Submission to the UN Special Rapporteur on disinformation and freedom of opinion and expression

Albert Zhang, Ariel Bogle and Dr Jacob Wallis

I. Introduction

Albert Zhang is a researcher, Ariel Bogle is an analyst and Dr Jacob Wallis is a senior analyst at the Australian Strategic Policy Institute's (ASPI) International Cyber Policy Centre (ICPC). ASPI does not take corporate positions on any issue, and the views expressed here are our personal opinions.

We would like to thank the UN Special Rapporteur Ms. Irene Khan for the opportunity to assist her in the preparation of a report on disinformation and the right to freedom of opinion and expression to the Human Rights Council.

Disinformation campaigns are a growing threat to global stability and democractic values, but in some countries, laws ostensibly aimed at countering such activities have been used to crack down on journalists and civil society groups. We are concerned by the increase in dissemination of disinformation by state and non-state actors in pursuit of financial, ideological and political goals. Manipulation of the information environment through the propagation of disinformation risks constraining the space available to democratic stakeholders, and particularly to marginalised groups, for authentic political expression.

II. Key challenges raised by disinformation

According to an Oxford Internet Institute report analysing recent trends in online propaganda, governments, political parties and state-affiliated actors attempted to manipulate public opinion online in 81 countries in 2020, up from 70 countries in 2019.[1] While much of the focus has been on disinformation efforts by nation states, ideologically-driven groups and other non-state actors play a significant role. As do emerging networks of public relation firms and influence-for-hire operations. In addition, the coronavirus pandemic has exacerbated existing social tensions and created many challenges for social cohesion. In some countries, people are increasingly working online, consuming digital media and interacting on social media, creating new vulnerabilities in a time of stress and disruption.

The study of disinformation is a growing field. Civil society organisations and academic researchers have done significant work identifying the techniques and impacts of malign online information operations, while governments and technology platforms are increasingly concerned with how such activities should be addressed and regulated. A robust ecosystem of global research is necessary to ensure the impact of disinformation and related laws on freedom of expression is properly scrutinised in an independent and sustainable manner. Carnegie's Partnership for Countering Influence Operations (PCIO) surveyed the influence operations research community in 2020 to identify some challenges faced by this emerging field.[2] In this submission, we focus on three key challenges, which were also identified by PCIO:

1. Lack of definitions and understanding
2. Lack of sustainable business model
3. Lack of access to data and transparency

A. Lack of definitions and understanding

There are many definitions provided for 'disinformation' or 'fake news' presented globally in legislation and by researchers, but a common consensus and nuanced understanding of these terms and related ideas remains elusive. Often different interpretations of disinformation manifest in specific contexts - disinformation during an election period, for example. This has resulted in the mislabelling

[1] Samantha Bradshaw, Hannah Bailey & Philip N. Howard, 'Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation', Working Paper 2021.1. Oxford, UK: Project on Computational Propaganda, 13 January 2021, https://comprop.oii.ox.ac.uk/research/posts/industrialized-disinformation/

[2] Victoria Smith & Natalie Thompson, 'Survey on Countering Influence Operations Highlights Steep Challenges, Great Opportunities', Carnegie Endowment for International Peace, 7 December 2020, https://carnegieendowment.org/2020/12/07/survey-on-countering-influence-operations-highlights-steep-challenges-great-opportunities-pub-83370

of different types of information operations, and in some cases, a very Western-centric approach. Some interpretations have included:

- The European Union (EU) Code of Practice on Disinformation defines disinformation as 'verifiably false or misleading information' which, cumulatively, (a) 'is created, presented and disseminated for economic gain or to intentionally deceive the public'; and (b) 'may cause public harm intended as threats to democratic political and policy making processes as well as public goods such as the protection of EU citizens' health, the environment or security'.[3]
- France, among the first countries to legislate a definition of 'fake news', refers to 'inexact allegations or imputations, or news that falsely report facts, with the aim of changing the sincerity of a vote'.[4]
- In 2020, the Australian Communication and Media Authority referred to disinformation as 'false and misleading information distributed by malicious actors with the intent to cause harm to individual users and the broader community' in its position paper to guide a code of conduct for technology platforms in Australia.[5]
- Instead of using 'disinformation', Singapore's Protection from Online Falsehoods and Manipulation Act 2019 refers to 'false statements of fact': 'A statement of fact is a statement which a reasonable person seeing, hearing or otherwise perceiving it would consider to be a representation of fact'.[6]

The labelling of disinformation is a challenge for governments, researchers and civil society organisations, especially when the veracity of content is difficult to verify and the intent is unclear. When China's foreign ministry spokesman Zhao Lijian posted an image on Twitter of an Australian soldier holding a knife to a child in reference to the 2020 Australian war crimes inquiry, for example, Australia's Foreign Minister Marise Payne reportedly referred to the image as 'the most 'egregious' example of social media disinformation she has ever witnessed'.[7] The image was an artwork, however,

[3] European Commission, Code of Practice on Disinformation, https://ec.europa.eu/digital-single-market/en/code-practice-disinformation

[4] PROPOSITION DE LO: relative à la lutte contre la manipulation de l'information, 20 November 2018, http://www.assemblee-nationale.fr/15/ta/tap0190.pdf

[5] Australian Communications and Media Authority, 'Misinformation and news quality on digital platforms in Australia: A position paper to guide code development', June 2020, https://www.acma.gov.au/sites/default/files/2020-06/Misinformation%20and%20news%20quality%20position%20paper.pdf

[6] Defined under Section 2(2) of the Protection from Online Falsehoods and Manipulation Act 2019 https://sso.agc.gov.sg/Act/POFMA2019#pr2-

[7] Stephen Dziedzic & Jane Norman, 'Scott Morrison demands apology from China over 'repugnant' tweet showing Australian soldier threatening to kill child', *ABC News*, 30 November 2020https://www.abc.net.au/news/2020-11-30/china-fake-image-australian-war-crimes-afghanistan-tensions/12934538

and was arguably not designed to deceive viewers of this fact. The killing of Afghan children by Australian soldiers during the war in Afghanistan remained an unproven allegation.[8] The image may not have been classified as disinformation within a narrow definition, and could be viewed more generally as state propaganda. This incident highlights the definitional complexity of categorising disinformation within rapidly evolving and often high-profile diplomatic discourse.

There is also a risk that ambiguous definitions may disproportionately burden the freedom of opinion and expression by censuring dissenting or disagreeable views. In some cases, new laws penalising and criminalising disinformation have allowed journalists to be targeted for expressing critical views of the government. In Bangladesh, for example, journalists have raised concerns about the potential for their work to be criminalised by laws that included prison terms for those who spread 'propaganda'.[9] Three journalists were also arrested in 2018 in Myanmar and later released for publishing stories critical of the Yangon regional government.[10] The triple were reportedly charged under Section 505(b) of the Myanmar Penal Code, which doesn't distinguish between false and true information and instead makes it an offence to make, publish or circulate any statement, rumour or report 'with the intent to cause, or which is likely to cause, fear or alarm to the public or to any section of the public whereby any person may be induced to commit an offence against the State or against the public tranquility'.[11]

On the other hand, a narrow focus on disinformation may lead policymakers to insufficiently address a spectrum of other malicious online activity. In 2020, ASPI ICPC's *Automating influence on Covid-19* report investigated Chinese-speaking actors that were broadly aligned with the political goals of the People's Republic of China. They had automated the sharing of articles on Facebook criticising the United State's handling of the coronavirus pandemic and scandals linked to then President Donald Trump.[12] The amplification was designed to mislead users of the popularity of the stories and emphasised racial divisions in the United States. The sources of these stories were, however, credible

[8] Georgia Hitch, 'What war crimes did Australian soldiers commit in Afghanistan and will anyone go to jail?', *ABC News*, 19 November 2020, https://www.abc.net.au/news/2020-11-19/afghan-war-crimes-report-released-what-you-need-to-know/12899880

[9] The Economist, 'Bangladesh's slide towards authoritarianism is accelerating', 4 October 2018, https://www.economist.com/asia/2018/10/04/bangladeshs-slide-towards-authoritarianism-is-accelerating?frsc=dg%7Ce

[10] Al Jazeera, 'Myanmar arrests three journalists over article on Suu Kyi protege', 10 October 2018, https://www.aljazeera.com/news/2018/10/10/myanmar-arrests-three-journalists-over-article-on-suu-kyi-protege

[11] Myanmar: The Penal Code, Online Burma/Myanmar Library, https://www.burmalibrary.org/sites/burmalibrary.org/files/obl/docs6/MYANMAR_PENAL_CODE-corr.1.pdf

[12] Elise Thomas, Albert Zhang & Dr Jake Wallis, 'Automating influence on Covid-19', ASPI ICPC, August 2020, https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2020-08/Automating%20influence%20on%20Covid-19.pdf?DxaB4psM9BvTNrhNQNTpu_jWNWmqPGXg=

news outlets so the information was not necessarily false. This activity may be more appropriately classified as social media manipulation, as opposed to disinformation.

