technology would help identify propagandistic content harbouring ill intentions against Singapore and our society.

Tech Topic 6. Extremism & Hate Detector

Technology that automatically detects online extremism and hate speech from social media text/audio/video. This includes anything that encourages hate or violence towards a person/group based on race, religion, ethnic origin, age, gender, sexual orientation, and/or disability. Detectors sensitive to Singapore and the region's specific

socio, linguistic and cultural values, such as racial and religious harmony, are particularly needed.

Tech Topic 7. Malicious Account/Bot Detector

Technology that detects inauthentic user accounts and identifies bots, particularly those that undertake malicious activities to generate online harms, such as illegal and ill-intended uses of logos of creditable organisations. This can include technology to detect large numbers of bots that are coordinated and controlled by a group of common users to amplify mis/disinformation.

For the Early Detection thematic tools, we will prioritise and select methods and systems that can present a significant advance over existing solutions, such as a) displaying high detection accuracy for real-world data, with applicability and robustness for the Singapore context, b) presenting features that are explainable to facilitate user adoption, and c) advancing the understanding of the underlying phenomenon.

3.5 **Research Theme 3: Mitigation & Empowerment** is a promising area where technology researchers, social, behavioural and communication scientists and practitioners can join forces to create theoretically and empirically sound tools for alerting, debunking, or even pre-bunking online mis/disinformation.

Tech Topic 8. Tools to Improve Debunking Effectiveness

Technology that algorithmically assists content generators and communication professionals in evaluating and recommending more effective mis/disinformation debunking messaging, such as the clarity, creditability, vividness, emotionality, empathy, and contextual dimensions of a crafted message intending to mitigate the negative impact of a piece of mis/disinformation which has already emerged or spread.

Tech Topic 9. Online Trust & Safety Policy Testing System

Technology that assists policymakers, analysts, content moderation leads and info operation teams to evaluate the impact of policy decision-making based on personas that represent the communities. This technology needs to factor in Singapore's local context, such as local demographics and psychographics and Singapore's values, to enable policy testing that is not feasible to generate and evaluate using conventional approaches.

Tech Topic 10: Tools to Enhance Media Literacy

Development and testing of “gamified education” tools such as mobile apps that seek to increase media literacy and critical thinking through gamification, quizzes, knowledge modules, etc. The tools and studies shall advance the understanding of our population's vulnerabilities from existing studies and examine with efficacy studies the extent to which such media literacy enhancement tools can increase “psychological defence" by educating citizens effectively and inoculating citizens against the effects of online falsehood and harms.

For the Mitigation & Empowerment thematic tools, we will prioritise projects that present highly novel and effective methods based on sound psychological knowledge and smart design for different demographic groups [e.g., 25-32].

3.6 Please refer to **Annex A** for more details on the Research Themes and the respective Tech Topics under the Research Themes.

3.7 Proposals can either tackle one or more of the tech topics of interest. Please clearly indicate the Tech Topic(s) that the proposal will address on the Proposal Cover Page.

3.8 With solid justification and rationale, proposals are allowed to propose tech topics not identified in Annex A. Such proposals must clearly explain the value of the proposed research approach in meeting the broader aims of the OTS Open Grant Call, provide comparative evaluations against state-of-the-art approaches (including, if necessary, the ten identified tech topics), and eventually be demonstratable and robust when evaluated using real-world local context data.

4. FUNDING SUPPORT

4.1 The OTS Open Grant Call offers two categories of funding support:

- Category 1: funding up to S$500K (inclusive of 30% indirect cost) for a proposal focusing on addressing one Tech Topic

- Category 2: funding up to S$3M (inclusive of 30% indirect cost) for a proposal focusing on addressing two or more Tech Topics

4.2 This Open Grant Call offers funding support for up to 2.5 years.

4.3 Applicants are strongly encouraged to collaborate with industry partners (e.g., mainstream media companies, digital platform companies, independent fact-checkers, etc.) and government public sector agencies to develop innovative solutions to address the grant call objectives and demonstrate strong potential (or, more preferably, actual technology deployment and use cases) for real-world applications within and beyond Singapore.

