National Council of Women of New Zealand
Te Kaunihera Wahine o Aotearoa

GENDER EQUAL NZ
MAKING EQUALITY REALITY

office@ncwnz.org.nz
www.ncwnz.org.nz

C/o PSA
PO Box 3817
WELLINGTON 6140
Attn: NCWNZ

31 August 2023

S23.13

Consultation Document: Safer Online Services and Media Platforms Submission to the Department of Internal Affairs

Introduction

1. The National Council of Women of New Zealand, Te Kaunihera Wāhine o Aotearoa (NCWNZ) is an umbrella group representing around 60 affiliated organisations and 300 individual members. Collectively our reach is over 200,000 with many of our membership organisations representing all genders. NCWNZ has 13 branches across the country.
2. NCWNZ's vision is a gender equal New Zealand and research shows we will be better off socially and economically if we are gender equal. Through research, discussion and action, NCWNZ in partnership with others, seeks to realise its vision of gender equality because it is a basic human right.
3. This submission has been prepared in consultation with the membership and in particular, with three of NCWNZ's Action Hubs:
    - International;
    - Safety, Health and Wellbeing; and
    - Influence, Impact and Decision-making.
4. In writing this submission, we have also drawn on NCWNZ's long history of commitment to eliminating all forms of violence against women and children, and striving for a gender equal Aotearoa. We understand the links between the social and cultural conditioning that drive sexism, gender discrimination, racism, xenophobia and violence against women and children, and wider violence and conflict both domestically and on the international stage. In this submission, NCWNZ is providing a gender lens to the review, which appears to be nominally included in the discussion document.
5. On 4 August 2023, we invited 25 national NGOs (representing many thousands of members across the country) to a meeting to discuss the issues and heard how

widespread and shared the concern was about harm in the online environment. We have shared a copy of this submission with them, and we know that many had earlier written to the department with their concerns as part of your review process.

6. Students against Sexual Harm (SASH) specifically asked to include the following comment in this submission:

> "We fully support the ideas and issues considered in this document. SASH is especially pleased with the depth in which a definition is given to unsafe and harmful content as we believe the incorrect categorisation of this leads to online harm. We are also pleased with the research that has been put into spotlighting the attacks against women in politics perpetrated online. We concur with the report where it states that the proposed model of enforcing codes of practice will not be sufficient and needs to be stronger."

7. In addition, the NZPPTA wanted to highlight the impact that online harm has on both students and teachers. This can range from online hate campaigns, toxic content, unauthorised recording and sharing of videos from the classrooms and bathrooms, addiction to social media, and violent, self-harm and suicide content. Teachers are also concerned about long-term ramifications for young people that are sharing problematic content online at an age where they may not appreciate the long-term consequences or permanence of that content in cyberspace.

8. This submission has also been informed by listening to the national and international speakers at the recent conference hosted by Diplosphere[1]: Images of the Future: Daily Life in a World Governed by AI.

9. We are appreciative of Te Tari Taiwhenua (the Department of Internal Affairs) engagement with the public in this important review and welcome the opportunity to make a submission.

Summary and Core Recommendations

10. We welcome the review – regulation is both necessary and important.

11. Any Government that is serious about eliminating violence against women and children needs to regulate the online environment. UNFPA has provided useful guidance on the meaning of technology-facilitated violence with examples[2], which is a useful point for considering how women are impacted in different ways by the online environment and how abuse is facilitated, assisted, or achieved through technology.

[1] Diplosphere Conference 2023. Images of the Future: Daily Life in a World Governed by AI. https://www.diplosphere.org/conference

[2] UNFPA. 2023. Measuring Technology-Facilitated Gender-based Violence: A discussion paper, p. 5, 16. https://www.unfpa.org/publications/measuring-technology-facilitated-gender-based-violence-discussion-paper

12. We are concerned that an overly narrow focus on content regulation in this review is likely to lead to lost opportunities and system gaps. At a minimum, the new regulator needs to have the ability to conduct and commission research looking at broader causes and impacts of harm from the online environment, with the statutory right to present this information to Parliament and a process for the Government to respond to any recommendations (for example, in the same way that the Law Commission reports are tabled and there is a government response). This will be particularly important as new technology develops (for example, Artificial Intelligence and machine learning) and our understanding of technology expands (for example, through transparency reports and research).

13. We know that things that happen online don't just stay online, as we've seen so recently with tragedies both here in New Zealand and overseas, such as the March 15 terrorist attacks, the Colorado shooting after a wave of increased anti-LGBTQ+ hate speech, and the subculture of incel violent extremism leading to the rape and murder of women.

14. Children have particular vulnerabilities and sensitivities in the online environment that need to be recognised in legislation. It is important to note that they are not the only group affected by online harm and everyone should benefit from new laws regulating the online environment. In this submission we have highlighted some of the impacts that the online environment has on women, including how it has been weaponised, and the failure of technology companies under the current arrangements to ensure that their products and services are safe or responsive. We have also drawn attention to relevant international obligations.

15. Fundamentally, it is important to create a regulatory regime that builds and strengthens public trust, safety, and confidence in the online environment. Human rights and democracy (including democratic institutions, representatives, and elections) need to be protected – both online and offline. Similarly, the public has a right to be protected from harmful disinformation that can often escalate into violence against individuals and groups, and damages the public understanding and response to big issues like climate change and COVID-19.

16. We have made a number of recommendations in this submission. In particular, we strongly recommend that legislation be progressed in the areas of safety by design, transparency and accountability of technology companies (and their senior executives) to an independent media regulator and the courts. Only effective legislation, transparency requirements and accountability can change the incentive and disincentive structure and the business calculations that companies do when singling out engagement and data extraction as a single metric for success. In making this recommendation we have drawn upon the extensive research of the international NGO, the Center for Countering Digital

Hate (CCDH), which has developed a STAR Framework[3] that sets these components in more detail. THE STAR Framework has four core elements:

- Safety by design
- Transparency of algorithms, rules enforcement and economics (advertising)
- Accountability to an independent media regulator and the courts
- Responsibility of technology companies and senior executives.

17. We cannot stress enough how important this review is and why legislation is needed. We look forward to working with the department as it develops its proposals and throughout the legislative process.

Terminology

18. ***Community***: Is used as short-hand throughout this document to encompass members of the public, iwi/Māori, businesses, charities, community groups and other organisations.

19. ***Department***: refers to the Department of Internal Affairs.

International Conventions and Commitments

20. There are a number of international human rights conventions and commitments that New Zealand has made that we want to draw officials' attention to for the purposes of this review, which are useful to consider when weighing different interests and options for reform.

The Universal Declaration of Human Rights[4]

21. Amongst other things, this Declaration sets out a basic principle that all human beings are born free and equal in dignity and rights. Every second, this is undermined in the online environment where there is the proliferation of misogyny, hate and abuse, and harassment – which all too often escalates into violence.

Convention to Eliminate all Forms of Discrimination against Women (CEDAW) [5]

22. This international convention sets out the key ways and areas that the Government needs to focus on to eliminate discrimination against women. In particular, we wanted to draw your attention to the following articles, which are relevant to this review:

    Article 3: "States Parties condemn discrimination against women in all its forms."

    Article 4: "Adoption by States Parties of temporary special measures aimed at accelerating de facto equality between men and women shall not be considered discrimination as defined in the present Convention, but shall in no way entail as a consequence the maintenance of unequal or

[3] Center for Countering Digital Hate. 2022. STAR Framework: A Global Standard for Regulating Social Media. https://counterhate.com/research/star-framework/

[4] United Nations. 1948. Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights

[5] United Nations. 1979. Convention on the Elimination of All Forms of Discrimination against Women New York, 18 December 1979. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women

separate standards; these measures shall be discontinued when the objectives of equality of opportunity and treatment have been achieved".

Article 5(a): "States Parties shall take all appropriate measures: To modify the social and cultural patterns of conduct of men and women, with a view to achieving the elimination of prejudices and customary and all other practices which are based on the idea of the inferiority or the superiority of either of the sexes or on stereotyped roles for men and women";

Article 7: "States Parties shall take all appropriate measures to eliminate discrimination against women in the political and public life of the country and, in particular, shall ensure to women, on equal terms with men, the right:

a) To vote in all elections and public referenda and to be eligible for election to all publicly elected bodies;

b) To participate in the formulation of government policy and the implementation thereof and to hold public office and perform all public functions at all levels of government;

c) To participate in non-governmental organizations and associations concerned with the public and political life of the country."

Article 8: Representation - "States Parties shall take all appropriate measures to ensure to women, on equal terms with men and without any discrimination, the opportunity to represent their Governments at the international level and to participate in the work of international organisations."

Article 12(1): "States Parties shall take all appropriate measures to eliminate discrimination against women in the field of health care in order to ensure, on a basis of equality of men and women, access to health care services, including those related to family planning".

Article 13: "States Parties shall take all appropriate measures to eliminate discrimination against women in other areas of economic and social life in order to ensure, on a basis of equality of men and women, the same rights, in particular: (c): The right to participate in recreational activities, sports and all aspects of cultural life."

Note that New Zealand is awaiting its next session with the United Nations' Committee on the Elimination of Discrimination against Women (the CEDAW Committee) and how the Government is responding to online harm and the need to regulate the online environment is likely to be an issue raised by the National Council of Women of New Zealand in its alternative report to the Committee.

Commission on the Status of Women

23. This year's Commission on the Status of Women (CSW67) at the United Nations in New York was specifically focused on "Innovation and technological change, and education in the digital age for achieving gender equality and the empowerment of all women and girls". NCWNZ (as part of the Pacific Women's Watch delegation) fed into the development of the final conclusions[6] agreed to by Governments. There are a number of

[6] Commission on the Status of Women 67th session 2023. Innovation and technological change, and education in the digital age for achieving gender equality and the empowerment of all women and girls. https://documents-dds-ny.un.org/doc/UNDOC/LTD/N23/081/71/PDF/N2308171.pdf?OpenElement

conclusions from CSW67 that are highly pertinent to this review, which we have included as **Appendix A** of this submission.

Online Harm against Women

24. Online harm against women comes in many forms.

25. UN Women and the World Health Organisation have provided the following definition of "technology-facilitated gender-based violence" (TFGBV):

> "Technology-facilitated violence against women is any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms."[7]

26. The United Nations Population Fund (UNFPA) have elaborated on this definition with examples in their discussion paper *Measuring Technology-facilitated Gender-based Violence, released this year*[8].