B. Lack of sustainable business model

Currently there are limited sustainable business models for the analysis of online campaigns of disinformation and social media manipulation. This work is vital to drive transparency around actors and impact, as well as to inform policy decision-making and provide a level of deterrence.

A range of civil society entities partner with the major social media platforms to provide analysis of information operation datasets prior to public release. The opportunity for independent analysis is a valuable approach in that it ensures there is expert-led analysis in the public domain to inform media commentary and public understanding. However, there is no clear business model underpinning this work. A small number of predominantly American and European think tanks, research institutes and corporate entities - such as Atlantic Council's Digital Forensic Research Lab, Stanford Internet Observatory, Graphika and the Australian Strategic Policy Institute - undertake the bulk of disinformation analysis driven by direct access to data from the platforms.

A complex set of skills is required to undertake this analysis and there is enormous potential to build capacity in parts of the world where the major social media platforms are less focused and civil society is under threat. Emerging and fragile democracies in particular must strengthen their capacity to detect, analyse and respond to disinformation campaigns. There are particular challenges in this context, given civil society organisations in some regions can face threats from regimes who may themselves be deploying disinformation campaigns targeted at their own domestic populations. It is also vital that work be supported that engages directly with affected communities to provide solutions that work in specific cultural and political contexts.

The potential implications of disrupted electoral events or manipulated political discourse are becoming increasingly apparent, even in contexts where there is a long history of free and fair elections, and protections of the right to free speech.

C. Lack of access to data and transparency

Access to data remains a major barrier for disinformation scholarship, and diminishes our understanding of impact on freedom of expression and other issues. For large-scale influence campaigns, manual inspection of potentially misleading accounts and content is mostly unfeasible. Currently, researchers typically rely either on computational data collection methods or datasets released by platforms like Facebook and Twitter. These datasets may contain information on text, number of interactions, hashtags, topics, images, videos, audio files, creation dates of posts or accounts, and other relevant data.

One method of collecting real-time data from social media platforms is through their application programming interfaces (APIs), which allow researchers to automate the request for data. Although, most platforms have limited their APIs and some companies have removed useful features to prevent exploitation. Facebook restricted its API following the Cambridge Analytica controversy to mitigate microtargeting of its users.[13] However, the effectiveness of this initiative remains unknown as companies can still solicit some types of data from a growing private data-broker industry, while researchers studying disinformation are priced out of commercial products or may have ethical reasons not to use them. There are open-source tools that assist with data collection but they can create biases and may only capture small portions of the overall activity.

Compared to the industry as a whole, Twitter is a leader in the maintenance of a public archive of data sets derived from state-linked information operations. This archive offers the most comprehensive access to state-attributed data in the public domain. In contrast, Facebook will provide datasets of state-linked information operations on its platform to select think tanks for analysis in advance of public disclosure.

When disinformation operations and inauthentic coordinated campaigns are detected online, social media platforms may take appropriate measures to delete, contain and block content, but this evidence may also be important for authorities to obtain. Facebook initially resisted pressure at the International Court of Justice to obtain posts and communications by members of Myanmar's military and police, for example, who were accused of spreading hate speech about the Rohingya population and by some parties, of genocide.[14] The UN's Independent Investigative Mechanism on Myanmar was eventually sent a 'first data set which partially complie[d] with [their] previous requests'. Archiving evidence of state-sponsored information operations will be essential to holding states to account to protect human rights online.

To store data for research purposes and also preserve evidence of disinformation campaigns, social media companies could also create a 'human-right locker' or a 'digital locker', as suggested by Joan Donovan, research director of the Shorenstein Center on Media Politics and Public Policy at the

[13] Axel Bruns, 'After the 'APIcalypse': social media platforms and their fight against critical scholarly research', *Information, Communication & Society*, 11 June 2019, https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1637447

[14] Poppy McPherson, 'Facebook shares data on Myanmar with United Nations investigators', *Reuters*, 26 August 2020, https://www.reuters.com/article/us-myanmar-facebook-idUSKBN25L2G4

Harvard Kennedy School, and Gabrielle Lim, a researcher with the Technology and Social Change Research Project.[15] The locker could contain data on accounts, posts, media and other content that was once public and involved in disinformation or influence operations. It would allow social media companies to delete misleading and harmful content from their platforms while sharing data to approved individuals and organisations to research and investigate.

