Principles: Human oversight and determination, Responsibility and accountability
Values: Human Rights
Stakeholders: Media, Private sector, Technical community

Behind the scenes of content moderation

Amanda, has just been employed as a “process executive” through a third- party firm for a social media platform, has been one of many content moderators given the heavy task of flagging posts containing hate speech, violence, misinformation. Amanda’s job is by no means luxurious, it has taken a heavy toll on her mental health since she started. She finds a few racist jokes, some neo-Nazi posts, she even came across a video of a man being beaten to death. Amanda spends less than 30 seconds on each item she finds and will do this up to 400 times a day.

One day during her shift, Amanda finds a post in her queue of someone saying that “autistic people should be sterilized”. Amanda decides to flag the post because she has deemed it to be offensive and to be a form of hate speech.

Unfortunately, her company’s policy is such that autism is not considered to be a protected characteristic in the same way that gender and race are so technically the post does not violate policy. Amanda is reprimanded for her mistake and this is being counted against her “accuracy score”. Amanda is new and has made previous errors during her shift.

Amanda decides to appeal the decision she made on the autism post, but she knows that it will most likely be overturned. Amanda is one of the many who’s job it is to purge the site of its most heinous content. Her job puts her in the face of traumatizing material on a daily basis and is expected to conduct herself with as much accuracy as possible.

Evidently, this is not possible, and she will most likely make similar mistakes again. Her job as a fact-checker on a platform with endless forms of violence will continue to take a toll on her. Perhaps this is where AI can stand to relinquish the burden from human moderators.

- Social media platforms in the private sector need to strike a balance of using AI and human moderators. Using AI where possible will be cost-effective and reduce the amount of potentially psychologically harmful content humans will be subjected to. At the same time, human input can provide the moderation process a more comprehensive result with full cultural and emotional context.

- Developers in the technical community should refine and employ the use of image recognition technology. Computer vision allows for the rapid analysis and identification of visual content and is conducted with almost perfect accuracy, thereby alleviating the workload of human moderators.

Human oversight & determination, respect and protection of human dignity, responsibility & accountability.

Know more about this case:

- "Trauma Floor", The Verge, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

- "How Facebook is using AI to combat Covid-19 misinformation and detect hateful memes", The Verge, https://www.theverge.com/2020/5/12/21254960/facebook-ai-moderation-covid-19-coronavirus-hateful-memes-hate-speech

Additional resources:

- "How to handle content moderation with the human factor in mind", imagga, https://imagga.com/blog/how-to-handle-content-moderation-with-the-human-factor-in-mind/

- "The human cost of online content moderation", Jolt Design, https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation