Taxonomy Expansion Proposal to GIFCT
R. Jabri-Markwell
16 March 2021

1. The goal of GIFCT's framework is to prevent violent extremists and terrorists from exploiting digital platforms. While this is an important aim, it does not capture the actors and online echo chambers that work to socialise individuals towards violence.
2. Also, the reliance on external designation lists ignores the contemporary reality of radicalisation and recruitment materials. Online echo chambers that socialise individuals towards violence include many examples of the materials that GIFCT is trying to prevent from being shared. These materials are rarely detected by platforms as they are buried within comment threads and lack organisational labels.
3. Proposals to expand designation and proscription lists have struggled with the political and legal difficulty of defining 'extremist ideology' or 'extremist rhetoric' where there are no explicit or imminent calls to violence. The scope for 'terror-scaping' ideas, organisations, or individuals, merely because they present as extreme, unpopular, or fringe, is a genuine concern, especially for marginalised communities that are already subject to over-policing and may have legitimate grievances with nation-states.
4. The involvement of political or foreign policy objectives in criteria for an expanded taxonomy is problematic and may overwhelm or stifle the development of a working solution.
5. Therefore it is imperative to develop a capability for assessing materials that have a close nexus to socialising individuals towards extremist violence in a way that is ideology-neutral, clearly defined and prioritises human rights.
6. Previous attempts of policymaking in this area tend to oscillate between very general approaches (e.g., UK's failed Bill to ban extremist speech in 2015[i]) and specific guidelines, often adopted by platforms, to list the types of hate speech or incitement that will not be accepted. The latter approach misses organisations or websites that serially attempt to socialise individuals towards extremist violence, especially when they skirt beneath the threshold of hate speech or criminal incitement[ii] (for example, through disinformation).
7. Platforms are motivated to assess one piece of material at a time, rather than patterns of behaviour over time of hateful online echo chambers. The material relied upon can dehumanise in aggregate over time in ways that are not apparent if assessing each piece individually.
8. A recent review[iii] by the UK Independent Commission for Countering Extremism recommended establishing a legal framework to counter hateful extremism, which it has defined as:

> activity or material directed at an out-group" (e.g., Muslims) who are perceived as a threat to an in-group (e.g., a Far-Right group) "motivated by or intending to advance a political, religious or racial supremacist ideology: a. To create a climate conducive to hate crime, terrorism, or other violence; or b. Attempt to erode or destroy the fundamental rights and freedoms of our democratic society as protected under Article 17 of Schedule 1 to the Human Rights Act 1998 ('HRA').

9. Their report emphasises that this is a working definition, not a legal one. It also recommended treating hateful extremism with as much priority as terrorism.
10. Defining extremist material or activity at law is more fraught, especially in articulating a clear actus reus (the act, as opposed to the intent). This ambiguity can create anxiety about big state or big tech interference in freedom of speech. Thus, instead of defining extremist material or activity, the GIFCT ought to consider targeting a technique relied upon by most violent extremist movements.
11. Dehumanisation offers an enduring, internationally accepted[iv] , and well-defined[v] concept, grounded in genocide prevention studies[vi]and increasingly in the literature on countering violent extremism.[vii] Most violent extremist movements tend to rely upon the dehumanisation

of an 'out-group' to their 'in-group' audience. We also know that at least Facebook and Twitter use dehumanisation as one of their policy frames.

12. Dehumanisation works to help a person overcome normal moral objections they may have to enact violence against another person or group. The target group is placed in a subhuman or inhuman category and constructed as an existential threat – thus, violence against them becomes proper, necessary, even righteous.

13. Dehumanisation is carried out through language and discourse, portraying the target group as:

Subhuman	**Mechanically inhuman**	**Supernatural**
The material presents the class of persons to have the appearance, qualities or behaviour of an animal, insect, form of disease or bacteria, or the material suggests that the whole class of persons is polluting, despoiling, or debilitating society.	The material presents the class of persons be inanimate or mechanical objects, or the material suggests the class of persons acts in concert to harm the in-group and are incapable of human thought or feeling.	The material presents the class of persons to be a supernatural threat.

