How AI can either exacerbate or prevent genocides: Reflection based on the 10 Stages of Genocide

Budapest Centre
FOR MASS ATROCITIES PREVENTION

Published by the Budapest Center for Mass Atrocities Prevention, February 2021

Authors: Marine Milard *(Research Assistant)* and Sophie Smith *(Research Assistant)* with the guidance of Dr. György Tatár (*Chair of the Budapest Centre)*

Throughout the twentieth century, genocide became a frequent occurrence with millions massacred. This has continued into the twentieth-first century, where perpetrators engage in the intentional mass killing of particular groups in society. Such events are not sudden occurrences but take place under particular circumstances developed over time.

To help recognize the evolution of genocidal processes and prevent future tragedies, American scholar Gregory H. Stanton developed the theory of "ten stages of genocide," which describe the different stages leading up to a genocide. The stages are as follows: (1) classification; (2) symbolization; (3) discrimination; (4) dehumanization; (5) organization; (6) polarization; (7) preparation; (8) persecution; (9) extermination; and (10) denial.[1] This process is not necessarily linear, and stages may occur in parallel to each other.

The authors of the article attempt to demonstrate the role that artificial intelligence (AI) can play throughout that process in terms of how it can exacerbate the situation or prevent its escalation. In particular, AI in relation to (1) the media and (2) surveillance is discussed given that both appear to be the most common features within the ten stages. While there are of course other AI tools that may be employed throughout the genocidal process, they will not be the focus of the paper. The paper merely attempts to introduce its readers to and raise awareness of the ten stages of genocide, providing a detailed overview of said stages, in addition to how AI vis-à-vis the media and surveillance may play a role in the process.

The article is the third piece in the series of reflections, prepared by the group of interns of the Budapest Centre, which aims to illustrate the role of AI in fighting mass atrocities. The authors hope that the document will also contribute to the research planned by the Budapest Centre within the Initiative "Multipolar Task Force."

Artificial Intelligence and the Media

[1] Gregory H. Stanton, "The Ten Stages of Genocide," *Genocide Watch*, n.d. Retrieved from: https://www.genocidewatch.com/tenstages.

With the growth of social media, AI technologies have been designed and widely applied in the online arena. These types of tools can both foment and exacerbate genocide, while simultaneously help prevent such crimes in regard to the ten stages of genocide. Specifically, the AI tools employed in the media sector relate to stages: (1) classification; (2) symbolisation; (3) discrimination; (4) dehumanization; and (6) polarization.

Media as an AI tool for facilitating perpetration of genocide

AI tools used on social media can play a significant role in exacerbating the aforementioned stages of genocide. The first stage, classification, refers to the creation of an "us" versus "them" dichotomy between people of different race, ethnicity, religion or nationality. Such divisions may foment in the form of removing or denying a group's citizenship, stripping them of their civil and human rights. This was evident in Burma's 1982 citizenship law, which deprives Rohingyas of national citizenship. Proceeding from classification, the second stage, symbolization, develops in which names or symbols are designated to the classifications to distinguish between the different groups. Such symbols, together with hatred, can lead to the dehumanization of a group. This was the case in the Cambodian genocide, where people from the Eastern Zone were forced to wear a blue scarf. The third stage, discrimination, occurs when the dominant group employs law, custom and political power to strip non-dominant groups of their rights, such as their voting rights. This is driven by an exclusionary ideology that promotes the monopolization or expansion of the dominant group's power while victimizing the weaker groups. An example of such discrimination is the denial of citizenship to Rohingya Muslims in Myanmar, which resulted in genocide, as well as the mass displacement of refugees. Once discrimination has taken effect, dehumanization, the fourth stage, commences, in which human rights are stripped from those perceived as "different." Indoctrination of the dominant group equates the non-dominant group to animals, vermin, insects or diseases to depersonalize them. This occurs through the spread of hate propaganda, circulated in print, on the radio, on television, and on social media. In tandem, the sixth stage, polarization, refers to hate groups beginning the spread of hate propaganda and enforcing other measures that violate the human rights of the targeted group, such as laws forbidding social interaction, arresting leaders in the opposition group or passing emergency laws. For example, throughout the genocide in former Yugoslavia, Serbian news station Pale Television relayed false information, broadcasting that Muslims were a threat to Serbs. Last, stage ten, denial, occurs during and proceeding the genocide when the perpetrators deny and attempt to disguise their crimes through, for example, intimidating witnesses and blocking investigations.