27. The CSW67 conclusions in Appendix A outline a number of online harms against women, including (without limitation):

- how technology can be weaponised against women, resulting in a breach of human rights, privacy, abuse, harassment, violence and impacting women's right to freedom of expression, movement and representation;
- how algorithms and other technology in the online environment can perpetuate and create discrimination, stereotypes, and a loss of equal opportunities for women;
- the need to ensure that online products and services are safe through, for example testing and risk assessments, and recognising how this may impact on everything from access to healthcare, to democratic institutions, to education and employment opportunities;
- the need for law and regulations that promote algorithmic and other forms of transparency;
- the need for accountability of both bad actors and the technology companies in the online environment, and the different actors and companies who are profiting from this harm;
- the disproportionate impact that new technology can have on women, young women, and children;
- the need for women to be involved in every stage of the design, development, monitoring, and evaluation of both technology, including new technology such as artificial intelligence, and regulations; and

[7] UNFPA. 2023. Op cit, p. 5.
[8] Ibid p 16 for examples.

- the connection between online harm and offline harm, including women's peace, security, and the right to live free from abuse, harassment, violence and victimisation.

28. We see a number of clear case studies, both here and abroad, which demonstrate both the perniciousness and immediate harm that is caused online. We are aware that this takes many forms, and that there is an intersectional element frequently at play as well, where women of colour, gender diverse, LGBTQ+ and disabled women may experience more harm and be subject to more abuse online. A few case studies that we want to highlight:

    - **Online abuse, hate and harassment**: This can be both public on posts and behind the scenes in private messages. For example, this study from the international NGO the Center for Countering Digital Hate (CCDH) found that 1 in 15 messages to the women in the study breached Instagram's community standards and "Instagram failed to act on 9 in 10 abusive messages and violent threats over DM reported using its tools and failed to act on any image-based sexual abuse within 48 hours"[9].
    - **Young women targeted with eating disorder, self-harm, mental health and suicide content within minutes of joining TikTok**: See, for example, this study[10] from CCDH.
    - **Women in political and high-profile roles are facing daily abusive messages and hate** – see for example these comments[11] from former Australian Prime Minister, Julia Gillard from 2016 after the murder of Jo Cox a British MP who was killed by a person who had radicalised on a diet of online disinformation and hate. We know that this targeted abuse is also frequently seen against other women in high profile roles, for example, Siouxsie Wiles[12], Jacinda Ardern[13], women MPs generally in Aotearoa[14] and

[9] Center for Countering Digital Hate. 2022. Hidden Hate: How Instagram fails to act on 9 in 10 reports of misogyny in DMs. https://counterhate.com/wp-content/uploads/2022/05/Final-Hidden-Hate.pdf

[10] Center for Countering Digital Hate. 2022. Deadly by Design: TikTok pushes harmful content promoting eating disorders and self-harm into young users' feeds. https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf

[11] Hunt E. 2016. Julia Gillard says online abuse deters women from political careers. https://www.theguardian.com/world/2016/oct/12/julia-gillard-says-online-abuse-deters-women-from-political-careers

[12] Covid-19 scientist Siouxsie Wiles reveals appalling social media abuse. 2020. https://www.nzherald.co.nz/nz/covid-19-scientist-siouxsie-wiles-reveals-appalling-social-media-abuse/7GHUGNG5KRU4WVQ44D67MTYHYM/

[13] Jacinda Ardern resigns: Social media 'cesspit' blamed for growing threats, abuse towards politicians. 2023. https://www.rnz.co.nz/news/national/482820/jacinda-ardern-resigns-social-media-cesspit-blamed-for-growing-threats-abuse-towards-politicians

[14] Women MPs subjected to 'real and widespread' sexism, harassment and violence - survey. 2019. https://www.1news.co.nz/2019/01/31/women-mps-subjected-to-real-and-widespread-sexism-harassment-and-violence-survey/

overseas[15] including high profile women during elections[16] and in politics[17], and that this can be a more intense experience for black, migrant and women of different ethnicities, wāhine Māori[18], Pasifika[19], the trans[20] community and disabled women[21].

- **Dangerous gendered disinformation, e.g. women's health**: For example, Google search made $10 million over the past two years by allowing misleading advertisements about fake abortion clinics that aim to stop women from having the procedure[22]. This is part of a bigger subset of disinformation that impacts on women's health[23], and gendered disinformation more generally, which is intersecting with violent extremism and national security[24].

General Comments

29. We agree with the Department that addressing online harm requires a comprehensive response from every sector in society. Everything important is impacted by what happens online – from our response to big issues like climate change and COVID-19, to social inclusion and elections, to addressing misogyny, racism, and different forms of online abuse and violence.

30. Until relatively recently, Governments internationally have declined to regulate social media companies, search engines and other parts of the online infrastructure. But the evidence against the idea that these companies have no role or are somehow neutral in the proliferation of online harm has mounted up. Problems are increasingly coming to light through whistleblower testimony, independent studies like those canvassed in this submission, and public hearings - such as congressional committee hearings in the States.

---

[15] Perraudin F, Murphy S. 2019. Alarm over number of female MPs stepping down after abuse. https://www.theguardian.com/politics/2019/oct/31/alarm-over-number-female-mps-stepping-down-after-abuse

[16] Simmons C, Fourel Z. 2022. Hate in Plain Sight: Abuse Targeting Women Ahead of the 2022 Midterm Elections on TikTok and Instagram. https://www.isdglobal.org/isd-publications/hate-in-plain-sight-abuse-targeting-women-ahead-of-the-2022-midterm-elections-on-tiktok-instagram/

[17] NDI. 2019, Tweets that Chill: Analyzing Online Violence Against Women in Politics. https://www.ndi.org/tweets-that-chill.

[18] Amnesty International UK. Black and Asian women MPs abused more online. https://www.amnesty.org.uk/online-violence-women-mps

[19] NZ Herald. 2018. Samoan author speaks out about online threats of rape and violence https://www.nzherald.co.nz/nz/samoan-author-speaks-out-about-online-threats-of-rape-and-violence/WI2IQ3QYIHEWA4TST37ESFJYRA/

[20] The Disinformation Project. 2023. Working paper: Transgressive transitions. https://thedisinfoproject.org/2023/05/05/working-paper-transgressive-transitions/

[21] eSafety Commissioner. 2022. How adults with intellectual disability experience online abuse. https://www.esafety.gov.au/research/how-adults-intellectual-disability-experience-online-abuse

[22] Korn J. 2023. Google earned $10 million by allowing misleading anti-abortion ads from 'fake clinics,' report says. https://edition.cnn.com/2023/06/15/tech/google-anti-abortion-ads-ccdh/index.html

[23] Sherman J. 2022. What is gendered health misinformation and why is it an equity problem worth fighting? https://meedan.com/post/what-is-gendered-health-misinformation-and-why-is-it-an-equity-problem-worth

[24] Di Meco L, Wilfore K. 2021. Gendered disinformation is a national security problem. https://www.brookings.edu/articles/gendered-disinformation-is-a-national-security-problem/

Countries are increasingly seeing the need to legislate as a way of forcing issues like transparency and holding Big Tech companies accountable for the harm that they are causing, contributing to and amplifying. Only effective legislation, transparency requirements and accountability can change the incentive and disincentive structure and the calculations that companies do when singling out engagement as the single metric for success. We know that things that happen online don't just stay online as we have seen so recently with tragedies in NZ and overseas, such as the Colorado shooting after a wave of increased anti-LGBTQ+ hate speech and the subculture incel violent extremism.

31. The world has not been agile enough to understand and respond to the harms arising from the online environment. Any regulatory framework and regulator needs to be nimble and resourced to the impact of changes in technology, society and both domestic and world events. There is a certain amount of "learning by doing" that needs to happen, with a model that will and should continue to evolve through alignment with human rights and democratic principles, a commitment to continuous improvement and evidence-based research.

32. We are concerned that an overly narrow focus on content regulation is likely to lead to lost opportunities and system gaps. At a minimum, the new regulator needs to have the ability to conduct and commission research looking at broader causes and impacts of harm from the online environment, with the statutory right to present this information to Parliament and a process for the Government to respond to any recommendations (for example, in the same way that the Law Commission reports are tabled and there is a government response). This will be particularly important as new technology develops (for example, Artificial Intelligence) and our understanding of technology expands (for example, through transparency reports and research).

Artificial Intelligence needs to be within scope of regulation

33. At the moment, this technology is being developed without a clear understanding of how decisions are being made by the AI (e.g. what patterns it is making when making calculations) and what the risks of it are (including how AI and machine-learning technology created for one purpose may be adopted and misused in a different context). Kissinger, Schmidt and Huttenlocher (2021)[25] outline some of the current risks and limitations of artificial intelligence in their book "The Age of AI: And Our Human Future". We draw attention to some of the key passages from this book (included in **Appendix B)**, which help to explain the nature, risks and challenges of AI and why there needs to be a multi-sectoral approach to monitoring risks and developing regulation. The authors specifically discuss how this technology has been adopted by search engines and social media platforms, and the risks of using this technology and making their services public without properly understanding how the algorithm makes decisions, including

[25] Kissinger A, Schmidt E, Huttenlocher D. 2021. The Age of AI: And Our Human Future. New York : Little, Brown and Company.

convergence and promotion of particular harmful content.. They note that the incentives within the current system are to rush to market rather than testing to a high standard. This is consistent with congressional testimony from ex-staffers at Big Tech companies in the United States in late 2022, which noted that all of the incentives in the system and employment prioritised engagement metrics and speed rather than safety, for example, engineering project management, bonuses and career progression.

34. Other countries and regions, such as the European Union and the United States, are already legislating (such as the EU AI Act[26]) or creating frameworks for the regulation and development of Artificial Intelligence (such as the White House's Blueprint for an AI Bill of Rights[27]). While the New Zealand Government launched a joint initiative[28] with the World Economic Forum about Redesigning the Regulation of Artificial Intelligence in 2019, at the time of writing this we have not been able to find the final report of this project. We note that there has been some work involved in understanding the impact of AI in autonomous weapon systems[29], through the Christchurch Call algorithm workstream[30] and that MBIE has been doing some work with partners in the AI Forum[31]. It would be good to understand what future plans the Government has in this space. We understand, from hearing speakers at Diplosphere's conference, that there is an inter-agency working group looking at AI. As per above, this needs to be a cross-sector conversation.