14. Much dehumanisation occurs gradually and cumulatively through disinformation and conspiracy narratives. The presence of explicit dehumanising language (eg cancer, disease, rats) is unnecessary to dehumanise an out-group to an in-group audience.
15. While it may be tempting to set the threshold higher at incitement to violence, incitement to violence is a difficult and inappropriate threshold here given
    a. Platforms (and criminal contexts) demand it poses an imminent threat – creating an impractical evidentiary burden for whole communities targeted by the material. Measuring the 'tipping point' for danger appears only to be workable where extremist violence or genocide has already occurred, and incitement can be retrospectively measured. Imminent harm is more useful in criminal contexts involving threats against individuals.
    b. A person who uses violent language is often reacting to dehumanising materials directed to them as an in-group audience member. The most prevalent and harmful forms of weaponization of digital platforms are not by organisations or websites openly inciting, threatening, or glorifying violence, but inducing and inspiring it through dehumanising materials about outgroups to in-group audiences.

16. It is proposed that the taxonomy be expanded to materials from a website or organisation that overtime create an aggregate harm of dehumanising an 'outgroup' to an in-group audience.
17. As there is no external designation list that is based on this criterion, this proposal requires the conception of a process that is not reliant on such lists.[viii]
18. The absence of a connected designation or proscription list should not preclude GIFCT strategy to contend with the eco-systems that socialise individuals towards violent extremism.
19. An organisation or website that is found to meet the standard of aggregate harm should be marked with a cautionary label by GIFCT, to invite assessment by member platforms for link blocking, demoting, de-platforming or labelling. That is unless procedural fairness can be assured by GIFCT, in which case, it could explore making decisions about particular organisations and websites.
20. At the same time, whilst GIFCT respects that content moderation is the business of respective platforms, the GIFCT ought to promote understanding of aggregate harm and its predictors so that platforms are not assessing dehumanisation on a post by post/ tweet by tweet/video by

video basis. To complement this proposal, platforms ought to be incentivised to more competently and consistently escalate forums on their platforms for assessment.

21. My first draft of a framework to guide assessments of aggregate harm is attached in Annexure A. It contains universal predictors that could be used to determine if an organisation or website is engaged in a project of dehumanisation of an out-group.
22. This policy work has been developed through an intensive study of Facebook and Twitter platforms. There is research that is currently under review for publication.
23. I would be pleased to collaborate with others to enrich a similar or connected proposal or this one, on the recommendation of GIFCT.

Taxonomy Expansion Proposal to GIFCT
R. Jabri-Markwell
16 March 2021

ANNEXURE A

Framework for determining whether an actor has *over time* dehumanised a group of persons identified on the basis of a protected characteristic.[ix]

These predictors indicate aggregate conduct to dehumanise an identified group:

1. *Dehumanising conceptions on the actor's website in relation to an identified group.* This may be expressed explicitly on the website through language or narratives[1] that portray the identified group as subhuman[2], mechanically inhuman[3] or supernaturally inhuman[4].[5]

2. *The features of material that are serially published, specifically*
   i) The subjects or participants routinely identified in material. Analysts will be looking for signs of essentialising an identity as part of a dehumanising discourse about an 'outgroup'. For example, their identity (eg Muslim, Jew, Black) is routinely emphasised in material to collectively attribute guilt for a specific member's heinous crimes, or to suggest over time that all members of that group act in concert.
   ii) Hostile verbs or actions (eg stabs, sets fire) *attributed to those subjects to cumulatively associate them with sub- humanity, barbarism, or serious threat to the in-group.*
   iii) Use of explicitly dehumanising descriptive language or coded extremist movement language with dehumanising meaning (eg invader, a term used in RWE propaganda to refer to Muslims as a mechanically inhuman and barbaric force). Note that the majority of headlines in our study did not rely on dehumanising descriptors or coded language though.
   iv) Proportion of actor's material that act as 'factual proofs' to particular narratives about this identified group. Here, narratives could be defined as narratives that have been used previously to justify atrocities or violence against this identified group.
   v) Presence of 'baiting' content to in-group audience. Rhetorical techniques like irony to draw an even more hateful response towards the identified group.
3. *Evidence in the user comment threads of a pattern of hate speech against a group on the basis of a protected characteristic.* This would include blatantly dehumanising remarks, iteration of

---

1 For example, demographic invasion and replacement narratives about Islam and Muslims are grounded in dehumanising conceptions of Muslims, as evidenced by the responses they illicit. This included the portrayal of Muslims as:
- mechanically inhuman 'theological automatons' who are 'unified in thought and deed' to carry out demographic invasion (Lee 2015). Significantly, it follows that there is no way to tell if Muslims are truly peaceable or not, and therefore all Muslims are a threat.
- Subhuman in their inherent violence, barbarism, savagery, or in their plan to infiltrate, flood, reproduce and replace (like disease, vermin without explicitly using those terms).

