Submission to Special Rapporteur's Call for Input to Report on Freedom of Expression and Gender Dimensions to Disinformation
Professor Lorna Woods, OBE

A - CONCEPTUAL ISSUES

1. A preliminary issue is the unhelpfulness of elements of the definitions of mis/dis/mal-information commonly used when considering policy-based or technology-based interventions. While these definitions focus on inaccurate or misleading information, a key element of any definition in this context, the concern arises in relation to the factors used to distinguish between mis/dis/mal-information. Specifically, the definitions of misinformation and disinformation focus on the mindset of the person disseminating the information. This is appropriate where we consider *ex post* responses based on the individual activities/publications of an individual – especially where possible penalties (eg civil action for defamation) are an issue. This allows an assessment of the proportionality of the response.

2. An approach based on intent does not, however, take into account the impact of the misinformation. Such impact does not necessarily correlate to speaker's intent. It is possible to destroy a person's life accidentally or carelessly and such content also contributes to a 'toxic' environment more generally[1]. For this reason, there is a preference for definitions such as "the spread of deceptive or inaccurate information and images against women in politics, following story lines that often draw on misogyny and distrust of women in politics, frequently referring to their sexuality".[2] Here there are two key aspects both relating to the characteristics of the information:
   - that the information is inaccurate and/or misleading; and
   - the information ties into misogyny and mistrust of women – essentially gender-based.

   While disinformation is not limited to express or pure fact claims, it conveys a message about the ways things are; it should include opinion linked to facts or evidence. It should not extend to views that can only ever be subjective, emotional or solely convey threats.

3. While much of this material is sexualised, it is submitted that this should not be an element of the base definition as it might exclude some material that is demeaning and undermining of women but not yet sexualised (eg the "get back to the kitchen" line of thinking, women are only fulfilled in motherhood, or undermining a woman's qualifications or experience). Note gendered disinformation could be about women (or other groups)[3] as a group or about a specific individual (and the two interconnect). Gendered disinformation includes direct attacks but should also include messaging that exploits or reinforces gender inequalities – and might even extend as far as stereotypical hyper-partisan stories. Within an overarching classification of online gender-based disinformation, it may be important to distinguish

1 Judson et al, 'Engendering Hate: The Contours of State-Aligned Gendered Disinformation Online' (Demos, 2020), https://demos.co.uk/wp-content/uploads/2020/10/Engendering-Hate-Report-FINAL.pdf.

2 Lucina Di Meco, "Why Disinformation Targeting Women Undermines Democratic Institutions", *Power 3.0*, 1 May 2020, https://www.power3point0.org/2020/05/01/why-disinformation-targeting-women-undermines-democratic-institutions/, accessed 5 July 2023; contrast eg Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, & Kaufmann, Z.. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online (Wilson Center, 2021) and Judson et al n2.

3 Women are more likely to be targeted than men; trans people also attract a high level of online abuse. See e.g. UNHRC, 'Report of the Special Rapporteur on Violence Against Women, Its Causes and Consequences on Online Violence Against Women and Girls from a Human Rights Perspective' (18 June 2018) UN Doc A/HRC/38/47, para 28.

between these different contexts, styles and actors; more work needs to be done in this area.[4] It is also important to note the intersectional aspects that play into gender-based disinformation.[5]

4. The online environment brings specific aspects to this problem. Similar to online gender-based violence (or online and technology facilitated gender-based violence (OGBV)), online gender-based disinformation has specific characteristics, including reach, scale, speed and impact. The information technologies available now can be exploited or weaponised to target campaigns on the basis of gender and may link in to other extermist communities.[6] While understanding perpetrators' motivations and understanding their techniques and networks is helpful to understanding the nature and extent of the harm at individual and societal levels as well as for developing mitigation strategies, this potential for weaponisation should not be seen as part of the definition of online gendered disinformation itself. There is a difference between the type of information and the way it is used and modes of dissemination – as well as a distinction between the content and the actors involved (though some types of actors may have a preference for some sub-types of gendered disinformation and distribution techniques).

5. Gender-based misinformation is often seen through the lens of women politicians, journalists and those in other parts of public life and the silencing effect it may have on them, as well as its limiting effects on their effectiveness as they have to deal with the distraction of online gender-based misinformation. This is likely to affect other women and discourage them from engaging in public life, adversely impacting on democracy and its institutions. While this is certainly a concern, it should be noted that this problem does not affect just the 'obvious' public life jobs but affects public-facing roles more broadly – those working in higher education for example, may find the public engagement now necessary for academic progression more challenging than male colleagues. In schools, news reporting suggests female teachers are being challenged and undermined by male pupils seemingly influenced by the gendered disinformation pedalled by Andrew Tate. Even beyond this, women may have their credibility undermined within their communities, potentially leaving some isolated in the face of offline abuse.