These five stages can be exacerbated through the use of social media and, in turn, through AI tools in the media realm.[2] In line with stages one through three, social media platforms can exacerbate divisions and discrimination to amplify an exclusionary ideology. This occurs as such platforms allow for widespread and instant dissemination of hate propaganda, in accordance with stages four and six. The Myanmar genocide exemplifies this given that Facebook was systematically employed by the Myanmar military to incite an "ethnic cleansing" against Rohingya Muslims. It helped instigate stage one, creating an "us" versus "them" dichotomy with "them," the minority, being labeled as "illegal Bengali immigrants."[3] Moreover, the platform allowed for the dissemination of fake news, such as fabricated allegations of rape of Buddhist women by Rohingya men.[4] It ultimately portrayed the Rohingya minority as an existential threat desiring an Islamic "takeover" with mortal harm to

[2] Simon Adams, "Hate Speech and Social Media Preventing Atrocities and Protecting Human Rights Online," *Global Centre for the Responsibility to Protect*, February 26, 2020. Retrieved from: https://www.globalr2p.org/publications/hate-speech-and-social-media-preventing-atrocities-and-protecting-human-rights-online/.

[3] United Nations Human Rights Council, *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar*, 39th session, A/HRC/39/CRP.2 (September 17, 2018).

[4] United Nations Human Rights Council, *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar*.

Buddhists.[5] The Rohingya were classed as "pests," "dogs" or other animals, dehumanizing the minority and increasing polarization within society.[6]

Within the rise of social media comes the increasing application of AI tools. AI technology can be used to generate deepfakes, which contributes to the spread of misinformation and conspiracy theories that consolidate an exclusionary ideology and deepen divisions and dehumanize the non-dominant group.[7] AI algorithms can be employed by social media platforms to flag and remove hate speech. However, they often fail to account for the highly context-dependent nature of hate speech, thus rendering removal algorithms ineffective. This was the case in Myanmar, where Facebook's algorithm was unable to account for the context in Myanmar, and, thereby failed to flag hateful content. In fact, Facebook was blamed for the rise in divisions, tensions and violence in Myanmar.[8]

In parallel, AI algorithms often lead to filter bubbles, which are not only capable of personalizing the users' online searches to their preferences, but also supporting and reinforcing their beliefs. Rather than offering a wider range of opinions, the criteria applied by the algorithms may be biased and further promote adverse attitudes and misjudgements. Users, thus, view information through a narrow and possibly biased lens, which amplifies polarization within society.[9] Therefore, AI has the ability to exacerbate the stages of genocide through both facilitating and accelerating the spread of hatred and deepfakes, tailoring the content of messages specifically to vilify targeted groups. That is particularly dangerous in countries where critical thinking is not part of the education curriculum and the youth are under the influence of monopolized state propaganda and have limited access to alternative domestic news.

Media as an AI tool for the prevention of genocide

While AI can exacerbate the aforementioned stages, it can equally contribute to the existing prevention mechanisms in each stage. In general terms, prevention during the first stage, classification, can take place through establishing universalistic institutions, which cut across ethnic, religious or racial divisions and encourage an inclusive and tolerant society. Legal measures may also effectively counter symbolization and hate speech. However, legal measures require strong support from both the political elite and the population to be effective as, without such support, the outlawed attitude, speech or symbol may merely be replaced with an alternative phrase or symbol that continues poisoning the atmosphere. Regarding the next stage, discrimination may be prevented by prohibiting any form of discrimination, granting all groups in society, regardless of race, ethnicity, religion or nationality, full political empowerment and citizenship rights. In tandem, for the prevention of dehumanization, constitutional protection against incitement to genocide is necessary. As a general rule, local and international leaders should explicitly and without delay, denounce hate speech as unacceptable. For effective prevention, any incitement of genocide, including hate propaganda in the media, should be demonstratively punished and banned. In the case of stage six, polarization, it is necessary to protect moderate leaders and human rights groups along with the targeted groups, while taking strong measures against potential perpetrators, to directly affect their daily lives and activities (seizing the assets of the oppressors and denying them visas for international travel). Last, to address denial, international tribunals, such as the Yugoslav tribunals, or national courts can be established to

[5] Ibid.

[6] Ali Siddiquee, "The portrayal of the Rohingya genocide and refugee crisis in the age of post-truth politics," *Asian Journal of Comparative Politics* 5, no. 2 (2019): 89–103. Retrieved from: https://journals.sagepub.com/doi/pdf/10.1177/2057891119864454.