5. ELIGIBILITY CRITERIA

5.1 At the point of application, the Principal Investigators (PIs) and Co-Principal Investigators (Co-PIs) must hold a full-time appointment and be salaried in a local publicly funded institution. The full-time appointment is defined as at least 9 months of service a year based in Singapore or 75% appointment.

5.2 The PI and Co-PIs must be based in Singapore. Collaboration with foreign organisations and experts in the capacity of Collaborator is allowed. Research work should be done in Singapore and should not be carried out overseas unless expressly approved by the grantor.

5.3 Where applicable, we encourage the integration of relevant real-world use cases or social and behavioural research to complement the technology-centric R&D work under this grant call to ensure the practicality, user-centricity and acceptability of the solutions proposed.

5.4 Proposals that other government agencies already fund will not be considered under this grant call. PIs shall declare their other funding sources and participation in other funding initiatives during application. Proposals with similar scope, currently funded or under evaluation by other funding initiatives, will not be considered.

5.5 In addition to the above criteria, applicants should not have any outstanding reports from previous A*STAR grants, NMRC grants and other national competitive grants.

6. EVALUATION CRITERIA

6.1 Selection of successful proposals will be based on, but not limited to, the following evaluation criteria:

i. The novelty, intellectual and innovative merit of the proposal
ii. The competitive advantage of the proposed solution(s) in meeting the aims of the OTS Open Grant Call
iii. Potential technological impact based on the proposed deliverables, in particular, the likelihood of technology being adopted and used by the industry through licensing
iv. The feasibility of implementing and deploying the proposed methodology
v. The track record of the PI and the team

6.2 The proposals will be reviewed by international scientific reviewers and the OTS Open Grant Call Proposal Review Panel based on the evaluation criteria. PIs of shortlisted proposals may be invited to give presentations and answer questions from the Proposal Review Panel.

6.3 All decisions are final, and no appeals will be entertained.

7. TIMELINE

7.1 The OTS Open Grant Call applications will open on **Wednesday, 1st Nov 2023** and close on **Wednesday, 10th Jan 2024 1700hrs (SG Time)**.

This grant call is targeted at the following review timeline (subject to change):

Evaluation Panel Meeting	– 15 March 2024
Announcement of grant call results	– 15 April 2024
Official project kick-off	– 1 July 2024

8. SUBMISSION INSTRUCTIONS

8.1 PIs should use the proposal and budget templates provided and follow the instructions stated in the templates.

8.2 Only the Principal Investigator from each research team can submit the application. The applicant will be the primary contact for the research team.

8.3 The application must be endorsed by the relevant institutional authority/director of research (or equivalent) of PIs and Co-PIs whenever applicable. Incomplete applications will not be accepted.

8.4 The application must be submitted through the iGrants system (igrants-app.a-star.edu.sg) before **Wednesday, 10th Jan 2024, 1700 hrs (Singapore Time).** Late applications will not be accepted.

8.5 Inquires can be addressed to CATOS-OTS@hq.a-star.edu.sg.

Annex A: Research Themes and Tech Topics of Interest

Online harms are massive in scale, dynamically evolving, and often beyond the capabilities of a few entities to control. Fake images, audio and videos are increasingly sophisticated and difficult to discern, with freely available tools requiring little expertise to use. Social botnets and troll farms are used to automate the spread and amplify their effects. Propagandists deliberate with techniques such as loaded emotive language and appeal to fear and prejudice, which may cause substantial harm without using objectively false information. More subtly, hostile actors use increasingly coordinated techniques, such as building seemingly innocuous communities before posting harmful content later. We need more robust tools to sense and detect the subtler, weaker signals at a very early stage before harm is caused.

Effective use of OTS tech capabilities can be wide-ranging, including but not limited to tackling mis/disinformation in health and science, assisting newsrooms in fact-checking, and tackling mis/disinformation in more complex societal and political issues. The following tech topics are of particular interest for the OTS Open Grant Call.