6. Given the breadth of the definition of both online gender-based violence, which can include – according to UNHCR[7] - sexual, physical, mental and economic harm inflicted in public or in private, and online gendered disinformation, there is considerable overlap between the two terms. They may be driven by the same societal factors and both link to discrimination. It can be difficult to distinguish between harassment and abuse and the spread of disinformation. Nonetheless, they are not coterminous. Disinformation has a connection to fact claims. A key element of gender-based violence is the existence of violence. While this

4 See e.g. classifications proposed by the National Democratic Institute - eiter, K., Pepera, S., and Middlehurst, M. Tweets That Chill: Analyzing Online Violence Against Women in Politics. National Democratic Institute, 2019, https://www.ndi.org/sites/default/files/NDI%20Tweets%20That%20Chill%20Report.pdf and by ISD in relation to TVEC content: Jacob Davey et al, A Taxonomy for the Classification of Post-Organisational Violent Extremist and Terrorist Content (ISD, 2021), https://www.isdglobal.org/isd-publications/a-taxonomy-for-the-classification-of-post-organisational-violent-terrorist-content/

5 Thakur, D., & Hankerson, D. L. (2021). Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender (Center for Democracy & Technology, 2021) https://cdt.org/insights/facts-and-their-discontents-a-research-agenda-for-online-disinformation-race-and-gender/

6 See eg Marc-André Argentino et al, She Drops: How QAnon Conspiracy Theories Legitimize Coordinated and Targeted Gender Based Violence (ISD, 2022), https://www.isdglobal.org/isd-publications/she-drops-how-qanon-conspiracy-theories-legitimize-coordinated-and-targeted-gender-based-violence/

7 https://www.unhcr.org/what-we-do/protect-human-rights/protection/gender-based-violence

must be broadly understood, especially in the online context to encompass non-physical threats and harms, there should be violence (or threat of violence). In the more indirect examples of gendered disinformation (essentially demeaning stereotypes and sexism[8]) it is hard to see the violence threshold crossed. It is arguable, however, that if the same tropes and stereotypes were linked to a particular person then the psychological harm could be seen as the result of violence through the act of targeting. This points to another potential distinction between OGBV and gender based disinformation. In OGBV the content is directed to a target directly or indirectly; disinformation addresses the community, though again the distinction is far from clear-cut. There may well be overlaps especially given the impact of the standards and beliefs in an environment affect the ability of people speak out as well as contributing to the 'gender digital divide'[9] more generally. Gendered misinformation therefore can be seen as contributing to the "manifestation of historically unequal power relations between men and women, which have led to domination over and discrimination against women by men and to the prevention of the full advancement of women"[10], a description applied to gender-based violence more generally.

B – RESPONSES OF STATES AND OTHER ACTORS

7. In the UK, the Online Safety Bill[11] is currently reaching the final stages of the legislative process. Introduced to "make the UK the safest place in the world to go online", according to the UK Government, the Bill introduces a regulatory regime which imposes duties on service providers in relation to certain types of criminal content and content harmful to children. The Bill contains no obligations for providers to take any action in relation to content that affects adults that does not reach the criminal threshold even if that content is contrary to the civil law (eg privacy, defamation, data protection) or contravenes other regulatory standards (eg advertising standards (which includes rules on stereotypes)). Significantly, disinformation, unless it constitutes a relevant criminal offence[12] or is harmful to children[13], will lie outside the regime. While the regime has often been portrayed as being about takedown, the obligations are instead to have systems and processes in place to achieve certain ends – essentially a lessening of the chance that users encounter criminal content. This is what has been termed a systemic approach to regulation. The key duties are to undertake risk assessments and to put in place steps to mitigate and manage the risk of harm to users. Services are required to take freedom of expression into account (and to a lesser extent privacy/data protection) when doing this. While the child risk assessment does recognise that groups with some characteristics might be more at risk than others, the Bill has on the whole been silent as to the gendered nature of harm and its silencing effect. The

8 See e.g Council of Europe definition in Recommendation CM/Rec(2019)1 of the Committee of Ministers to member States on preventing and combating sexism (C/MRec(2019)1, 27 March 2019), https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168093b26a: Any act, gesture, visual representation, spoken or written words, practice or behaviour based upon the idea that a person or a group of persons is inferior because of their sex, which occurs in the public or private sphere, whether online or offline, with the purpose or effect of: i. violating the inherent dignity or rights of a person or a group of persons; or ii. resulting in physical, sexual, psychological or socio-economic harm or suffering to a person or a group of persons; or iii. creating an intimidating, hostile, degrading, humiliating or offensive environment; or iv. constituting a barrier to the autonomy and full realisation of human rights by a person or a group of persons; or v. maintaining and reinforcing gender stereotypes.