[7] Lisa Maria Neudert & Nahema Marchal, *Polarisation and the use of technology in political campaigns and communication* (Brussels: European Parliament, 2019). Retrieved from: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634414/EPRS_STU(2019)634414_EN.pdf.

[8] United Nations Human Rights Council, *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar*.

[9] Sukhayl Niyazov, "The Real AI Threat to Democracy," *Towards Data Sciences*, November 14, 2019. Retrieved from: https://towardsdatascience.com/democracys-unsettling-future-in-the-age-of-ai-c47b1096746e.

prove the crimes and punish the perpetrators, in combination with local justice, truth commissions and public school education to pave the way toward reconciliation.

Within these prevention methods, AI can play a significant role. AI algorithms are able to monitor communication at each level of social life, in online and offline fora alike, assess the content and indicate misinformation, deepfakes, discriminatory words and incitement of hatred. AI may also be utilized to enforce the ban on hate speech on social media platforms such as Facebook.[10] In fact, AI can help counter this cycle of misinformation and polarization through creating algorithms that present content to counter the user's existing beliefs, thus allowing the user to digest more balanced content.[11] It is important, however, that such algorithms are adapted to avoid any bias, such as racial discrimination, given that there have been several cases in which algorithms have presented a bias against minority groups. For example, one study concluded that twitter's AI algorithms were 1.5 times more likely to label a tweet by an African American as hate speech compared to other users.[12] To counter this, it is necessary to adopt a context-specific algorithm that accounts for the context in which the "offensive" term is used, looking at whether specific characteristics of hate speech are present.[13]

Artificial Intelligence in Surveillance

AI technologies used for surveillance or military purposes can also help further genocidal processes or exacerbate such crimes. Regarding the ten stages of genocide, this applies to the following stages: (5) organisation; (7) preparation; (8) persecution; and (9) extermination.

Surveillance as an AI tool for facilitating perpetration of genocide

AI tools used in monitoring systems, global surveillance or even in the military domain can play a key role in the development of the stages of genocide mentioned above. The fifth stage, organisation, starts after what was described as « dehumanisation » of the targeted groups and marks a turning point in the intensification of the genocidal process. The crime is starting to get organised by a state, a military or a militia at this step. Special units are often armed and trained to spy on, arrest and murder a specific group. The seventh stage, known as "preparation," refers to the drawing of genocidal plans and the use of euphemisms to hide the true intentions of the perpetrators - with terms such as "self-defence," "counter-terrorism," "ethnic cleansing" or even "purification." Moreover, the population is indoctrinated with fear of the targeted group and the actions taken against the group can be seen as justified by the whole population.

In the eighth stage, persecution, victims are identified and isolated - sometimes in ghettos, concentration camps or in areas devastated by famine to starve them. The victims are deliberately deprived of basic human rights, tortured, killed, abused and used for atrocious scientific experiments. It is the stage in which the mass atrocity massacres truly begin. Victims of the Cambodian genocide perpetrated by the Khmer Rouge government were placed in prisons to get tortured and killed - the best known being the S-21 prison. The ninth stage, extermination, is the intensification of the "persecution" stage. It is the stage of mass killings where cultural or religious properties, symbols or

[10] Lisa Maria Neudert & Nahema Marchal, *Polarisation and the use of technology in political campaigns and communication;* Facebook AI, "How AI is getting better at detecting hate speech," *Facebook AI,* November 19, 2020. Retrieved from: https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/.

[11] Mary L Martialay, "New algorithms could reduce polarization driven by information overload," *Techxplore,* July 31, 2020. Retrieved from: https://techxplore.com/news/2020-07-algorithms-polarization-driven-overload.html.

[12] Maarten Sap, Dallas Card, Saadia Gabriel, Yen Choi & Noah A. Smith, "The Risk of Racial Bias in Hate Speech Detection," *Association for Computational Linguistics* (2019): 1668-1678. Retrieved from: https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf.

[13] Caitlin Dawson, "Context Reduces Racial Bias in Hate Speech Detection Algorithms," *USC Viterbi*, July 7, 2020. Retrieved from: https://viterbischool.usc.edu/news/2020/07/context-reduces-racial-bias-in-hate-speech-detection-algorithms/.

goods may also be destroyed to eliminate the group’s existence from history. The burning of books written by authors from the religious or ethnically targeted groups is an example, together with the destruction of temples or museums.