Core elements of legislation

35. We strongly recommend that legislation be progressed in the areas of safety by design, transparency and accountability of technology companies (and their senior executives) to an independent media regulator and the courts. Only effective legislation, transparency requirements and accountability can change the incentive and disincentive structure and the business calculations that companies do when singling out engagement and data

[26] EU AI Act: first regulation on artificial intelligence. 2023. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence?&at_campaign=20226-Digital&at_medium=Google_Ads&at_platform=Search&at_creation=RSA&at_goal=TR_G&at_advertiser=Webcomm&at_audience=ai%20eu&at_topic=Artificial_intelligence_Act&at_location=FR&gclid=Cj0KCQjwz8emBhDrARIsANNJjS5oBtU0dXC4QxBlJWuG-J92A7VI02aut_dyE7m0DCqmI0XxRVB7yb0aAu-YEALw_wcB

[27] The White House. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work For The American People. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[28] Te Tari Taiwhenua | Department of Internal Affairs. 2019. Artificial Intelligence (AI) workshop to start national conversation. https://www.dia.govt.nz/press.nsf/d77da9b523f12931cc256ac5000d19b6/51ee9bf29e9fd572cc2584a80001c762!OpenDocument

[29] Twyford P. 2021. Shaping the future: Autonomous Weapons Systems. https://www.beehive.govt.nz/speech/shaping-future-autonomous-weapons-systems

[30] Ardern J. 2022. Christchurch Call Initiative on Algorithmic Outcomes. Https://Www.Beehive.Govt.Nz/Release/Christchurch-Call-Initiative-Algorithmic-Outcomes

[31] Ai Forum New Zealand. 2018. Artificial Intelligence: Shaping a Future New Zealand. https://www.mbie.govt.nz/dmsdocument/5754-artificial-intelligence-shaping-a-future-new-zealand-pdf

extraction as a single metric for success. The CCDH has developed a STAR Framework[32] that sets these components in more detail:

Safety by design: ensuring that products and services are safe and have safety as a key component from design to implementation and amendment stage. Safety by design means being proactive at the front-end to anticipate risks, adequate testing, and constant review and evaluation of outcomes to feed in any necessary changes to systems and products. Safety by design is assisted by requirements for risk assessments, a proactive duty of care to ensure products are safe, transparency requirements, accountability and responsibility mechanisms. The Australian eSafety Commissioner[33] has written extensively on this issue and prepared guidance and resources for businesses.

Transparency of algorithms, rules enforcement and economics (advertising): these requirements would ensure that technology companies have:

- **Transparency of algorithms**: so that technology companies understand and are transparent about what is happening on their platforms and through their services, which enables everyone to better understand where there may be emerging and growing problems that have negative impacts and outcomes for the public. At a minimum, algorithmic transparency should include:
  - Search algorithms and data– such as autocompleting a keyword and metadata used;
  - Recommendation algorithms and data– which curate content that a user may be interested in;
  - Ad-tech algorithms and data– that target users based on demographics and behaviour to optimise advertising; and
  - Moderation algorithms and data– that target content, users and groups that breach the law or the platform's / search engine's terms and conditions/community standards. This should include internal metrics, such as the violative view rate.

    Algorithmic transparency will be assisted if independent researchers have API access and there is a contestable fund for independent research - levied from the technology companies.
- **Transparency of rules enforcement:** so people understand what the standards are, how decisions are made and that there will be consequences if they breach those rules. This includes having accessible and responsive complaints systems for users. Transparency in this area recognises the importance of everyone's freedom of speech, including the chilling effect and offline harm that can result from unchecked

[32] Center for Countering Digital Hate. 2022. Op cit

[33] Australian eSafety Commissioner. Safety by Design puts user safety and rights at the centre of the design and development of online products and services. https://www.esafety.gov.au/industry/safety-by-design

online harassment, hate and violence on children, young people, women and minorities.

- **Transparency of economics (advertising)**: Ad libraries and understanding what online ads are funding is an opaque and inconsistent area for social media platforms and search engines - yet they are critical for the development of an online harm ecosystem, in many cases resulting in the direct funding of websites and channels that promote hate, violence and dangerous disinformation. In addition, the presence of mainstream advertising on those sites can have an legitimising effect on the messages on that site - normalising extreme content.

  CCDH explains that transparency in this area means:

  > "specifically, understanding where, when, by whom, and using which data. One option for achieving this is to require advertisers to publicly declare, on their websites, the domains where their ads appear. This creates a driver for corporate accountability, i.e. that consumers' money is not being funnelled to content that fundamentally harms individuals, communities and society. This type of information is often provided to advertisers by brokers, some of which are updated in real time. This requirement would simply ensure that advertisers disclose the URLs of the pages on which their adverts appear, but not other information, such as performance data or targeting criteria."[34]

  They explain that this is needed because currently:

  > Each year, respectable companies and their customers unwittingly funnel millions of pounds directly to the Internet's most malicious and subversive actors and messages. Misinformation and hate sites are almost entirely funded by online advertising— often paid for by unsuspecting mainstream organisations who don't know what content their brand is appearing next to, and thereby funding. Their adverts are placed by third-party Brokers, such as Google's Adsense business, which then allocate adverts to particular sites to fulfil predetermined target demographic (age/gender/location) and psychographic (attitudinal and behavioural) profiles. The use of algorithms to select which ads appear where to fulfil a target profile has led to these services being called "programmatic advertising."[35]

Accountability to an independent media regulator and the courts: In this submission we have expressed strong support for a media regulator that is an independent Crown entity - arm's length from government and independent from industry. We have also provided substantial comments on the role and functions of the regulator.

[34] STAR Framework - pp16-17.

[35] CCDH. 2022. STAR Framework - Pg. 17.

We are concerned that an overly narrow focus on content regulation in this review is likely to lead to lost opportunities and system gaps. At a minimum, the new regulator needs to have the ability to conduct and commission research looking at broader causes and impacts of harm from the online environment, with the statutory right to present this information to Parliament and a process for the Government to respond to any recommendations (for example, in the same way that the Law Commission reports are tabled and there is a government response). This will be particularly important as new technology develops (for example, Artificial Intelligence and machine learning) and our understanding of technology expands (for example, through transparency reports and research). UNFPA has provided a useful guidance on the meaning of technology-facilitated violence with examples[36], which is another useful point for considering how women are impacted in different ways by the online environment and abuse is facilitated, assisted or achieved through technology. Any government that is serious about addressing violence against women will need to regulate the online environment.

Responsibility of technology companies and senior executives: This means that there are consequences for failing to fulfil statutory duties and requirements in the code of conduct. We support the concept of codes proposed in the consultation document, but recommend that these be drafted by the new independent media regulator in consultation with the community. There should be significant penalties for breaching the codes and core statutory duties. We agree that there should be a New Zealand-based representative for the company and recommend that there should be a "good character" / "fit and proper person" type test, such as applies to lawyers and in other professions. This is justified given the enormous power that these individuals and companies wield on public discourse, elections, social cohesion, public understanding of important issues, health, safety and democratic institutions.

The statute should set out principles and core requirements relating to online safety regulation and must include a duty of care, transparency requirements and the ability for the independent media regulator to inspect, audit, and review decisions of technology companies.

[36] UNFPA. 2O23. Measuring Technology-Facilitated Gender-based Violence, pg.5 https://www.unfpa.org/publications/measuring-technology-facilitated-gender-based-violence-discussion-paper. pg. 16.

Responses to Specific Questions from the Department

Definitions in the proposals

1. What do you think about the way we have defined unsafe and harmful content? (page 18)

36. In defining harmful content, it is also important to consider:

- **What the cumulative impact of individual pieces of content on individuals and communities may be**. For example, Facebook whistleblower Frances Haugen released Meta's internal research at Instagram[37], which found that 13.5% of teen girls said that Instagram makes thoughts of suicide worse and 17% of teen girls said that Instagram makes thoughts of eating disorders worse. Similarly, there is a cumulative effect of having a daily dose of online hate in messages or being shared a constant stream of misogynist content through an algorithmic feed. This can have a conditioning effect on both views and behaviour, a pattern that is intensified with little to no friction based on the way that these companies have designed their algorithms and platforms.
- **How content is used for the purposes of radicalisation and can lead to more extreme content and communities**: for example, it is common for people or groups that are connected with terrorism and violent extremism to use humour and memes on main platforms that channel a user through to more marginalised or encrypted platforms[38]. The recommender algorithm on social media also plays a part in connecting extreme and vulnerable communities together, including, for example Facebook automatically creating groups and suggesting people to connect to, on Instagram and Tik Tok recommending content and accounts to follow, and all of these platforms as well as Google and Youtube sharing extremist content in search results.
- **How content that references or connects with an event or characteristic may be triggering.** For example, after a violent attack, the creation of content connected to that event with video footage or words used during the event can be triggering.
- **The format that the content is received**: for example, most content that is shared online is not currently subject to any kind of age rating or warnings, which does apply on many other media platforms in New Zealand. This can be particularly problematic when that harmful content is recommended to you through, for example, Google search functions or the algorithms on social media. Suddenly a person may be presented with content that they had not sought and that may be distressing. One minute they may be looking at kitten videos and the next minute the algorithm may be recommending self-harm content.

[37] Keith M. 2021. Facebook's internal research found its Instagram platform contributes to eating disorders and suicidal thoughts in teenage girls, whistleblower says. https://www.businessinsider.com/facebook-knows-data-instagram-eating-disorders-suicidal-thoughts-whistleblower-2021-10

[38] See for example: NCTC, DHS, FBI. 2022. Use of Memes by Violent Extremists. https://www.dni.gov/files/NCTC/documents/jcat/firstresponderstoolbox/128S_-_First_Responders_Toolbox_-_Use_of_Memes_by_Violent_Extremists.pdf

- **The need to consider how online platforms bring together different communities in a harmful way.** For example, connecting victims and offenders together based on engagement with content, how children and adult strangers may engage and share content, how vulnerable groups of people may be connected together and sharing harmful content.
- **Coordinated efforts to create narratives that can have harmful impacts, for example, this study**[39] from ISD Germany shows how myths and disinformation about COVID-19 and the vaccine were shared and became popular across platforms. The researchers note that "On Telegram, the readership of various channels increased by up to 471%. Relevant Facebook pages also saw an average growth of 21% to a total of over 4.5 million followers between April 2020 and April 2021; a development that could also be observed on other platforms."
- **Harm can arise from both paid and unpaid content**: Frequently, the format that content is presented on social media and in search results on search engines like Google, is presented without a clear delineation between paid and unpaid content, i.e. paid content is an advertisement. This can have a number of issues including polluting the information ecosystem and people's understanding of core issues like COVID-19 or climate change, to abuse, violence and negative stereotypes and encouraging harmful behaviours.
- **Consumer protection from stereotypes and bias**: How access to/amplification of content can reinforce negative stereotypes or the algorithmic bias can mean that individuals are restricted in being able to access products or promoted harmful products and services based on unlawful characteristics, for example, promoting a club as being for "White People Only" or offering different mortgage interest rates at a bank based on someone's religion. This is an area that has received attention in the US, and is included as one of the principal areas in the White House's Blueprint for an AI Bill of Rights[40].
- **Content can be automatically generated and may not be directly created by an individual or group**: This is already the case, for example, with "bots" that generate harmful hate speech and disinformation. One well-known example involved the US Presidential election in 2016, where Twitter disclosed[41] (two years after the fact in the face of wide public upset) that 50,000 Russia-linked accounts used its service to post

[39] Winter H, Gerster L, Helmer J, Baaken T. 2021. Disinformation Overdose: A study of the Crisis of Trust among Vaccine Sceptics and Anti-Vaxxers. https://www.isdglobal.org/isd-publications/disinformation-overdose-a-study-of-the-crisis-of-trust-among-vaccine-sceptics-and-anti-vaxxers/

[40] The White House. Algorithmic Discrimination Protections: You should not face discrimination by algorithms and systems should be used and designed in an equitable way. https://www.whitehouse.gov/ostp/ai-bill-of-rights/algorithmic-discrimination-protections-2/

[41] Swaine J. 2018. Twitter admits far more Russian bots posted on election than it had disclosed. https://www.theguardian.com/technology/2018/jan/19/twitter-admits-far-more-russian-bots-posted-on-election-than-it-had-disclosed

automated material about the 2016 US election and that this had reached at least 677,775 Americans. This problem was not unique to Twitter as a platform, for example, a report from the University of Oxford's Computational Propaganda Project and the social network analysis firm Graphika found that there was in fact "a vast campaign spearheaded by the Internet Research Agency (IRA)[42] - a Russian company that has been described by the United States Intelligence Community as a "troll farm" with ties to the Russian government. The report says Russia had a particular focus on targeting conservatives with posts on immigration, race and gun rights. There were also efforts to undermine the voting power of left-leaning African-American citizens, by spreading misinformation about the electoral process." We are likely to see more and different applications of automated content with the continued development of Artificial Intelligence.