Partnerships between academics, think tanks and industry have been a good start, but more can be done. We encourage digital technology platforms to not only archive and publicly release datasets, but also invest in efforts to create interactive environments for researchers with non-technical skills to explore data sets. The size of some takedown datasets may also exceed terabytes due to accompanying videos and images and due to a lack of computational resources, many researchers may be prevented from analyzing these data. Archived data should then be designed to be accessible and navigable for non-technical people.

While granting researchers greater access to data, we also recognise the ethical considerations in protecting the right to individual privacy, among other issues. Codes and standards of practice should be agreed upon by governments, academica and civil society of how datasets are archived, shared and published. It is also important that such work remain independent, and that social media companies are not able to control or place limitations on the work of researchers in this area, beyond what is needed to preserve user privacy.

III. Government Policies

Governments are responding to the spread of online disinformation, but approaches vary. Like many other countries, Australia has approached the spread of disinformation and other undesirable content online with a variety of legislative and regulatory solutions, creating an arguably complex landscape for stakeholders. After the 2019 Christchurch mosque shootings where the perpetrator live streamed the attack on Facebook, the federal government passed laws that penalised providers of online services or content for failing to report and remove “abhorrent violent material”.[16] Further, in December 2020, a draft online safety bill was released for consultation.[17] This legislation would give Australia’s e-safety

[15] Joan Donovan & Gabrielle Lim, ‘The Internet Is a Crime Scene’, *Foreign Policy*, 20 January 2021, https://foreignpolicy.com/2021/01/20/internet-crime-scene-capitol-riot-data-information-governance/

[16] Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019, https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=s1201

[17] Department of Infrastructure, Transport, Regional Development and Communications, Consultation on a Bill for a new Online Safety Act, 23 December 2020, https://www.communications.gov.au/have-your-say/consultation-bill-new-online-safety-act

commissioner greater powers to order the take-down of harmful content, although some critics argue parts of the proposal are vague and too far reaching.[18]

Another emerging component of Australia's online content framework is a voluntary code of practice for online disinformation. This emerged from the Australian Competition and Consumer's Digital Platforms Inquiry, which recommended a code to address complaints about disinformation as well as new rules that would force Facebook and Google to pay media outlets for the use of their content.[19] The government asked the major digital platforms to develop the voluntary code, overseen by Australia's media regulator.[20] In late 2020, however, the Australian Communications and Media Authority reportedly complained the draft code produced by DIGI and other groups[21] did not meet expectations and consultation is continuing.[22]

The European Union (EU) has implemented a Code of Practice on Disinformation and this in some ways serves as a model for the Australian regulatory environment, however the EU's latest assessment of the performance of the code suggests that there is room to improve. The EU's position is that the implementation of the Code has produced positive outcomes in terms of enhancing transparency and accountability, yet the assessment[23] also found:

- "the absence of relevant key performance indicators (KPIs) to assess the effectiveness of platforms' policies to counter the phenomenon;
- the lack of clearer procedures, commonly shared definition and more precise commitments;
- the lack of access to data allowing for an independent evaluation of emerging trends and threats posed by online disinformation;

[18] Samantha Floreani, 'Proposed online 'safety' measures may do more harm than good', *The Canberra Times*, 15 February 2021, https://www.canberratimes.com.au/story/7124518/proposed-online-safety-measures-may-do-more-harm-than-good/?cs=14264

[19] Australian Competition & Consumer Commission, 'Digital platforms inquiry - final report', 26 July 2019, https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report

[20] Australian Communications and Media Authority, 'Australian voluntary code(s) of practice for online disinformation', 20 October 2020, https://www.acma.gov.au/australian-voluntary-codes-practice-online-disinformation

[21] 'Consultation on disinformation industry code', DIGI, 19 October 2020, https://digi.org.au/disinformation-code/

[22] Zoe Samios & Lisa Visentin, 'ACMA: Tech giants' code to handle fake news fails to meet expectations', *The Sydney Morning Herald*, 26 October 2020, https://www.smh.com.au/politics/federal/acma-tech-giants-code-to-handle-fake-news-fails-to-meet-expectations-20201026-p568oq.html

[23] European Commission (2020) Assessment of the Code of Practice on Disinformation – Achievements and areas for further improvement, https://ec.europa.eu/digital-single-market/en/news/assessment-code-practice-disinformation-achievements-and-areas-further-improvement

- missing structured cooperation between platforms and the research community;
- the need to involve other relevant stakeholders, in particular from the advertising sector."