9 UNHRC, 'Promotion, Protection and Enjoyment of Human Rights on the Internet: Ways to Bridge the Gender Digital Divide from a Human Rights Perspective' (5 May 2017) UN Doc A/HRC/35/9, para 3.

10 Declaration on the Elimination of Violence Against Women, UNGA Res 48/104 (20 December 1993), preamble.

11 https://bills.parliament.uk/bills/3137

12 The National Security Bill will introduce an offence of foreign interference to tackle state based disinformation campaigns.

13 Health disinformation is one example.

most recent set of Government amendments, however, include provision for the regulator to introduce guidance about those duties in relation to women and girls and the disproportionate risk of harm they face. Although this is not express on the face of the text, this could address at least some aspects of online gendered disinformation (though it will be limited in relation to the scope of the existing duties).[14]

8. The principle of systemic regulation was derived from work done by Carnegie UK Trust, which proposed a single statutory duty of care on a service provider.[15] This borrowed from health and safety at work regimes where the obligation on the operator is to take reasonable steps to prevent against foreseeable harm and based on the insight that design features, business model and user tools have an indirect impact on content and that platforms have some responsibility for those choices. The mechanism for this is risk assessment followed by appropriate risk mitigation (with ongoing review). In the online environment, the physical environment is replaced by the software making up the design architecture of the service, together with the way the business is run – so what is its business model, how much does it invest in product safety and user complaints? Crucially, the regime does not focus on individual items of content but looks at the system across which that content flows. In emphasising architecture over take down, a greater range of interventions are possible, which may allow for more proportionate responses from a freedom of expression point of view.[16]

9. The statutory duty of care is a principle operating at a high level of abstraction and needs more detail through guidance, industry standards or codes of practice. Carnegie UK argued that while individual content domains may be differently affected by particular design features, and different user tools may be helpful depending on context, all content flows through the same distribution chain in a service. This means that the same questions arise about risk of features and operational choices should be asked, even if the impact of design choices and consequently mitigation measures might differ across content domains. On this basis it seems that a common framework could be developed by reference to an information flow model (and Carnegie UK proposed a four-stage model- below[17]). The framework would form the basis for a company approach to risk assessment and mitigation. This framework could be deployed across multiple content domains and jurisdictions. In adopting this cross-cutting approach, design-based risk mitigation measures can be seen to have cross-domain – and cross harm – effects. The approach may therefore be more efficient for service providers in tackling specific harms across a range of content domains and – potentially – across jurisdictions. Carnegie UK then worked with a number of civil society organisations specialising in ending violence against women and girls to contextualise the outline of the code into the context of violence against women and girls.[18] It did not seek to

14 Amendment 152, marshalled list of Amendments to the Online Safety Bill, House of Lords Report Stage, 4th July 2023. Amendment text here: https://bills.parliament.uk/bills/3137/stages/17765/amendments/96016

15 Will Perrin and Lorna Woods, *Online Harm Reduction – a Statutory Duty of Care and a Regulator* (Carnegie UK Trust, 2019), https://carnegieuktrust.org.uk/carnegie-uk-online-safety-bill-resource-page/

16 Lorna Woods, The Carnegie Statutory Duty and Fundamental Freedoms (Carnegie UK, 2019), https://carnegieuktrust.org.uk/carnegie-uk-online-safety-bill-resource-page/; this point was subsequently recognised by the UN Special Rapporteur for Freedom of Expression: Paper A/74/486 19 October 2019 para 51 https://www.undocs.org/A/74/486

17 Carnegie UK: Model Code: a reference model for regulatory or self-regulatory approaches to harm reduction on social media (2023): https://carnegieuktrust.org.uk/publications/model-code-a-reference-model-for-regulatory-or-self-regulatory-approaches-to-harm-reduction-on-social-media/

18 Carnegie UK Trust, EVAW, Glitch, Nspcc, Refuge, 5Rights Foundation, Prof Clare McGlynn and Prof Lorna Woods, Violence against Women and Girls (VAWG) Code of Practice, (2022), https://www.endviolenceagainstwomen.org.uk/wp-content/uploads/2022/05/VAWG-Code-of-Practice-16.05.22-

limit itself to online violence only but sought to recognise that online violence is often part of offline threats and violence.