2. Does the way we have defined unsafe and harmful content accurately reflect your concerns and/or experiences relating to harmful content? (page 18)

37. The starting point should be that <u>everyone</u> is impacted by harmful content online and every thing that is important is impacted by the online environment. The Government response and corresponding action from technology companies needs to have this front and foremost. Whether it relates to elections, climate change and COVID-19 disinformation, online misogyny and other forms of online hate, radicalisation, terrorism and violence, or even the way that we can access products and services.

38. While children have particular vulnerabilities and sensitivities in the online environment that needs to be recognised in legislation, they are certainly not the only group affected and should not be the only group to benefit from new laws regulating the online environment.

39. As outlined in this submission, we are also particularly concerned about how the online environment can be weaponised against different groups in society, including women and girls. The Brookings Institute have studied this at a global level and found:

> "While sexist attitudes are integral to understanding violent extremism and political violence[43], social norms per se don't explain how attacks against women in politics have been weaponized for political gain and cynically coordinated by illiberal actors that take advantage of algorithmic designs and business models that incentivize fake and outrageous content. A new wave of authoritarian leaders and illiberal actors around the world use gendered disinformation and online abuse to push back against the progress made on women's and minority rights. This movement seeks to push women politicians and activists aside, reignite gender stereotypes and misogyny, and strategically take advantage of technology as a tool in

[42] Lee D. 2018. The tactics of a Russian troll farm. https://www.bbc.com/news/technology-43093390

[43] FBA, PRIO, UN Women. 2020. The Sexism and Violence Nexus. https://fba.se/en/about-fba/publications/the-sexism-and-violence-nexus/

these campaigns. Vladimir Putin in Russia[44], Rodrigo Duterte in the Philippines[45], Viktor Orban in Hungary[46], and Recep Tayyip Erdogan in Turkey[47] are among the political leaders who have used gendered disinformation campaigns to attack women in politics, aggressively challenge feminism, and attack liberal values.

40. These efforts are part of a larger strategy to weaken the human rights system. According to the UN Human Rights Council[48], the erosion of women's human rights "is a litmus test for the human rights standards of the whole of society," and this tech-enabled backlash against women's rights has broader ramifications for global peace and security[49].

41. State-aligned gendered disinformation campaigns are used as a deliberate tactic to smother opposition voices, erode democratic processes, and silence demands for government accountability[50].

42. UNFPA has been doing a lot of work with partners globally on understanding and responding to TFGBV. They explain this issue and have defined it as follows:

    "...this kind of digital violence is committed and amplified through the use of information and communications, technologies or digital spaces against a person based on gender. It is facilitated through the design and use of existing as well as new and emerging technologies (both hardware and software). It is always evolving.

    Technology-facilitated gender-based violence takes many forms, including sextortion (blackmail by threatening to publish sexual information, photos or videos); image-based abuse (sharing intimate photos without consent); doxxing (publishing private personal information); cyberbullying; online gender and sexual harassment; cyberstalking; online grooming for sexual assault; hacking; hate speech; online impersonation; and using technology to locate survivors of abuse in order to inflict further violence, among many others. (Click here for a glossary of digital-violence terms.) It carries significant health, safety, political and economic consequences for women and girls, for their families and communities, and for society as a whole. As

[44] Ferris-Rotman A. 2018. Putin's War on Women: Why #MeToo skipped Russia. https://foreignpolicy.com/2018/04/09/putins-war-on-women/

[45] Liotta E. 2019. Ranking The Worst Sexist Comments President Duterte Has Made About Women. https://www.vice.com/en/article/xwn4d3/duterte-sexist-comments-women-philippines

[46] Walker S. 2018. We won't keep quiet again': the women taking on Viktor Orbán. https://www.theguardian.com/world/2018/dec/21/hungary-female-politicians-viktor-orban

[47] O'Grady S. 2014. Erdogan Tells Feminist Summit That Women Aren't Equal to Men. https://foreignpolicy.com/2014/11/24/erdogan-tells-feminist-summit-that-women-arent-equal-to-men/

[48] United Nations. Human Rights Council. 38th session. 2018. Report of the Working Group on the issue of discrimination against women in law and in practice. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/132/85/PDF/G1813285.pdf?OpenElement

[49] Dharmapuri S, Shoemaker J. 2021. Peace & Security and the Digital Ecosystem: Five Emerging Trends in the Technology and Gender Policy Landscape. https://oursecurefuture.org/publication/women-peace-security-and-digital-ecosystem-five-emerging-trends-technology-and-gender

[50] Di Meco L, Wilfore K. 2021. Gendered disinformation is a national security problem. https://www.brookings.edu/articles/gendered-disinformation-is-a-national-security-problem/

women and girls self-censor to prevent technology-facilitated gender-based violence, their voices are silenced and democracies suffer."[51]

43. Locally, the Disinformation Project has published initial research on the rise of online misogyny in New Zealand[52].

44. Like other forms of online harm, the abuse does not stay online but has a very real risk of normalising misogyny and escalating discrimination and violence against women.

About our proposed new framework to regulate platforms

3. Have we got the right breakdown of roles and responsibilities between legislation, the regulator and industry? (page 32)

45. We strongly support the creation of an independent media regulator (independent Crown entity), which is arm's length from the Government. To ensure that it is able to gain and maintain public trust on important regulatory decisions, this body needs to be completely separate from the technology companies, any real or perceived political interference, and any perception that enforcement decisions are compromised because its governance is overseen by a government department that has operational enforcement functions that depend on decisions from the regulator.

46. The current system of multiple regulatory bodies has led to a situation where it is not clear to the public who they should and can make a complaint to in what situations, whether that person can actually do anything useful for members of the public that are affected by harmful content, and does not deal with the issue of media convergence. An accessible, single point of entry for consumers is an important part of a modern media regulatory system. Research, education, investigation, monitoring and auditing should also be core parts of the new regulator's role, and they should be able to partner with organisations to fulfil these functions.

47. We strongly recommend that the Government adopts a "safety first" approach to regulation of online technology companies rather than an industry-led approach to, for example, the development of the Codes of Practice. Given the poor track record of self-regulation, the continued opaqueness of the operations of these companies, and the overriding business incentives for driving engagement and profit over consumer safety - these companies should not be holding the pen on the Codes of Practice. NetSafe's experience with developing a Code generally did not receive buy-in from the community because it looked like it favoured industry at the expense of public safety and other public interests. Similarly, the Australian experience of using an industry-led method has resulted in the e-Safety Commissioner rejecting two of the final submitted Codes and

[51] Department of Internal Affairs | Te Tari Taiwhenua. 2023. Safer Online Services and Media Platforms. https://www.dia.govt.nz/diawebsite.nsf/Files/online-content-regulation/$file/Safer-Online-Services-and-Media-Platforms-Discussion-Document-June-2023.pdf

[52] Taylor K, Hannah K, Hattotuwa S. 2022. Dangerous speech, Misogyny, and Democracy. https://thedisinfoproject.org/2022/11/29/dangerous-speech-misogyny-and-democracy/

reserving judgement on an additional third Code[53]. This process has taken a substantial amount of time only to result in Codes that are not fit-for-purpose and have been outdated. We should learn from this experience.

48. We want to see this new regulatory framework succeed and would recommend that the new media regulator is responsible for developing the Codes in consultation with the community. Community includes technology companies but also extends, for example, to women's groups like NCWNZ, iwi / Māori, unions, schools, academics, experts, businesses and groups representing or advocating for consumers, public health, the environment, different faiths and ethnicities, LGBTQ+, human rights and disabled people. This approach is more efficient and credible, and draws on the twin pillars of technical expertise and democratic participation and engagement thus better ensuring that the public interest goals of an online safety regulatory regime are met.

49. Having the media regulator writing and having the final say on the Codes sets a very clear message about what the expectations and requirements are. The New Zealand-based media regulator is best placed to be in a convening role for consulting with all parts of the community compared to global corporations and will have and develop subject matter and technical expertise through these engagements and exercising their other regulatory functions, such as monitoring and making censorship decisions. This responsibility will also help to ensure the media regulator remains up-to-date with technological developments, government and society priorities.

50. We largely agree with the other roles and responsibilities outlined in the consultation document, although we would like to see the new regulator have gender analysis capability, including at the governance level. Further thought should be given to whether the new regulator (alongside NZ Police and Customs NZ) should be responsible for enforcement and what the benefits and costs are.

51. The public also needs to have tools to be able to identify emerging forms of online hate and disinformation, so that the regulator and the system more generally can benefit from these insights. We recommend that the public has access across platforms as outlined in the CCDH STAR framework[54] and that the codes include one for technology companies regarding data access for independent researchers[55], such as that proposed by UK civil society working with legislators on the Online Safety Bill.

[53] Australia. eSafety Commissioner. 2023. eSafety Commissioner makes final decision on world-first industry codes. https://www.esafety.gov.au/newsroom/media-releases/esafety-commissioner-makes-final-decision-on-world-first-industry-codes

[54] Center for Digital Hate (CCDH). 2022. STAR Framework: CCDH's Global Standard for Regulating Social Media. https://counterhate.com/research/star-framework/

[55] Center for Countering Digital Hate (CCDH) et al. 2023. Letter to UK Government: Data Access in Online Safety Bill. https://counterhate.com/wp-content/uploads/2023/06/Coalition-letter-OSB-data-access-amend-13_06_23-3.pdf

4. Do you agree that the government should set high-level safety objectives and minimum expectations that industry must meet through codes of practice? (page 32)

52. Yes. The Government should also receive advice from the independent media regulator on these issues. Legislation should include minimum standards that all platforms are required to meet.