Research Theme 1	**PREVENTION**
Tech Topic 1 Content Provenance & Authenticity for Trustability by Design	Existing solutions for combating misinformation, which rely primarily on passive detection and retrospective tracing, are often inefficient and contribute little to nurturing a cohesive community where people voluntarily and actively engage in trust-building missions with collective effort. This tech aims to establish a robust, autonomous and sustainable trust-centric ecosystem based on unique online trust identities for the general public and official sources. Features enabled by such an ecosystem include: (1) Users and official sources are each associated with a unique trust identity of integrity and sustainability (2) Content can be securely bound with its source's trust identity during the entire life cycle of generation, propagation, consumption and citation (3) Misinformation would be subdued by discouraging its generation with a lasting negative impact on its source’s associated trust identity, and curbing its propagation by reluctant sharing and spontaneous flagging by users to boost their trust identities (4) Communities of strong trust can be cultivated with networks of users in which those of strong trust identities safeguard critical information diffusion paths and collectively shield others from the attack of misinformation Key research topics and features required include but are not limited to: • Decentralised trust identity with integrity and uniqueness enabled by identity-data-binding and distributed ledger technology

	• Content signature and provenance by asymmetric encryption, digital watermarking, privacy-enhancing verification and distributed ledger technology • User and content trust index system • Data assetisation and tokenomics-based incentive design Evaluation metrics – Fault tolerance with threat models, user adoption rate, user activity level
Tech Topic 2 Mis/DisInformation Source Attribution	Tracing back to the originator and spreaders is a challenging investigative task but can be the most effective strategy in preventing future situations. This tech topic aims to build a real-time platform for monitoring and detecting mis/disinformation creators and spreaders. Key research topics and features required include but are not limited to: • A "digital narrative" that helps to present a timeline on how a piece of mis/disinformation unfolds in the various platforms on the Internet • Identify the source and the major spreaders if it is misinformation, given suspicious content identified by the user • Integrate existing solutions to build a workable prototype • Potential research and development on platform-specific malicious bot/account detection Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation report, data that present the User Acceptance, and other relevant metrics.
Research Theme 2	**EARLY DETECTION**
Tech Topic 3 Multimodal Deepfakes Detector	Deepfake detection has attracted much attention with increasing complexity due to the rapid evolvement of techniques and the threat landscape. Existing deep fake solutions primarily focus on detecting manipulation and fabrication surrounding human faces and voices. The objective is to build a robust deep fake detector that detects audio-visual content manipulation. The content manipulation can be in (but is not limited to) facial/non-facial region, speaker/voice manipulation, lip-synching, etc. Key research features shall consider one or more of the following: • Manipulation detection related to well-known public figures (politicians, actors/actresses, CEOs, etc.). As they are public figures, they are frequently subject to vicious deepfake attacks • Manipulation detection for the non-facial region

	• Prevention for adversarial attacks on the deepfake detection • Cross-modality manipulation (e.g., the use of a non-manipulated image but with a deceiving claim via text) • Discern malicious vs. non-malicious intent behind the author who posts the content Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation report and data that present the AUC, F1, Accuracy, Processing Speed (latency, e.g., 9 seconds), User Acceptance, and other relevant metrics when evaluated with realistic data, and compared against existing state-of-the-art solutions. Model generalizability and explainability should be considered.
Tech Topic 4 Non-Factual Claims Detector	Manual fact-checking is time-consuming, facing operational challenges in scalability and comprehensive coverage in claim verification. Much research has been conducted, and datasets have been developed for automatic fact-checking. However, fact-checking is a complicated process. Claims can be found in different media (text, audio or video) and platforms (website, social media, broadcast). As the checking process is not straightforward, it lacks close collaboration between human fact-checking and automatic platforms. We believe using AI technologies to assist a human expert in the different steps of fact-checking to verify the claim as early as possible is critical in combating online harms. This track aims to solicit excellent research and development efforts in fact-checking from different RIs, IHLs and companies. It also integrates the multiple steps of fact-checking in a single framework to provide people with technologies that help them verify or debunk viral misinformation. The technologies to be developed include, but are not limited to, identifying claims worth fact-checking, retrieving relevant evidence and related claims to fact-check a claim, and verifying a claim, aiming to build up some of the following capabilities. • Multimodal and multilingual evidence and related claims extraction. To extract multimodal and multilingual external evidence and associated claims to help fact-checkers decide the factuality of a given claim • Human-assisted multimodal multilingual fact verification. To incorporate both explainable and non-explainable approaches to assist fact-checkers in their verification Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation reports and data that present F1, Precision, Recall, AUC, and other relevant metrics. Model generalisability and explainability should be considered.

Tech Topic 5 Propaganda Detector	Propaganda is the “deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behaviour to achieve a response that furthers the desired intent of the propagandist”. The impact can be severe if the propagandists' targeted parties do not have the tools to quickly identify and act upon propaganda that is targeted to undermine their agenda. For the general public, an accurate, robust and explainable propaganda detector engine can also be a potentially valuable media literacy and critical thinking training tool to defend from intentional disinformation. The technological approach to propaganda is niched and nascent, with a high prospect for innovation and research opportunities. Key research topics and features required for this tech should include but are not limited to: • General-purpose propaganda detection • Technical-level detection including, but not limited to, loaded language, name-calling & labelling, flag-waving, sowing of doubt • Propaganda data curation and detection for different users or user group profiles • Propaganda detection engines that are trained, validated and tested to be robust for local content Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation reports and data that present F1, Precision, Recall, AUC, and other relevant metrics. Model generalisability and explainability should be considered.
Tech Topic 6 Extremism & Hate Detector	Social networking platforms allow users to publish and share content instantly, with or without intention and deliberation about the possible consequences. The lack of editorial oversight and the sheer amount of content makes it challenging to control the content that may cause harm to an individual, group, or society. For example, Cardiff University’s HateLab project found that increased online hate speech can increase the number of physical-world crimes committed against minorities. Social media platforms have recently realised the seriousness of online hate speech and have introduced strict policies to address this problem. However, given the volume of data and the lack of understanding of local culture/dynamics, it becomes complicated for these platforms to remove such content promptly. The objective is to automatically detect online extremism and hate speech from social media text/audio that encourages hate or violence towards a person/group based on their race,

	religion, ethnic origin, age, gender, sexual orientation, and/or disability. Key research topics and features required for this tech should include but are not limited to: • Technology that takes into account the difference between global vs. local values and concerns, such as racial and religious harmony, that are of particular need for Singapore and the region • Hate speech classification system from online text, audio • Explore video hate speech • Specific to racism, sexism, ageism etc., which can cause substantial harm to society and brands Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation report and data presenting F1, Precision, Recall, AUC, and other relevant metrics. Model generalisability and explainability should be considered.
Tech Topic 7 Malicious Account/Bot Detector	Social bots are social media accounts controlled, at least in part, through software. It allows a single entity (or a group of entities) to control many accounts to carry out autonomous actions like reply/post/follow, etc., based on triggers or scripted patterns. Such an entity may easily mix automatic and manual behaviours to avoid detection or be accused of being inauthentic. Malicious social bots have been used to manipulate social media users by amplifying misinformation or disinformation, committing financial fraud, suppressing or disrupting speech, spreading malware or spam, trolling/attacking victims, and other types of abuse. Key research topics and features required for this tech should include but are not limited to: • Robust semi-autonomous social bot detector applicable to multiple social media platforms • Multilingual and multimodal social bot detector • Technology to detect large-scale, coordinated botnet activities controlled by a group of common users, to amplify mis/disinformation Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation reports and data that present F1, Precision, Recall, AUC, and other relevant metrics. Model generalisability and explainability should be considered.
Research Theme 3	**MITIGATION & EMPOWERMENT**
Tech Topic 8	Debunking mis/disinformation with the general public, even when scientific and objective evidence or facts and evidence-

Tools to Improve Debunking Effectiveness	based narratives are available, is highly challenging for many reasons. According to industry feedback and suggestions, debunking news tended to be far less consumed than mis/disinformation. Technological tools that can make debunking and counter-mis/disinformation messaging more consumed and more effective will be handy and benefit public agencies as well as media companies. Key research topics and features required include but are not limited to: • Messaging evaluation functions such as alerts on creditability, clarity, vividness, emotional connectedness/empathy, confidence, agency and inclusiveness to assist communication professionals and practitioners in devising more effective debunking strategies • Efficacy studies of programmes that train and utilize both digital and human fact-checkers. Topics can be built around assessing the comparative effects of expertise versus crowdsourcing – e.g. panels of experts combined with everyday persons for fact-checking and information verification • Connecting science with the technological means of communication in crisis and uncertainty Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation reports and data that present the User Acceptance, and other relevant metrics.
Tech Topic 9 Online Trust & Safety Policy Testing System	Online harms take many forms and are particularly unevenly impacting certain vulnerable groups, such as children, people who are less tech-savvy, and underrepresented minorities. Social media platform companies rely on human expertise and local context input on content moderation. This is a costly process, and exposing humans to troubling, harmful content is often unethical and unsustainable. Developing a Singapore local context-aware policy testing system built on personas will be advantageous to allow policy managers, moderation leads, and operations teams to generate, test, and deploy localised community-specific inputs for Trust & Safety policy decision-making. Efforts such as the “Undersight Board” are a good example of this attempt. Key research topics and features required include but are not limited to: • Ability to build personas that are representative of Singapore's local communities and contexts • Scientifically grounded components that support policy users to generate, simulate, and evaluate the impact of harms and Online Trust & Safety policies

	Technology solutions or outcomes from this tech topic are expected to be delivered at least TRL 5, with validation report, data that present the User Acceptance, and other relevant metrics.
Tech Topic 10 Tools to Enhance Media Literacy	Past studies have found media literacy to be the most effective at safeguarding people against online threats. Therefore, research is needed to understand the variety of media literacy programmes in Singapore and how they target various types of online threats. This research area should aim to understand the scope and focus of existing programmes as they target the various population groups. Topics for research can include the following: • What are the effects and the limits of these programmes as they pertain to new and emerging threats, such as deepfakes, AI-generated misinformation, etc. • How do such programmes reach and enable the safeguarding of population groups? Where are the likely gaps in reaching specific vulnerable populations? Evaluation will focus on the extent to which the study and tools advance the understanding of our population's vulnerabilities from existing studies and examine the effectiveness through Efficacy studies, and the extent to which such media literacy enhancement tools can increase “psychological defence" by educating citizens effectively and inoculating citizens against the effects of online falsehood and harms.

Related references/resources:

Prevention

[1] Harrison, A. (2017, August 6). Can you trust the mainstream media?, The Guardian. https://www.theguardian.com/media/2017/aug/06/can-you-trust-mainstream-media

[2] Kim Andersen, Adam Shehata & Dennis Andersson (2021). Alternative News Orientation and Trust in Mainstream Media: A Longitudinal Audience Perspective, Digital Journalism, https://doi.org/10.1080/21670811.2021.1986412

[3] Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. Journal of Applied Psychology, 87(4), 611–628. https://doi.org/10.1037/0021-9010.87.4.611

[4] K.C. Toth, A. Anderson-Priddy, (2019). Self-sovereign digital identity: A paradigm shift for identity, IEEE Secur. & Priv. 17 (3) (2019) 17–27. https://doi.org/10.1109/MSEC.2018.2888782

[5] Y. Liu, D. He, M.S. Obaidat, N. Kumar, M.K. Khan, K.K.R. Choo, (2020). Blockchain-based identity management systems: a review, J. Netw. Comput. Appl. 166 (2020) 1–11, 102731. https://doi.org/10.1016/j.jnca.2020.102731

[6] Chaban A.V. (2020, November 30) Can blockchain block fake news and deep fakes? IBM Blogs, https://www.ibm.com/blogs/industries/blockchain-protection-fake-news-deep-fakes-safe-press/

[7] Harrison K. & Leopold A. (2021, July 19). How Blockchain Can Help Combat Disinformation https://hbr.org/2021/07/how-blockchain-can-help-combat-disinformation

[8] Ruan P, Dinh T, Lin Q, Zhang M, Chen G and Ooi B, (2021). LineageChain: a fine-grained, secure and efficient data provenance system for blockchains, The VLDB Journal, 30:1. https://doi.org/10.1007/s00778-020-00646-1

[9] Wang, Z., Byrnes, O., Wang, H., Sun, R., Ma, C., Chen, H., ... & Xue, M. (2023). Data Hiding With Deep Learning: A Survey Unifying Digital Watermarking and Steganography. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2023.3268950

[10] Y. Wang, Y. Pan, M. Yan, Z. Su and T. H. Luan. (2023). A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions. IEEE Open Journal of the Computer Society, https://doi.org/10.1109/OJCS.2023.3300321