10. The model code has many similarities with the UN Guiding Principles on Business and Human Rights (Ruggie) and the OECD Guidance for Multinational Enterprises and would be consistent with those approaches. Any model code reflecting these principles should include a recognition of the responsibility of the service provider as regards the results of its business choices, as well as a commitment to (product) safety by design (including safety testing). The provider must also commit to the risk assessment and mitigation process, determining appropriate metrics to assess the appropriateness and success of the mitigation plan, that take into account human rights. While the ICCPR is an obvious starting point, CEDAW and its general recommendations are relevant as, in relation to girls, is the UNCRC and specifically General Comment 25. In reviewing progress, service providers must engage with relevant experts and organisations representing groups adversely affected by the relevant content. This is particularly important given the male-domination in STEM subjects and consequently the construction of online environments.[19]

11. The four-stage information flow model, which reflects the role of the platforms in creating and influencing the flow of content from their users, comprises the following:
    a. access to the service and content creation;
    b. discovery and navigation;
    c. user response tools; and
    d. platform response.

12. Within each of the four stages a number of different considerations nest. More detail on each of these can be found in the Coalition VAWG Code.
    a. **Access to the service and content creation** includes tools available to users to create content (e.g. filters, nudification apps and mechanisms for labelling content), as well as restrictions (e.g. limits on frequency of posting) but also includes the user sign-up process and the terms of service for use of the platform. So questions around anonymity, multiple accounts, the acceptability of bot accounts and disposable accounts could all be considered here as well as the adequacy of the terms of service (assessed either against national law or international law standards, as appropriate). Networks of accounts and groups (such as those constituting the manosphere) could be assessed at this stage, looking at both actors and their behaviours. Security of accounts (in the light of doxxing and hijacking of accounts) is also important, as are default privacy settings. Terms of service need to recognise the nature of the problem and be appropriately granular; here it is key to recognise the gendered nature of a sub-set of disinformation and not treat it as just disinformation. The main focus in community standards or terms of service tends to lie on user-facing provisions; advertising content policies should not, however, be forgotten; nor the impact of advertising revenue sharing business models on user content creation. There is a tendency for outrageous and strongly emotive content to receive a lot of engagement; rewarding this sort of content is problematic. Gender-based disinformation could be created under this reward system. Some users may not intend harm but nonetheless

Final.pdf

19 David Sullivan, 'Business and Digital Rights: Taking Stock of the UN Guiding Principles for Business and Human Rights in the ICT Sector' (Association for Progressive Communications, June 2016), www.apc.org/en/pubs/business-and-digital-rights-taking-stock-un-guidin

may intend to post content that contributes to gendered disinformation; nudges can be targeted putting friction into the creation process.

b. **Discovery and navigation** covers all sorts of recommendation tools, and features for organising content and people to follow such as hashtags and feeds highlighting trending issues, as well as search functions/autocompletes. Some hashtags are abusive; there should be some sort of oversight or review process (even if automated) for the use of hashtags. Even when not directly abusive, they can be demeaning: Canadian politician Catherine McKenna was linked to the hashtag 'Climate barbie'.[20] As regards autocompletes, in 2013 UN Women did work on what came up as autocomplete starting "women shouldn't" or "women should", resulting in significant misogyny.[21] oday the word abortion triggers the auto-complete "is bad". The service provider should review their recommender systems, whether in relation to content or to other users to follow, especially their automated systems; it should should check automated systems for bias (e.g. arising from training data). As part of this a service provider could consider whether to provide appropriate information to its users about the accuracy (or otherwise) of information (eg flagging content that has been fact-checked). Advertising delivery systems also fit here, including advertiser sign-up processes (KYC), ad content policy and audience segmentation tools. It should consider how to institute oversight over the segments used for personalisation and have policies in place to identify unacceptable or unethical labels, such as might emerge through automation. A final consideration is the extent to which content from other sites should be permitted: the concern smaller platforms that have no interest in compliance and which use mainstream platforms to enlarge their own user base. While a service cannot have responsibility for the choices of other services, it might consider whether it is providing a bridge for them to new reach new audiences, potentially contributing to the mainstreaming of idea contained in gendered disinformation.