5. Do you agree with how we have defined 'platforms'? Do you think our definition is too narrow, or too broad? If so, why? (page 32)

53. The retailer exception may cause issues and there is an inadvertent risk of, for example, a hate website setting up an online shop as a way of escaping regulation. For example, the well-known online hate and disinformation site Breitbart[56] has a store on its website and Etsy and Amazon have both been known to sell extremist products. This study[57] from the Institute for Strategic Dialogue (ISD) identified five such e-commerce sites (Etsy, Redbubble, Zazzle, Teespring and Teepublic) that were selling products:

> "promoting everything from harmful misinformation about the COVID-19 pandemic, to antisemitism and anti-LGBTQ+ hate, to neo-Nazi narratives and symbols. While there is evidence that these platforms are in many cases removing the most egregious and obvious forms of bigotry, it is still extremely simple to find and purchase hateful products across the full range of these platforms."[58]

54. There is a commercial element to a lot of the online hate and disinformation information ecosystem, which often is seen through advertising revenues (for example, Gateway Pundit) but does extend to these more tangible revenue streams.

6. We are trying to focus on platforms with the greatest reach and potential to cause harm. Have we got the criteria for 'Regulated Platforms' right? (page 32)

55. We encourage officials to test how and whether their proposal supports a "safety by design"[59] and public health approach[60] to dealing with online harm issues.

56. Reach (in terms of numbers) is one important metric for assessing harm but is not the only one. It is important to think about how the information ecosystem and radicalisation operates online. It is very common for a person to go down a "rabbit hole" from a mainstream platform to more extreme and smaller sites by engaging with click bait. For example, we are aware that there is an information ecosystem that is dedicated to the incelosphere and that this combines a mix of search engines like Google, platforms like Youtube and dedicated websites which are both public and encrypted. The regulation

---

[56] https://store.breitbart.com/

[57] Squirrell T, Martiny C. 2022. Profiting from Hate: Extremist Merchandise on Redbubble, Etsy, Teespring, Teerepublic and Zazzle. https://www.isdglobal.org/isd-publications/profiting-from-hate-extreme-merchandise-on-redbubble-etsy-teespring-teerepublic-and-zazzle/

[58] Center for Countering Digital Hate (CCDH). 2021. Gateway Pundit generated up to $1.5 million from election misinformation. https://counterhate.com/research/gateway-pundit-generated-up-to-1-5-million-from-election-misinformation/

[59] Australia. eSafety Commissioner . Safety by Design. https://www.esafety.gov.au/industry/safety-by-design

[60] PERIL. How We Work. https://perilresearch.com/how-we-work/

needs to cover the full gambit and the regulator / Police need the power and the tools to disrupt this violent extremism network, and to prevent young and vulnerable people from being encouraged down these pathways. An important question to answer when mapping this out is where the intervention points are.

57. Incels is just one example. Laura Bates has written extensively on the online "manosphere" in her book *Men who Hate Women*[61], which is her findings from undercover research online and found a "spider-web" of forums and groups ranging from everything from "pick-up artists" to men who hurt women, and how the different movements impact women in both the online and offline world by using the same radicalisation playbook as white supremacists and other extremist groups.

7. Do you think we have covered all core requirements needed for codes of practice? (page 39)

58. See recommendation above about the need for a code of practise for technology companies providing independent researchers with access to data.

59. In addition to the proposals, it is also important that codes include requirements on expectations relating to:

    - A transparent and responsive complaints system;
    - Transparency on paid posts and advertising libraries; and
    - Risk assessments and other ways that platforms will be discharging their duty of care, including when products and services may be amended (e.g. changes to the algorithm).

60. This needs to sit alongside legislative duties that set minimum standards including a duty of care to users and obligations when working with the regulator (e.g. during an audit process or when information is requested by the regulator).

61. There is a major benefit in developing subject specific codes, e.g. misogyny, and consulting experts and civil society as part of this process.

9. Do you think some types of platforms should be looked at more closely, depending on the type of content they have? (page 39)

62. Yes. But not limited to content it should also be the way that the content is experienced and how certain content is amplified, moderated or generated. For example, while TikTok markets itself as a dance and entertainment app, we know that this same platform is amplifying self-harm and suicide content to 13-year-old users and that people who identify as having an eating disorder are more at risk of receiving this content via TikTok's "For You" algorithm. The platform is designed to move quickly through short video clips and the consumer experience is incredibly immersive.

63. Some websites are designed for a criminal purpose, to harm others, or to encourage others to harm others. For example, we are aware of websites that have been designed

[61] Bates L. 2020. Men Who Hate Women: From incels to pickup artists, the truth about extreme misogyny and how it affects us all. Simon & Schuster

to encourage people to commit suicide, commit rape or doxying individuals. Other sites are a coordination spot for criminal activity. For example, how Andrew Tate[62] worked with followers on his "Hustler's University" site to flood TikTok's algorithm and amplify his posts and misogynistic and dangerous content.

10. Do you think the proposed code development process would be flexible enough to respond to different types of content and harm in the future? Is there something we're not thinking about? (page 43)

64. It may be necessary to act quickly on types of content, and impacts of new technology or before, during or after a major event (such as before an election or after the March 15 terrorist attacks). It may be that there needs to be a "temporary measures" code requirement or amendment to the code on a temporary basis in an emergency and / or until formal consultation processes on a final code are complete. There should be safeguards and independent oversight around when this is exercised.

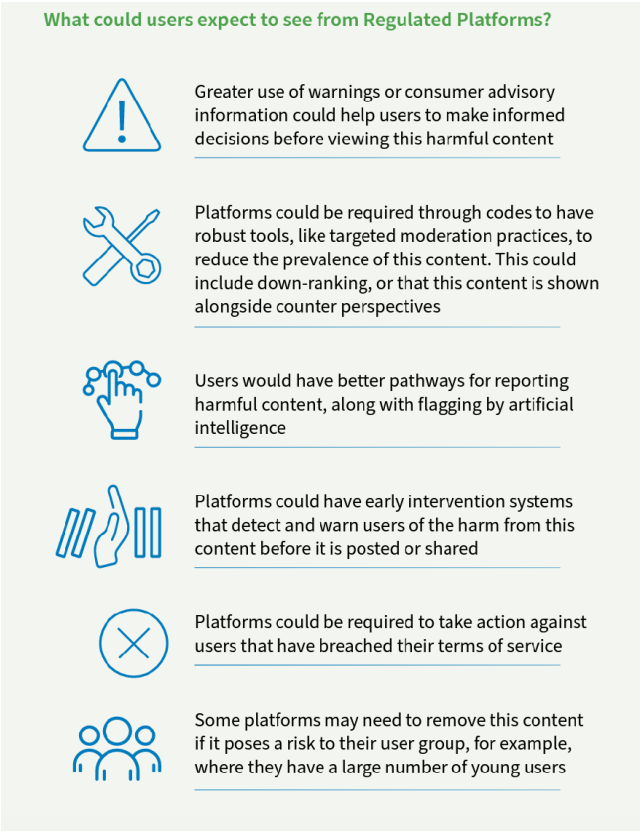
11. What do you think about the different approaches we could take, including the supportive and prescriptive alternatives? (page 43)

65. We support the prescriptive approach. Self-regulation and "regulation light-touch" have both palpably failed. We need to have a legislative framework that changes the incentives and disincentives in the system. Both the EU and the UK are legislating further than the supportive approach because they can see the very real impact on their citizens and democratic institutions. Gentle words of encouragement from a regulator and placing the focus of new obligations on consumers (which includes children) rather than companies that are profiting is not the fix we need and fails to recognise both the nature and extent of the harm. While media literacy, for example, should be part of the education function of a new media regulator, it is not sufficient or effective at preventing the amplification of misogynistic content or the spread of a livestream video of the March 15 terrorist attack.

12. Do you think that the proposed model of enforcing codes of practice would work? (page 48)

66. We note that you have used violent misogynist content as an example of how the new system may work with the following possible actions:

[62] Das S. 2022. Inside the violent, misogynistic world of TikTok's new star, Andrew Tate. The Guardian. https://www.theguardian.com/technology/2022/aug/06/andrew-tate-violent-misogynistic-world-of-tiktok-new-star



67. Given the link between online hate and offline violence, we are not convinced that the above (often voluntary platform) actions are going to be sufficient to mitigate the risk of harm to girls, young women, women, and gender-diverse communities.

68. It is important to emphasise why this content matters. This content is harmful because it silences women and women's right to freedom of expression, it creates hostility against women, it escalates into disadvantage, hate and violence. It normalises women being treated as less than human. Given the "othering" process that is created through online hate and misogyny, it is not sufficient for there to be removal *only if it poses a risk to their user group or for platforms to possibly have robust tools*. For example, there may be a Reddit forum or part of the Incelosphere that is dedicated to hate against women and we know that in some cases users that identify as women are prevented from joining and participating in these forums. The online hate does not stay online and the hate is not specific to those users but has a very real impact on people in the offline world. Toxic, violent communities that are created and allowed to fester unabated are at risk of inflicting harm on themselves and others. There are whole communities, for example, that idolise the violent extremist incel Eliot Rodger[63] and have shared his "manifesto". It is common for incels to add ER on posts as a way of referencing him and his beliefs.

[63] 2014 Isla Vista killings. https://en.wikipedia.org/wiki/2014_Isla_Vista_killings

69. This area of regulation needs to be strengthened. We note that the UK's Online Safety Bill[64], the Netz DZ Act in Germany and the EU's Digital Services Act all have significant penalties for companies failing to meet their obligations under the legislation.

13. Do you think the regulator would have sufficient powers to effectively oversee the framework? Why/why not? (page 48)

70. It is common for regulators to have litigation powers in overseas jurisdictions, for example, the Federal Trade Commission in the US has a mandate for enforcing consumer protection laws including any relating to social media. There remain judicial review protections if the Code itself goes beyond existing legislation. In New Zealand, we note that the Environmental Protection Agency[65], though a Crown Agent, does have this enforcement role as a regulator. The NZ Police and Customs NZ also have a key enforcement role in this process.

71. We would expect that in any new regulatory regime, the new regulator would be heavily involved if not leading a prosecution. We would want to understand a bit more what the Government considers the risks to be in giving the regulator this function.

14. Do you agree that the regulator's enforcement powers should be limited to civil liability actions? (page 48)

72. See above.

15. How do you think the system should respond to persistent non-compliance? (page 48)

73. There should be escalating penalties and actions that can be taken by the regulator in cases of persistent non-compliance, and in serious cases, and the regulator should be transparent about this in its reports.

16. What are your views on transferring the current approach of determining illegal material into the new framework? (page 54)

74. There is a need to ensure that harmful pieces of content that are banned (such as child sexual abuse material) under the current legislation continue to be banned under the new legislation.