Early Detection

[11] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J. (May 2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. https://arxiv.org/abs/2001.00179

[12] Juefei-Xu, F., Wang, R., Huang , Y., Guo, Q., Ma, L., Liu, Y. (March, 2022). Countering Malicious DeepFakes: Survey, Battleground, and Horizon. https://arxiv.org/abs/2103.00218

[13] Liu, P., Lin, Y.,He, Y., Wei, Y., Zhen, L., Zhou, J.T. et al. (2021). Automated deepfake detection. ArXiv: https://arxiv.org/pdf/2106.10705.pdf

[14] Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., ... & Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. arXiv preprint arXiv:2103.07769.https://doi.org/10.24963/ijcai.2021/619

[15] Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 10, 178-206. https://doi.org/10.1162/tacl_a_00454

[16] Vo, N., & Lee, K. (2018, June). The rise of guardians: Fact-checking url recommendation to combat fake news. In The 41st international ACM SIGIR conference on research & development in information retrieval (pp. 275-284). https://doi.org/10.1145/3209978.3210037

[17] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Preslav Nakov. (2020). A Survey on Computational Propaganda Detection, July 2020, IJCAI. https://doi.org/10.24963/ijcai.2020/664

[18] Krishnamurthy, G., Gupta, R.K., and Yang, Y. (2020). SocCogCom at SemEval-2020 Task 11: Characterising and detecting propaganda using sentence-level emotional salience features, Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval 2020), pp. 1973-1801, 12-13 Dec 2020, Barcelona, Spain. https://aclanthology.org/2020.semeval-1.235

[19] Williams, M.L., Burnap, P., Javed, A., Liu, H. and Ozalp, S. (2020). Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. The British Journal of Criminology, Vol. 60(1), pp. 93-117. https://doi.org/10.1093/bjc/azz064

[20] Fischer, A., Halperin, E., Canetti, D., & Jasini, A. (2018). Why We Hate. Emotion Review, 10(4), 309–320. https://doi.org/10.1177/1754073917751229

[21] Uyheng, J., Carley, K.M. Characterizing network dynamics of online hate communities around the COVID-19 pandemic. Appl Netw Sci 6, 20 (2021). https://doi.org/10.1007/s41109-021-00362-x

[22] Yang , K.C., Varol, O., Hui, P.M., Menczer, F. (2020). Scalable and Generalizable Social Bot Detection through Data Selection. AAAI 2020. https://arxiv.org/abs/1911.09179

[23] Lynnette Hui XianNgDawn C.RobertsonKathleen M.Carley, (2022). Stabilising a supervised bot detection algorithm: How much data is needed for consistent predictions? Online Social Networks and Medi, https://doi.org/10.1016/j.osnem.2022.100198

[24] Botometer^{R}, Observatory on Social Media (OSoMe) and the Network Science Institute (IUNI) at Indiana University, Available: https://botometer.osome.iu.edu

Mitigation & Empowerment

[25] Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest. 2012;13(3):106-131. https://doi.org/10.1177/1529100612451018

[26] Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online, Scinece. 359(6380). pp. 1146-1151. https://doi.org/10.1126/science.aap9559

[27] Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. Psychological Science, 28(11), 1531–1546. https://doi.org/10.1177/0956797617714579

[28] Katherine Kricorian, Rachel Civen & Ozlem Equils (2022) COVID-19 vaccine hesitancy: misinformation and perceptions of vaccine safety, Human Vaccines & Immunotherapeutics, 18:1, https://doi.org/10.1080/21645515.2021.1950504

[29] World Economic Forum. (August 2023). Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

[30] Undersight Board: Ensuring respect for free expression, through generative judgment. https://undersightboard.github.io

[31] Lorig, F., Johansson, E., & Davidsson, P. (2021). Agent-based social simulation of the COVID-19 pandemic: A systematic review. JASSS: Journal of Artificial Societies and Social Simulation, 24(3). https://doi.org/10.18564/jasss.4601

[32] Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., & Silva, R. C. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. Chaos, Solitons & Fractals, 139, 110088. https://doi.org/10.1016/j.chaos.2020.110088