During these four stages, a large number of surveillance tools monitored through artificial intelligence technologies can be used to exacerbate the perpetration of genocidal crimes. To illustrate how AI in the surveillance realm can be used to further mass atrocities, the Chinese genocide of the Uighur community will be given. This Turkic Muslim community living mainly in the Xinjiang region (northwestern China) has been persecuted by the Chinese government for at least a decade. From extensive controls and restrictions on their cultural, religious and social activities to secret detention without any legal process in state-sponsored “re-education camps,” the Uighur community is currently undergoing what many human rights experts and government officials have called “a genocide.” Although this term is not yet officially recognised by the UN and many governments and is denied by the Chinese government, this Muslim minority is - unfortunately - a good example to illustrate the use of AI technologies in mass surveillance. The Chinese government employs a wide network of surveillance cameras using facial recognition thanks to an artificial intelligence created by the telecommunication giant, Huawei. This feature is capable of detecting the faces of a person from the Uighur minority and - most alarming - to alert Chinese officials in case of “unusual behaviour” or if a person goes beyond a certain authorised-area.

China also uses an intrusive surveillance smartphone application, called « Integrated joint operations platform » in order to track Uighurs. Every little aspect of their life is spied on; a lot of information is gathered at first, but police officials are also aware through this app of their geolocation data, the person with which they communicate, the use of forbidden software such as WhatsApp, the usage of unusual amounts of electricity, etc. In the Xinjiang region, machines developed with AI-technologies are also able to check and scan IDs at checkpoints and alert the police or army in case the application or the surveillance cameras detect what Chinese officials call “suspicious activity.” Millions of Uighurs are currently experiencing the misuse of these AI-technologies. The same technologies for surveillance could be used in other areas of the world or to target other ethnic or religious groups. AI provides the capability to exacerbate the stages of genocide through mass surveillance of a targeted group and signalling individual activities to authorities.

Surveillance as an AI tool for prevention of genocide

Traditional mechanisms to prevent each stage of genocide are first presented in this section. Afterwards, the role of AI in preventing those stages are drawn on, as this technology can both exacerbate each step and contribute to better prevent them.

The fifth stage of genocide, its organisation, can be prevented by legal sanctions against powerful leaders of the state or militias organizing the genocide. Arms embargoes can be imposed on the countries or the regions involved in mass atrocity crimes. More global economic sanctions can be imposed and may have a greater impact than the arms embargos themselves. Sanctioning only the perpetrators of the genocide and not the country as a whole might be more efficient in stopping the genocide or slowing it down, without causing any harm to the targeted population or the whole population.

This is possible today thanks to the Magnitsky Act for instance, which allows the U.S. government to sanction human rights offenders by freezing their assets and banning them from entering the United States of America. The U.S. government has, in July 2020, sanctioned through the Magnitsky Act two Chinese governmental entities (the Xinjiang police station and the regional state-owned paramilitary organisations XPCC) and six officials of this region for human rights abuses against the Uighur community in the Xinjiang region.

Regarding the seventh stage, preparation, human rights experts suggest that commissions should be set up by the UN Human Rights Council to investigate violations and to recommend how to prosecute those with genocidal intent. This is rarely done as national law has to be taken into account and the "will" of perpetrating genocide is not enough alone to be judged in international law, as the act has also to be present. At the stage of persecution, the international community should mobilise in order to assist and help the victims. This means sending humanitarian aid or welcoming refugees to safe countries.

Finally, in order to prevent the ninth stage - extermination - rapid armed intervention should be mobilized by the UN or regional organizations into the country where the mass atrocities are being perpetrated while keeping safe areas for refugee escape corridors, established and monitored through armed international protection. This armed intervention should be carried out by a multilateral force supported by all the members of the UN Security Council (the P5 + 10 non-permanent members). If the UNSC is not able to coordinate its actions or agree, regional alliances could intervene rapidly and save lives.

Within these mechanisms of prevention, artificial intelligence can play a role in supporting them and in halting the genocide. For instance, as we saw above, AI can be used to detect future victims of genocide, monitor every aspect of their lives, etc. In turn, this technology could be used to identify the perpetrators of genocide and sanction them. If surveillance methods, coupled with AI, are able to identify the leaders or the militias perpetrating the genocide, Magnitsky Act sanctions can be imposed and could have an impact on the development of the genocide. More globally, AI technologies in the surveillance realm can be used to acquire proof that the genocide is happening (as it is very often kept secret) and to condemn it with consequent evidence.

Artificial intelligence can also be used to welcome victims of mass atrocities in other countries. For instance, the International Rescue Committee, alongside the Stanford University Immigration Policy Lab, has developed an AI algorithm that aims at helping immigrants or refugees enter a new country, and that also facilitates their integration by improving their employment rates (if the instructions for the algorithm are followed carefully). Finally, AI technologies can be used to guard safe areas and refugee escape corridors.