17. Should the regulator have powers to undertake criminal prosecutions? (page 54)

75. See comments above.

18. Is the regulator the appropriate body to exercise takedown powers? (page 56)

76. Yes.

[64] Shead S. 2022. UK government to speed up criminal sanctions for tech bosses with new online safety laws . https://www.cnbc.com/2022/03/16/online-safety-bill-tech-execs-face-jail-time-under-new-uk-rules.html

[65] Environmental Protection Agency. RMA enforcement. https://www.epa.govt.nz/industry-areas/compliance-monitoring-enforcement/rma-enforcement/

19. Should takedown powers be extended to content that is illegal under other New Zealand laws? If so, how wide should this power be? (page 56)

77. Yes. The regulator needs to primarily focus on public safety and online harm. We do not think it is the appropriate body to be enforcing copyright legislation, for example, as that would risk flooding its functions and detract from the focus. We would like to review areas that the Government considers relevant here.

20. If takedown powers are available for content that is illegal under other New Zealand laws, should an interim takedown be available in advance of a conviction, like an injunction? (page 56)

78. See above.

Potential roles and responsibilities under the proposed framework

21. What do you think about the proposed roles that different players would have in the new framework? (page 63)

79. See response to question three and other comments in this submission. Primarily, this relates to:

    - Strengthening the role of the regulator and the public / community
    - Rebalancing the role of industry, so that public safety is driving legislation rather than company interests.

80. We also have concerns about how effective (or not) Netsafe has been within the current system. There is an inherent conflict with them accepting money from technology companies and purportedly acting in the best interest of the public – this is why there was such a big push-back on the code that they developed with industry before they went out for consultation with the public and why that is a failed, low-trust model. The new regulator should absorb the education function that NetSafe has.

22. Have we identified all key actors with responsibilities within the framework? Are there any additional entities that should be included? (page 63) What would the proposed model achieve?

81. See earlier comments above the governance structure / organisation including gender analysis capacity.

24. Do you think that our proposals will sufficiently address harms experienced by Māori? (page 69)

82. A New Zealand regulator is more likely to understand the context and importance of Te Tiriti o Waitangi than global media companies. We also consider it to be a more appropriate way of fulfilling the Crown's obligations of partnership, participation, and active protection under Te Tiriti.

83. The recommendations that we have made in this submission to improve the proposals are intended to benefit all victims of online hate, harassment, abuse and violence, including Māori. It is important that there are core transparency, accountability and responsibility requirements in the new regulatory system and that companies are driven to consider and address how their products and services may negatively impact different

groups, including Māori. These recommendations will help to drive a safety by design approach.

25. What do you think about how rights and press freedoms are upheld under the proposed framework? (page 70)

84. Press freedom is important but like other forms of speech, there are limits that can and should be placed in a free and democratic society. For example, there still need to be rules in place so that content is age appropriate, fair, accurate and does not encourage crime or other forms of harm against individuals and groups. There is also a distinction between news content and content that happens to be produced by a media company. In an age of media convergence, news articles are appearing online in people's social media feeds or search results in the same way as other content. Many articles are promoted with "click bait" headlines and feed off the same economic model of data, engagement and advertising.

85. In addition, it is important to scrutinise:

- to what extent a media outlet is a genuine, independent news outlet that follows recognised journalistic ethics and is accountable to a reputable regulatory body; and
- To what extent a media outlet is knowingly spreading disinformation or online hate, including as part of a broader information ecosystem.

26. Do you think that our proposals sufficiently ensure a flexible approach? Can you think of other ways to balance certainty, consistency and flexibility in the framework? (page 70)

86. There is a need to be responsive to emerging harm, social and technological changes - and as our understanding of the impacts from the new transparency requirements and research grows over time. This is an area of law that needs to remain under review and the government should ensure that it prioritises resources and legislative time for any necessary amendments. The Codes will provide a good deal of flexibility in the system.

Additional Comments and Next Steps

87. The Department should consider the imposition of an industry levy to help fund the new regulation and the commission of independent research on online harm.

88. We would welcome the opportunity to meet with you to discuss the comments and recommendations in this paper, and as the Government further develops its policy and legislative proposals.

Appendix A: Relevant Conclusions from CSW67

Full list available here[66]. Note the paragraph numbers are those in the report.

12. The Commission reaffirms that the Beijing Declaration and Platform for Action recognized that it is essential that all women not only benefit from technology, but also participate in the process from the design to the application, monitoring and evaluation stages. It recalls that, in the political declaration on the occasion of the twenty-fifth anniversary of the Fourth World Conference on Women, Governments pledged to harness the potential of technology and innovation to improve women's and girls' lives and to close the development divide and the digital divide, including the gender digital divide, as well as address the risks and challenges emerging from the use of technologies.

15. The Commission recognizes the need to ensure that human rights are promoted, respected and fulfilled in the conception, design, development, deployment, evaluation and regulation of technologies and to ensure that they are subject to adequate safeguards in order to promote an open, secure, stable, accessible and affordable information and communications technology environment for all women and girls.

16. The Commission acknowledges that multiple and intersecting forms of discrimination and marginalization are obstacles to the achievement of gender equality and the empowerment of all women and girls in the context of innovation and technological change, and education in the digital age. It respects and values the diversity of situations and conditions of women and girls and recognizes that some women face particular barriers to their empowerment. It stresses that, while all women and girls have the same human rights, women and girls in different contexts have particular needs and priorities, requiring appropriate responses.

17. The Commission recognizes that while technology can be used to promote women's and girls' full realization of civil, political, economic, social and cultural rights, it can also be used to perpetuate gender stereotypes and negative social norms and create vicious cycles, in which inequalities are amplified and perpetuated through digital tools, and also recognizes the need to address the impact of structural barriers to the realization of those rights.

18. The Commission expresses concern about the unequal pace of digital transformation and access to technology within and among countries and the structural and systemic barriers, inter alia, gender stereotypes and negative social norms and the disproportionate share of unpaid care and domestic work, undermining the ability of women and girls to securely access information and communications technologies and the Internet and to become equipped with the knowledge, awareness and skills for their social empowerment and women's economic empowerment and connected at a level that allows for a safe online experience at an affordable cost, especially in developing and African countries.

[66] Commission on the Status of Women 67th session 2023. Innovation and technological change, and education in the digital age for achieving gender equality and the empowerment of all women and girls. https://documents-dds-ny.un.org/doc/UNDOC/LTD/N23/081/71/PDF/N2308171.pdf?OpenElement

19. The Commission recognizes that adolescent girls are part of the most digitally connected generation in history and can disproportionately face discrimination, violence that occurs through or is amplified by the use of technology, and other barriers in the context of innovation and technological change, and education in the digital age, which prevents them from accessing the full benefits of digital technologies and meaningful participation in society, and can create and exacerbate inequalities.

26. The Commission reaffirms that the full, equal and meaningful participation of women in decision-making processes and in leadership positions at all levels is essential to the achievement of gender equality and the empowerment of all women and girls, as well as the realization of their human rights and fundamental freedoms. It also reaffirms the importance of the participation and leadership of women in decision-making related to information and communications technologies, including policies and programmes to promote women's and girls' ability to use digital technologies and to address any potential negative impacts of such technologies.

38. The Commission notes with concern that new technological developments can perpetuate existing patterns of inequality and discrimination in the absence of effective safeguards and oversight, including in the algorithms used in artificial intelligence-based solutions. It notes that gender bias in technology affects individuals but also contributes to setbacks in gender equality and women's empowerment, and that therefore a gender-responsive approach should be taken in the design, development, deployment and use of digital technologies with full respect for human rights.

39. The Commission recognizes that, despite the opportunities, there is a need to addresses challenges associated with the misuse of new and emerging digital technologies which can be designed and/or used to incite violence, hatred, discrimination and hostility, inter alia, racism, xenophobia, negative stereotyping and stigmatization of women and girls. The Commission expresses concern that women, and particularly girls, often do not and/or cannot provide their free, explicit and informed consent to the collection, processing, use and storage of their personal data or to the reuse, sale or multiple resale of their personal data, as the collection, processing, use, storage and sharing of personal data, including sensitive data, have increased significantly in the digital age.

40. The Commission recognizes that the way many digital platforms are designed, maintained and governed has given rise to disinformation, misinformation and hate speech, which can undermine the fulfilment of women's and girls' rights, including the right to freedom of opinion and expression and to participate in all spheres of public life, and, in this regard, also recognizes that teaching children and young people digital literacy and skills, competencies for positive engagement with digital technologies and respect for gender equality helps to address issues related to online safety, privacy and various forms of violence, including gender-based violence that occurs through or is amplified by the use of technology, and empowers youth, including young men and boys, to become agents of change for gender equality.

41. The Commission emphasizes that serious harm and discrimination against women and girls triggered by the use of new and emerging digital technologies call for regulations that take into account the voices and experiences of women and girls to improve accountability requirements to address any human rights violations and abuses and enhance transparency on how to use and protect data and address the potential human rights violations and abuses caused by the use of such products and services, taking into account the Guiding Principles on Business and Human Rights.

42. The Commission recognizes that social media has transformed how information is shared globally, providing women and girls with new channels to share content and opinions, as well as to come together to raise awareness and mobilize, and therefore stresses the need to facilitate and expand, particularly for women and girls, the accessibility and affordability of safe, secure and inclusive online platforms and digital technology, including by investing in and creating effective regulatory frameworks, including for content moderation and reporting mechanisms, that are fully compliant with relevant obligations under international human rights law.

43. The Commission recognizes that the promotion of and respect for women's and girls' right to privacy, according to which no one shall be subjected to arbitrary or unlawful interference with his or her privacy, family, home or correspondence, and the right to the protection of the law against such interference, are important to the prevention of all forms of violence, including sexual and gender-based violence, abuse and sexual harassment, cyberbullying and cyberstalking, as well as any form of discrimination which can occur in digital and online spaces. It is deeply concerned at the negative impact that surveillance and/or interception of communications, including extraterritorial surveillance and/or interception of communications, as well as the collection of personal data, in particular when carried out on a mass scale, may have on the exercise and enjoyment of the human rights of women and girls.

44. The Commission notes that many emerging digital technologies remain widely unregulated and recognizes the need for effective measures, for all enterprises which own, manage and govern digital technologies and services, to tackle the challenges associated with the use of such technologies, including those that have adverse impacts on gender equality and the empowerment of all women and girls, and to introduce due diligence to identify, prevent and mitigate the risks and negative impacts of technology on women and girls.

45. The Commission recognizes that the use of artificial intelligence has the potential to transform the delivery of public services, societies, economic sectors and the world of work and to contribute to the achievement of gender equality and the empowerment of all women and girls, as well as their human rights and sustainable development. It also recognizes that the use of artificial intelligence can contribute to setbacks in these areas and have far-reaching implications and cause disproportionate negative impacts on women and girls, especially through new evolving technologies that create new forms of violence, such as deepfakes.