c. **User response** tools allow the user to curate and adapt the online environment, but this category also includes tools for engaging with content (like buttons, for example, or features to facilitate reposting and sharing) as well as the ease of making complaints. All tools should be easy to use by all groups of users likely to access the service. While tools allowing the user to curate their own environment can be empowering it is important to ensure that potential victims are not made wholly responsible for their own safety; women already have to do much 'safety work'.[22] The difference between the two positions may depend on the other features and tools available and the environment on the service generally. The reporting processes cover all content and behaviour (whether user-generated, service generated (e.g. autocompletes) or advertising-based).

d. **Platform response** includes moderation and complaints processes, including any user rights of appeal, crisis protocols and transparency reporting. An essential

20 EU Disinfo Lab, Gender-Based Disinformation: Advancing Our Understanding and Response, 20 October 2021, https://www.disinfo.eu/publications/gender-based-disinformation-advancing-our-understanding-and-response/

21 UN Women, UN Women ad series reveals widespread sexism, 21 Octobeer 2013, https://www.unwomen.org/en/news/stories/2013/10/women-should-ads; even now Google autocompletes to 'women shouldn't speak in church'.

22 Vera-Gray, F. and Kelly, L. 'Contested gendered space: public sexual harassment and women's safety work' (2020) *International Journal of Comparative and Applied Criminal Justice* https://www.tandfonline.com/doi/full/10.1080/01924036.2020.1732435

element in this system is that the service provider must have in place sufficient numbers of moderators, proportionate to the service provider size and growth and to the risk of harm, who are appropriately trained to review harmful and illegal content and who are themselves appropriately supported and safeguarded. Given the graphic and image-based nature of some of the gender-based disinformation, the provider could consider whether text-based reporting, especially that comprising drop-down menus are always appropriate. Standard reporting process could be particularly problematic in the context of image based sexual abuse. Perhaps special consideration should be given to the complaints of some groups of users – e.g. journalists or politicians (especially during an election period). This is not to suggest that substantively they should be treated differently but that a fast-track complaints mechanism might be necessary.[23] It should also provide the opportunity for non-users who are affected by content or behaviour on the service to report that content and/or behaviour. An appeals process needs to be available.

13. At each of the four stages, an intervention could be any one of: an ex ante design choice; the provision of tools or other mechanisms; or content specific responses. For example, in terms of discovery, a service could choose to optimise for authoritative sources; allow users more control to curate their own feed; or introduce suppression measures related to particular content or speaker or networks.

C- PROPOSALS FOR CHANGE

14. Consideration should be given to developing a code of practice, to be seen in the first instance as a voluntary or best practice guide. The model outlined above provides a starting point, enabling an initial consideration of the system design aspects that create an environment in which gendered disinformation can take hold and flourish. This should be developed in consultation with subject experts to understand how the problems of online gendered disinformation interact with system design, The experience of using this collaborative approach in the development of the Coaliton VAWG code of practice demonstrates that the engagement of subject experts will lead to a much deeper understanding of the problem areas and possible ways to develop further best practice. This iterative approach will also underline the robustness of the "Model Code" framework in identifying the core aspects of system design that can exacerbate online harms, while also allowing for a flexible or modular approach to that framework, according to the subject matter. The use of a code signals the value placed on best practice but also provides a benchmark against which providers can be measured.

Author Biography:

Lorna Woods is Professor of Law at the University of Essex and a member of the Human Rights Centre there. Starting her career in a technology, media and telecommunications practice in the

23 E.g. Jigsaw (Google Unit) developed with Twitter an open source tool to manage abuse, intended for journalists and public figures: Harassment Manager: https://jigsaw.google.com/harassment-manager/; it is discussed Jigsaw, Technology to help women journalist document and manage online abuse, 8 March 2022, https://medium.com/jigsaw/technology-to-help-women-journalists-document-and-manage-online-abuse-5edcac127872

City of London she has long-standing experience in the field of communications regulation and its intersection with human rights. Since 2018 she has worked with Carnegie UK Trust on a project concerned with the reducation of harm on social media, including proposals to fight against online gender-based violence. Her expertise in this field is well recognised. She has given expert evidence to numerous official and parliamentary inquiries in the UK and internationally, participated in an expert panel before the Colombian Constitutional Court, chaired an expert working group for the OSCE Special Representative on Freedom of the Media on AI and freedom of expression looking at hate speech, and received an OBE for her services to internet safety policy.