2 The material presents the class of persons to have the appearance, qualities or behaviour of an animal, insect, form of disease or bacteria. Or the material suggests that the whole class of persons are polluting, despoiling or debilitating society (a description used by Maynard and Benesch, above n 1, 80.

3 The material presents the class of persons be inanimate or mechanical objects; or the material suggests the class of persons acts in concert to harm the in-group and are incapable of human thought or feeling.

4 The material presents the class of persons to be supernatural threat.

5 Where an ideology is not explicitly identified by the site, as the Institute for Strategic Dialogue has done in these circumstances, a sample of the site's produced material could be subjected to qualitative assessment. The other factors listed above would assist in that assessment.

extremist ideology concerning the target group as an existential threat to the in-group, or glorification of, or incitement towards, violence against the target group. Where that pattern is evident in relation to a high proportion of links shared from one host website, this can be taken as a primary sign that the website is engaged in a project of hatred or dehumanisation. However, the absence of comments does not signify that dehumanisation has not been successfully enacted in the user's mind.

Taxonomy Expansion Proposal to GIFCT
R. Jabri-Markwell
16 March 2021

Bio of Author

Rita Jabri Markwell is a lawyer and political/policy adviser, who has had a career in federal politics policy development and advocacy. She is also a trained English and History teacher. Most recently, she has been working with the Australian Muslim Advocacy Network (AMAN) to interrogate how propaganda supportive of the Christchurch terror attack and demographic invasion theories continues to survive on mainstream platforms. This has involved much direct engagement, problem-solving and testing of various ideas with industry (Facebook and Twitter), research and civil society sectors. In 2020, she spearheaded a study of five actors on Facebook and Twitter to understand what predictors ought to be used by platforms to measure dehumanisation over time of out-groups to in-group audiences. AMAN is a national body working to secure the physical and psychological welfare of Australian Muslims, which includes developing policy solutions to online disinformation and hatred.

Taxonomy Expansion Proposal to GIFCT
R. Jabri-Markwell
16 March 2021

ENDNOTES

[i] John Ware, 'Why Britain must not let extremists operate with impunity', The Article, 24 February 2021 <https://www.thearticle.com/why-britain-must-not-let-extremists-operate-with-impunity>

[ii] See for example, Will Baldet, "How 'Dangerous Speech' Is The Mood Music For Non-Violent Extremism: How do we define websites, groups and individuals who stay the right side of our hate crime laws but whistle the tune which advances the rhetoric of violent extremism?", *Huffpost*, 9 May 2018; James Grierson, 'UK extremists 'exploiting gaps in law to push their agenda', The Guardian, 10 June 2020. <https://www.theguardian.com/society/2020/jun/10/uk-extremists-exploiting-gaps-in-law-to-push-their-agenda>; Lizzie Deardon, 'New laws needed to tackle 'shocking and dangerous' scale of extremism, review finds', The Independent, February 2021 <https://www.independent.co.uk/news/uk/home-news/extremism-laws-review-impunity-mark-rowley-b1806349.html>

[iii] Released 24 February 2021 <https://www.gov.uk/government/publications/operating-with-impunity-legal-review>

[iv] 'Genocide begins with 'dehumanization;' no single country is immune from risk, warns UN official', UN News, 9 December 2014.

[v] Nick Haslam. (2006). Dehumanization: An Integrative Review. Personality and social psychology review: an official journal of the Society for Personality and Social Psychology, 257.

[vi] Jonathan Leader Maynard and Susan Benesch, 'Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention' (2016) 9(3) Genocide Studies and Prevention: An International Journal 70.

[vii] Department of Security Studies and Criminology. (2020, October 9). Mapping Networks and Narratives of Online Right-Wing Extremists in New South Wales (Version 1.0.1). Sydney: Macquarie University.
Marczak N. (2018) A Century Apart: The Genocidal Enslavement of Armenian and Yazidi Women. In: Connellan M., Fröhlich C. (eds) A Gendered Lens for Genocide Prevention. Rethinking Political Violence. Palgrave Macmillan, London.

[viii] Given requirements around privacy, freedom of expression, due process and natural justice, it is undesirable to move towards an expanded designation or proscription list without it being grounded in a transparent administrative framework that allows for appeals and review.

[ix] GIFCT will be able to compile a list of protected characteristics recognised commonly by the member platforms. At a minimum it should refer to the UN list: United Nations Strategy and Plan of Action on Hate Speech Detailed Guidance on Implementation for United Nations Field Presences, September 2020, <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf>.