46. The Commission notes with concern the underrepresentation of women and girls, and the lack of or limited participation of women and, as appropriate, girls in the conceptualization, development, implementation and use of digital technologies, as well as the use and production of imbalanced and non-representative data, which can lead to inaccuracies and biases in algorithms, the training of smart applications and artificial intelligence-based solutions, and therefore to discrimination, including racial and gender-based discrimination. It also notes with concern that this impacts the accuracy of facial recognition technologies, including for women and girls, and exacerbates racial inequalities, and notes in this context the importance of effective remedies to address those inaccuracies.

47. The Commission expresses concern that the current innovation ecosystems do not sufficiently contribute to achieving gender equality and are characterized by an uneven distribution of power and financial resources, resulting in women being significantly underrepresented in decision-making, affecting their rights and opportunities in the digital age, and being unable to benefit from the millions of decent and quality jobs created by the digital transitions.

48. The Commission emphasizes that national strategies on technology and innovation should provide a cohesive basis for gender-responsive policies and programming that contribute to the empowerment of all women and girls and protect, promote and respect their human rights. It recognizes the need to take a whole-of society and multi-stakeholder approach so that each actor contributes to putting in place the conditions that will shape infrastructure, regulations, business, investments and educational systems and provide a more inclusive digital environment. It also recognizes multi-stakeholder efforts related to the achievement of gender equality and the empowerment of all women and girls and the realization of their human rights, taking note of all international, regional and national initiatives in this regard to advance the full, effective and accelerated implementation of the Beijing Declaration and Platform for Action.

49. The Commission strongly condemns all forms of violence against women and girls, which are rooted in historical and structural inequalities and unequal power relations between men and women. It reiterates that violence against women and girls in all its forms and manifestations, online and offline, in public and private spheres, including sexual and gender-based violence, such as sexual harassment, domestic violence, gender-related killings, including femicide, harmful practices such as child, early and forced marriage and female genital mutilation, as well as child and forced labour, trafficking in persons and sexual exploitation and abuse are pervasive, underrecognized and underreported, particularly at the community level. It expresses deep concern that women and girls may be particularly vulnerable to violence because of multidimensional poverty, disability and limited or lack of access to justice, effective legal remedies and psychosocial services, including protection, rehabilitation and reintegration, and to health-care services. It re-emphasizes that violence against women and girls is a major impediment to the achievement of gender equality and the empowerment of all women and girls and that it

violates and impairs or nullifies their full enjoyment of all human rights and fundamental freedoms.

52. The Commission also recognizes that girls are often at greater risk of being exposed to and experiencing various forms of discrimination and gender-based violence and harmful practices, including through the use of information and communications technology and social media. It further recognizes that the COVID-19 pandemic has resulted in girls spending more time online, which has been exploited by offenders and has therefore increased the need for measures and education to promote child safety.

53. The Commission expresses concern about the continuity and interrelation between offline and online violence, harassment and discrimination against women and girls and condemns the increase of such acts that are committed, assisted, aggravated or amplified by the use of technology. The Commission is deeply concerned by the magnitude of various forms of violence, including gender-based violence that occurs through or is amplified by technology, and the significant physical, sexual, psychological, social, political and economic harm it causes to women and girls, throughout their life course, infringing on their rights and freedoms, in particular for those in public life. It recognizes how such violence significantly increases the risk of depression and suicide, especially among adolescent girls.

54. The Commission further condemns gender-based violence and the emergence and rise of harmful behaviours and narratives which undermine and discredit women's and girls' online and offline expression, forcing women and girls to self-censor, close their accounts on digital platforms or reduce their interaction in online and offline spaces, limiting their full and meaningful participation in public life and the enjoyment of their human rights and fundamental freedoms.

55. The Commission recognizes the harm caused to girls and, especially when non-consensual, to women by the use, sharing or dissemination, or threat thereof, of intimate or personal sexually explicit content, whether real or simulated, such as photographs or videos, including through peer pressure to create, share or disseminate such content, as well as the short- and long-term repercussions for the victims and survivors as a result of such actions. It notes the fact that several countries have criminalized the online circulation of such content, ensuring that victims do not have to rely solely on other criminal law provisions.

56. The Commission expresses concern that women participating in public life, including politicians, voters, candidates, election administrators, judges, journalists, women in sport and members of women's organizations, face higher levels of violence, including in digital contexts, and especially on social media, which prevents them from exercising their equal right to participate in all spheres of public life, and notes with concern that there is a lack of preventive measures and remedies, which underlines the need for action by Member States in partnership with other stakeholders.

57. The Commission recognizes the need to foster a policy of zero tolerance in the digital environment for all forms of violence against women and girls, harassment, stalking,

bullying, threats of sexual and gender-based violence, death threats, arbitrary or unlawful surveillance and tracking, trafficking in persons, extortion, censorship and illegal access to digital accounts, mobile telephones and other electronic devices, in line with international human rights law. It also recognizes the multi-jurisdictional and transnational nature of such activities and the continual use and adaptation of digital technologies by perpetrators to avoid detection and investigation, and calls for active cooperation among different actors, including States and their law enforcement and judicial authorities, and the private sector, with regard to detecting crimes, reporting them to competent and relevant authorities for investigation, safeguarding electronic evidence of crimes and handing the evidence over to those authorities in a timely manner, and enhancing international cooperation involving electronic evidence in this regard. The Commission is concerned about the use of technologies, including the Internet, social media and online platforms, to perpetrate trafficking in women and girls, including for sexual and economic exploitation.

58. The Commission stresses the need to develop and, where it already exists, strengthen and implement legislation that prohibits violence against women and girls that occurs through or is amplified by the use of technology and to provide adequate protection for women and girls against all forms of violence in public and private spheres, and the need to improve the coherence of policy actions for the elimination and prevention of all forms of violence, including gender-based violence that occurs through or is amplified by the use of technologies, around principles focusing on victim- and/or survivor-centred approaches, with full respect for human rights, access to justice, transparency, accountability and proportionality. It expresses concern about the lack of comprehensive and accurate disaggregated data collection on the extent of the prevalence, forms and impact of such violence, resulting in fragmented and incomplete information.

60. The Commission reaffirms the human rights of girls and recognizes that the fulfilment of these rights is assisted through the development of digital literacy and skills among children, as well as their parents or legal guardians, teachers and educators, and through empowering girls to report and seek help in responding to online threats and bullying in adequate ways, and raising their awareness of online safety. It notes with concern the use of technologies to facilitate various forms of exploitation of girls, including for online child sexual exploitation and sexual abuse, and the production and distribution of child pornography, also known as child sexual abuse material.

61. The Commission recognizes that negative social norms, as well as gender stereotypes and systemic and structural barriers, are among the root causes of the gender digital divide, causing persistent gender gaps in science, technology, engineering and mathematics education and women's and girls' lifelong learning opportunities, which keep women from attaining and retaining decent and quality jobs. It also recognizes the importance of women's full, equal and meaningful participation in the technology workforce, including in fast-growing and well-paid careers such as in cloud computing, software and artificial intelligence development and data management, and as entrepreneurs, innovators, researchers and industry executives and leaders. It notes that policies and programmes to

achieve gender parity in science, technology, engineering and mathematics should place the responsibility for driving change on those who are responsible for creating supportive workplaces and educational settings in order to promote the representation of women and girls from different backgrounds.

65. The Commission reaffirms that the right to the highest attainable standard of physical and mental health is foundational to building the resilience of all women and girls. It underlines the need for strengthening access to gender-responsive, safe, available, affordable, accessible, quality and inclusive health-care services, including those related to mental health, maternal and neonatal health, menstrual health and hygiene management, and ensuring universal access to sexual and reproductive health-care services, including for family planning, information and education.

Adopting gender-responsive technology design, development and deployment

(bbb) Take proactive steps to include women and girls in the planning, coding and design of machine learning and artificial intelligence technologies, including through investments in education and the adoption and implementation of actions to eliminate biases and discrimination against all women and girls in algorithms;

(ddd) Mainstream a gender perspective in the financing, design, development, deployment, use, monitoring and evaluation of emerging technologies to prevent, identify and mitigate potential risks for all women and girls and in order to ensure their full and equal enjoyment of human rights; and take measures to design and carry out periodic impact assessments of the effects of the use of emerging technologies with respect to the achievement of gender equality and establish, as appropriate, due diligence mechanisms and develop regulatory approaches to improve these technologies, including on transparency and accountability;

Strengthening fairness, transparency and accountability in the digital age

(eee) Develop and implement legislation, in consultation with all relevant stakeholders, including international organizations, business enterprises and civil society, with preventive measures, effective sanctions and appropriate remedies, that protects women and girls against violations and abuses, including of the right to privacy;

(fff) Adopt regulations on evaluation and audit requirements for the development and use of artificial intelligence to provide a secure and transparent, high-quality data infrastructure and systems to prevent and address human rights violations and abuses, as well as gender bias;

(ggg) Take concrete measures to harness and design digital technologies for the common good and promote norms and mechanisms facilitating accessibility and the fair distribution of the benefits of digital technologies for sustainable development and gender equality, such as global data commons;

Preventing and eliminating all forms of violence, including gender-based violence that occurs through or is amplified by the use of technologies

(kkk) Eliminate, prevent and respond to all forms of violence against all women and girls in public and private spaces, online and offline, such as sexual and genderbased violence, including domestic violence, gender-related killings, including femicides, all harmful practices, including child, early and forced marriage and female genital mutilation, sexual exploitation and abuse and sexual harassment, as well as trafficking in persons and modern slavery and other forms of exploitation, through multisectoral and coordinated approaches to investigate, prosecute and punish the perpetrators of violence and end impunity, and take appropriate measures to create a safe, enabling and violence-free working environment for women, including by ratifying key international treaties that provide protection against gender-based violence and sexual harassment;

(lll) Ensure that the perspectives of women, and girls as appropriate, are taken into account in armed conflict and post-conflict situations and in humanitarian emergencies and that they effectively and meaningfully participate, on equal terms with men, in the design, implementation, follow-up and evaluation of policies and activities related to conflict prevention, peace mediation, peacebuilding and post conflict reconstruction, as well as take into account the perspectives of women and girls who are internally displaced and who are refugees; and ensure that the human rights of all women and girls are fully respected and protected in all response, recovery and reconstruction strategies and that appropriate measures are taken to eliminate all forms of violence and discrimination against women and girls in this regard;

(mmm) Support the important role of civil society actors in promoting and protecting the human rights and fundamental freedoms of all women; take steps to protect such actors, including women human rights defenders; integrate a gender perspective into the creation of a safe and enabling environment for the defence of human rights and to prevent discrimination, violations and abuses against them, such as threats, harassment, violence and reprisals; and combat impunity by taking steps to ensure that violations or abuses are promptly and impartially investigated and that those responsible are held accountable;

(nnn) Condemn and take all appropriate measures, including legal action, to combat the use of digital tools, including social media and online platforms, for the purpose of harassment, hate speech and racism against women and girls, trafficking in persons and all forms of sexual exploitation and abuse of women and girls, as well as for child, early and forced marriage and forced labour, and any non-consensual sharing of personal, sexually explicit content of women and the production and distribution of child pornography, also known as child sexual exploitation and abuse material;