Armed intervention can also use AI tools that promote human-machine interaction and technology that avoids civilian casualties through precise attacks. Such weapons need to be used according to international law, employing technology that is precise and transparent, makes distinctions between military and civilians and enables combatants to make judgements.

Conclusion

Artificial Intelligence can play a significant role in each of the ten stages of genocide, both in the exacerbation and prevention of genocide through social media and surveillance tools. State responsibility is at the heart of the application of such tools. Indeed, if a state is willing to perpetrate genocide, it will misuse these AI tools. However, states are not the only responsible parties. Tech companies play a crucial role in the use - or misuse - of AI. Accordingly, both actors - tech companies and states - should act together in establishing an agreement at a supranational level to prevent any exploitation of AI tools that could foment genocide. Within this context, international organisations possess a key role in promoting an agreement, and in contributing to its credibility and sustainability. Under such an **Agreement on the Design and Use of Artificial Intelligence**, the UN and other supranational organisations could closely monitor the AI situation in each country and raise an alarm if the misuse of AI is such that it could lead to genocide.

Bibliography

Adams, Simon. “Hate Speech and Social Media Preventing Atrocities and Protecting Human Rights Online.” *Global Centre for the Responsibility to Protect,* February 26, 2020. Retrieved from: https://www.globalr2p.org/publications/hate-speech-and-social-media-preventing-atrocities-and-protecting-human-rights-online/.

Dawson, Caitlin. “Context Reduces Racial Bias in Hate Speech Detection Algorithms.” *USC Viterbi*, July 7, 2020. Retrieved from: https://viterbischool.usc.edu/news/2020/07/context-reduces-racial-bias-in-hate-speech-detection-algorithms/

Dholakia, Nazish. “Interview: China’s ‘Big Brother’ App. Unprecedented View into Mass Surveillance of Xinjiang’s Muslims.” *Human Rights Watch*, May 2019. Retrieved from: https://www.hrw.org/news/2019/05/01/interview-chinas-big-brother-app

Facebook AI. “How AI is getting better at detecting hate speech.” *Facebook AI,* November 19, 2020. Retrieved from: https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/

Harwell, Drew and Dou, Eva. “Huawei tested AI software that could recognize Uighur minorities and alert police, report says.” *The Washington Post*, December 2020. Retrieved from: https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/

Martialay, Mary L. “New algorithms could reduce polarization driven by information overload.” *Techxplore*, July 31, 2020. Retrieved from: https://techxplore.com/news/2020-07-algorithms-polarization-driven-overload.html

Mozur, Paul. “One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority.” *The New York Times*, April 14, 2019. Retrieved from: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

Neudert, Lisa Maria and Nahema Marchal. *Polarisation and the use of technology in political campaigns and communication.* Brussels: European Parliament, 2019. Retrieved from: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634414/EPRS_STU(2019)634414_EN.pdf.

Niyazov, Sukhayl. “The Real AI Threat to Democracy.” *Towards Data Sciences,* November 14, 2019. Retrieved from: https://towardsdatascience.com/democracys-unsettling-future-in-the-age-of-ai-c47b1096746e.

Siddiquee, Ali. “The portrayal of the Rohingya genocide and refugee crisis in the age of post-truth politics.” *Asian Journal of Comparative Politics* 5, no. 2 (2019): 89–103. Retrieved from: https://journals.sagepub.com/doi/pdf/10.1177/2057891119864454.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. "The risk of racial bias in hate speech detection." *Association for Computational Linguistics* (2019): 1668-1678. Retrieved from: https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf.

Stanton, Gregory H. “The Ten Stages of Genocide.” *Genocide Watch*, n.d. Retrieved from: https://www.genocidewatch.com/tenstages.

United Nations Human Rights Council. *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar*. 39th session. A/HRC/39/CRP.2. September 17, 2018.

Unknown. “Supporting Refugees with Artificial Intelligence.” *Verdict_AI*, n.d. Retrieved from: https://verdict-ai.nridigital.com/verdict_ai_summer19/refugees_artificial_intelligence**.**

Yuang, Yuang “Xinjiang phone app exposes how Chinese police monitor Uighur Muslims.” *Financial Times*, May 2019. Retrieved from: https://www.ft.com/content/dfec4ac4-6bf5-11e9-80c7-60ee53e6681d.