Australia's own Indo-Pacific region is made up of established and emerging democracies. It is also a region of geo-political rivalry, domestic political misinformation and a shadow marketplace of influence-for-hire services. There are examples of innovative policy-making in the region in response to disinformation. There are also examples of responses that are politically opportunistic, and vague in terms of policy implementation due to the definitional challenges that we highlight above. Taiwan faces a persistent campaign of political warfare from the People's Republic of China that incorporates campaigns of disinformation, foreign interference and economic coercion. According to the V-Dem Institute[24], Taiwan faces more disinformation than almost any other country in the world. The Taiwanese government will act to counter-disinformation directly but favours a devolved and collaborative model that supports whole-of-society engagement with, and resilience to, the threat posed by covert disinformation campaigns linked to the People's Republic of China.[25]

Other states have passed legislative responses that threaten the healthy functioning of pluralist civil society. Some critics allege fake news legislation introduced in Singapore risks criminalising freedom of expression[26] and similar legislation in Malaysia[27] has recently been repealed. A series of legislative bills introduced in the Philippines,[28] as well as politically motivated yet commercially outsourced online trolling, inhibit the healthy functioning of Filippino civil society and constrain willingness to express views critical of the government. After examining recent elections in Southeast Asia, Jonathan Corpus Ong, an associate professor of global digital media at the University of Massachusetts, wrote:

"As social media becomes more central in political campaigning, especially in the pandemic moment, we should be alert to how politicians may use fears of the "infodemic" to introduce regulations with vaguely phrased definitions of fake news. Such government regulations deepen cultures of self-censorship, facilitate

[24] V-Dem Institute, 'Democracy Facing Global Challenges', 21 May 2019, https://www.v-dem.net/media/filer_public/99/de/99dedd73-f8bc-484c-8b91-44ba601b6e6b/v-dem_democracy_report_2019.pdf

[25] Flemming Rose, 'The Taiwan Election: Dealing with Disinformation while Protecting Speech', *Cato Institute blog*, 7 February 2020, https://www.cato.org/blog/taiwan-election-dealing-disinformation-while-protecting-speech

[26] Freedom House, 'Freedom on the net 2020: Singapore', https://freedomhouse.org/country/singapore/freedom-net/2020

[27] Freedom House, 'Freedom on the net 2020: Malaysia', https://freedomhouse.org/country/malaysia/freedom-net/2020

[28] Freedom House, 'Freedom on the net 2020: Philippines', https://freedomhouse.org/country/philippines/freedom-net/2020

a growing lack of trust in politicians, reduce transparency in governance, inhibit dissent and exacerbate long-existing inequalities in political participation."[29]

IV. Recommendations

The UN will necessarily decide on definitions for key terms in this space, including disinformation. In the formulation of these definitions, we recommend a globally representative selection of stakeholders are included to avoid a limited and Western-centric framing. This would support the UN by framing stakeholder dialogue and engagement in ways that promote the right to freedom of opinion and expression around the world.

We recommend the UN use its influence to promote greater coordination and collaboration across industry, civil society, academia and government. This would enhance empirical research to inform policy-making on the impact and effects of disinformation, provide an environment to define fundamental terms and agree on standard metrics by which to measure the impact of disinformation campaigns.

We recommend the development of a mechanism for researchers that would allow the analysis of data related to disinformation campaigns in context, even when the related content has been removed from public view. This environment would support the analysis of both data and platform algorithms. Algorithms are fundamental to both the amplification and suppression of content. This data preservation and sharing tool could be used to drive transparency around how the algorithms that surface content in certain social media environments influence perception. Furthermore, we suggest that alongside the development of tools and analytic environments, the UN funds capacity building work that will upskill civil society organisations in parts of the world that are under-represented in the analysis of disinformation campaigns.

[29] Jonathan Corpus Ong, 'How "Fake News" Regulations Can Serve Political Incumbents', Centre for International Governance Innovation, 2 November 2020, https://www.cigionline.org/articles/how-fake-news-regulations-can-serve-political-incumbents