(ooo) Strengthen the understanding and track patterns of forms of gender-based violence that occur through or are amplified by the use of technology in order to guide evidence-based policymaking and programming and comprehensively measure its impact;

(ppp) Develop, amend and expand legislation and policies and strengthen their implementation in consultation with relevant stakeholders, including victims and survivors

of violence and women's organizations, by including victim- and survivor informed responses and fast-track processes to prevent, eliminate and respond to all forms of violence against women and girls that occur through or are amplified by the use of technology, and institute measures to address such violence;

(qqq) Adopt comprehensive measures and programmes that seek to address forms of gender-based violence and human rights violations against women and girls which can occur through the use of technology, including but not limited to the use, and threats associated with the use, of the unauthorized distribution or manipulation of information or images, and any other forms of violence that may arise due to the continual development of technology;

(rrr) Provide support to victims and survivors of gender-based violence that occurs through or is amplified by the use of technology through the provision of service responses that avoid retraumatization, including comprehensive social, health, care and legal services and helplines; ensure women's and girls' equal access to justice, including by providing accessible, confidential, supportive and effective reporting mechanisms for incidences of such violence; increase women's legal literacy and awareness of available legal remedies and dispute resolution mechanisms; and provide civil and administrative alternatives for victims and survivors who have difficulty in gaining access to legal avenues owing to financial barriers or systemic discrimination, while recognizing the major contribution of civil society women's organizations that provide supporting services to survivors;

(sss) Develop effective gender- and age-responsive strategies, while bearing in mind the best interests of the child, for preventing and combating sexual exploitation and abuse of girls in digital contexts, including by ensuring that institutions providing services to girls are equipped with appropriate safeguards to prevent and intervene early, and for building protective factors in families, households and communities to impede offenders' efforts, both online and offline, taking into account the roles and responsibilities of their parents, legal guardians or other individuals legally responsible for them;

(ttt) Explore the potential of new technologies to support efforts to prevent and respond to sexual violence in armed conflict and to facilitate the participation of victims and survivors in criminal justice processes, as appropriate;

(uuu) Strengthen the capacity and improve policy coherence and coordination of government actors, including parliamentarians, policymakers, law enforcement officials, the judiciary, health and social workers and educators, and of civil society organizations, to develop knowledge, skills and digital expertise to prevent and eliminate violence against women and girls that occurs through or is amplified by the use of technology, including through institutional training, and provide victim- and survivor-centred support;

(vvv) Ensure that public and private sector entities prioritize the prevention and elimination of gender-based violence that occurs through or is amplified by the use of technology by implementing, through meaningful engagement with victims and survivors, safeguards and preventive measures that address multiple risk and protective factors related to violence, including improved content moderation and curation and the interoperability,

transparency, accessibility and effectiveness of reporting systems, including by establishing robust and reliable content removal processes that are fully compliant with relevant obligations under international human rights law.

Appendix B: Relevant Extracts from Kissinger, Schmidt and Huttenlocher's book The Age of AI: and Our Human Future (2021).

Issues
Nature of AI decision-making
• "AIs do not "explain" how or what they learned in human terms. Nor can developers ask an AI to characterise what it has learned… At best we can only observe the results an AI produces once it has completed its training. Accordingly, humans must work backwards. Once an AI produces a result, people – be they researchers or auditors – must verify that the AI is producing the results required."[67] • "Google's image-recognition software has infamously mislabelled images of people as animals and animals as guns. These errors are plain to any human but eluded the AI. Not only are AIs incapable of reflection, they also make mistakes – including mistakes that any human would consider rudimentary."[68] • "Alternatively, AI bias may result directly from human bias… this can occur in the labelling of outputs for supervised learning in the labelling of outputs for supervised learning - whatever misidentification the labeller makes, deliberate or inadvertent, the AI will encode. Or a developer may incorrectly specify a reward function used in reinforcement training. Imagine an AI trained to play chess on a simulator that overvalues a set of moves favoured by its creator. Like its creator, that AI will learn to prefer those moves even if they fare poorly in practise."[69] • "... quantity and coverage matter - training AIs on large quantities of highly similar images will result in neural networks that are incorrectly certain of an outcome because they have not encountered it before."[70] • Example: Tay (Microsoft's chatbot) "encountered hate speech and quickly began to mimic it, forcing its creators to shut it down." This was a situation where the algorithm was not "fixed" but was continuing to learn in a live public environment[71]. • AI is currently constrained by code in three ways: 1. Code sets the parameters of AI's possible actions – "these parameters might be quite broad, permitting a substantial range of autonomy and therefore risk"[72]

[67] Kissinger A, Schmidt E, Huttenlocher D. 2021. Op cit. p. 77-78.
[68] Ibid, p. 79.
[69] Ibid, p. 80.
[70] Ibid, p. 79-80.
[71] Ibid, p. 81.
[72] Ibid, p. 84.

2. "AI is constrained by its objective function, which defines and assigns what it is to optimize."[73]
3. "AI can only process inputs that it is designed to recognize and analyze."[74]

- "Once AI has been trained, it typically acts faster than the speed of human cognition… We are experiencing and facilitating changes that require our attention – in thought, culture, politics, and commerce – well beyond the scope of a single human mind or particular product or service."[75]
- "... some network platforms have assumed functions so significant as to potentially influence the conduct of national governance. … A network platform operating according to its standard commercial objectives and the demands of its users may, in effect, be transcending into the realm of governance and national strategy."[76]
- "As the tools for spreading disinformation become more powerful and increasingly automated, the process of defining and suppressing disinformation increasingly appears as an essential social and political function."[77]
- "The power to train defensive AI against an objective (or subjective) standard of falsehood - and the ability, if any can be developed, to monitor that AI's operations – would in itself become a function of importance and influence rivalling the traditional roles held by government. Small differences in the design of an AI's objective function, training parameters, and definitions of falsehood could lead to society-altering differences in outcomes."[78]
- "New users may adapt underlying algorithms for very different aims. A commercial innovation by one society could be adapted for security or information-warfare purposes by another."[79]

AI makes mistakes but is served up to the public regardless

- "And while developers are continually weeding out flaws, deployment [of a product to the market] has often preceded troubleshooting."[80]
- "Developing professional certification, compliance monitoring, and oversight programs for AI – and the auditing expertise their execution will require – will be a crucial societal project. In industry, pre-use testing exists on a spectrum. App

[73] Ibid.
[74] Ibid.
[75] Ibid, p. 98
[76] Ibid, p. 111.
[77] Ibid, p. 115.
[78] Ibid, p. 116.
[79] Ibid.
[80] Ibid, p. 79-80.

developers often rush programs to market, correcting flaws in real time, while aerospace engineers do the opposite: test their jets religiously before a single customer ever sets foot on board."[81]

- "… the robustness of AI auditing and compliance regimes is poor. In the real world, an unexpected failure can be more harmful, or at least more challenging, than an expected one … the inability of AI to check otherwise clear errors on its own underscores the importance of developing testing that allows humans to identify the limits of an AI's capacities, to review its proposed courses of action, and to predict when an AI is likely to fail …"[82]

AI used on social media

- "We rely on AI to assist us in pursuing daily tasks without necessarily understanding precisely how or why it is working at any given moment. We are forming new types of relationships that will have substantial implications for individuals, institutions, and nations - between AI and people, between people using AI-facilitated services, and between the creators and operators of these services and governments. Without significant fanfare – or even visibility – we are integrating nonhuman intelligence into the basic fabric of human activity. ... As more users are drawn to [large platforms, e.g. Google and Facebook], gatherings tend to result in a large base of users – sometimes … even billions. The network platforms increasingly rely on AI, producing an intersection between humans and AI on a scale that suggests an event of civilizational significance… network platforms seek to build their user bases and commercial partnerships in regions containing markets that are commercially and strategically significant to Washington and Beijing."[83]
- "As AI becomes increasingly critical to network platforms' functioning, it is also becoming, gradually and unobtrusively, a sorter and shaper of reality - and, in effect, an actor on the national and global stage."[84]
- "Election campaigns on social media undertaken by Russia and other powers – are a kind of digitized propaganda, disinformation, and political meddling with a larger scope and impact than in previous eras… A central paradox of our digital age is that the greater a society's digital capacity, the more vulnerable it becomes."[85]
- "AI-facilitated disinformation and psychological warfare, including the use of artificially created personae, pictures, videos, and speech, is poised to produce

[81] Ibid, p. 82.
[82] Ibid, p. 81.
[83] Ibid, p. 95-96.
[84] Ibid, p. 102.
[85] Ibid, p. 153.

unsettling new vulnerabilities, particularly for free societies. Widely shared demonstrations have produced seemingly realistic pictures and videos of public figures saying things they have never said."[86]

- "AI powers are in a position to deploy machines and systems exercising rapid logic and emergent and evolving behaviour to attack, defend, surveil, spread disinformation, and identify and disable one another's AI."[87]

The need for multi-stakeholder perspectives

- "What the consumer welcomes as a convenience, the national security official may view as an unacceptable threat or the political leader may reject as out of keeping with national objectives. ... The nature and scale of network platforms is beginning perspectives and priorities of different worlds in complex alignments, sometimes creating tension and mutual perplexity. In order for individual, national, and international actors to reach informed conclusions about their relationship to AI – and to one another – we must seek a common frame of reference."[88]
- "AI-enabled network platforms have the capacity to shape human activity in ways that may not be clearly understood - or are even clearly definable or expressible - by the human user. This raises essential questions: With what objective function is such AI operating? And by whose design, and within what regulatory parameters?"[89]
- "New concepts of understanding and limitations - between regions, governments, and network platform operators - must be defined. The human mind has never functioned in the manner in which the internet era demands. With its complex effects on defence, diplomacy, commerce, health care, and transportation posing strategic, technological, and ethical dilemmas too complex for any one actor or discipline to address alone, the advent of AI-enabled network platforms is raising questions that this should not be viewed as exclusively national, partisan, or technological in nature."[90]

[86] Ibid, p. 159.
[87] Ibid.
[88] Ibid, p. 99.
[89] Ibid, p. 109.
[90] Ibid, p. 132.

The risks and options for small countries

- "Previously sources of information and communication were typically local and national in scope - and maintained no independent ability to learn. Today, network platforms created in one country could become the arteries and lifeblood of another country, as the platform learns which consumers need certain products and as it automates the logistics of provision. In effect, such network platforms could become critical economic infrastructure, giving the country of origin leverage over any country that relies on it."[91]
- "For countries that do not produce homegrown network platforms, the choice for their immediate future seems to be between:
  1. Limiting reliance on platforms that could provide leverage to an adversary government;
  2. Remaining vulnerable – for example, to the potential of another government's potential ability to access data about its citizens; or
  3. Counterbalancing potential threats against each other."[92]

[91] Ibid, p. 128-129.

[92] Ibid, p. 128.