Title: Human Moderators behind the AI: Unseen Costs and Hidden Consequences

Abstract:

Content moderation is a topic of concern among many. Many companies have started to use AI for their platform moderation such as TikTok, Instagram, and other popular sites. Alongside this however comes a forgotten side of AI which contains its own issues: Who is moderating the AI's input? Human moderators are often subject to the numerous violations AI is seen to filter out. AI systems are excelling in automating content filtering, however they frequently falter in identifying some of the nuanced or content specific violations, leaving human moderators to bridge the gap. The result? Exposure to this type of content has been linked to severe mental health issues, including PTSD, anxiety and even depression. Moderators are often in work environments which provide little support for their needs. In this paper, I will be exploring the legal and ethical questions behind those who label AI data and the work they are subjected to.

Legal Question:

To what extent can tech companies be held legally accountable for the psychological harm suffered by human moderators tasked with managing labeling AI data content?

Introduction:

- The issue behind this is as a part of these jobs labelers have to review images and videos depicting graphic violence, hate speech, child abuse, and other traumatic material.
- This leaves individuals with emotional strain, burnout, fatigue, anxiety, depression and PTSD.
- The question is **can anybody be held responsible for this psychological damage if this is a part of the job they signed up for?**

Terms:

- What is **AI Data Labeling**?
    - AI Data Labeling is a human completing the tasks AI can't do just yet.
        - According to Amazon and its practices, data labeling is "**the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels to provide context so that a machine learning model can learn from it.**"

Relevant Laws:

While there may not be any laws to deal with this specific issue, there are laws around obscenity that should be considered:

- 18 U.S.C. § 1462- Importation or transportation of obscene matters
- 18 U.S.C. § 1465- Transportation of obscene matters for sale or distribution
- 18 U.S.C. § 1466- Engaging in the business of selling or *transferring* obscene matter
- 18 U.S.C. § 1466A- Obscene visual representations of the sexual abuse of children
- 18 U.S.C. § 2252C Misleading words or digital images on the Internet

Precedents (Key Cases):

- **Scola v. Facebook Inc, California Superior Court, San Mateo County, No. 18-civ-5135**
    - More than 10,000 content moderators "accused the company [Facebook Inc.] of failing to protect them from psychological injuries resulting from their exposure to graphic and violent imagery" (Reuters, 2024). Facebook did not admit to any wrong doing during the proceedings. A settlement of $85 million from Facebook

was to be paid to the moderators as well as $52 million to go towards a mental health treatment fund and other payments to class members.

- **Soto and Blauret v. Microsoft Corporation**
    - “Microsoft workers on the “online safety team” were forced to view photos and videos of “indescribable sexual assaults”, “horrible brutality”, murder and child abuse, resulting in severe post-traumatic stress disorder, according to a lawsuit” (Levin, 2017). Both Henry Soto and Greg Blauret state they developed symptoms of “P.T.S.D., including insomnia, nightmares, anxiety, and auditory hallucinations” (Chen, 2017). Soto reported he was finding it difficult to spend time around his young son due to the material he was forced to review at work. Microsoft disputed the claims stating the company “takes seriously its responsibility to remove and report imagery of child sexual exploitation and abuse being shared on its services, as well as the health and resiliency of the employees who do this important work.”

Examples and Hypotheticals:

In defending against Soto and Blauret, Microsoft stated they have resources in place for all employees who feel as though something in their work has caused mental health issues:

1. Mandatory psychological support: Mandatory monthly one-on-one meetings with a psychologist; monthly group meetings with a psychologist; training on how to limit the impact of viewing this imagery; quarterly psychological educational trainings for employees and their manager on how to recognize the symptoms of trauma and exercise self-care; referrals to personal mental health professionals.

2. State-of-the-art tools and technology to reduce the realism of the imagery: That blurs imagery; converts high resolution imagery to low resolution imagery; reduces large imagery to thumbnails; separates audio from video so the reviewer is not hearing and seeing at the same time; and converts imagery to black and white.
3. Strict guidelines to separate this portion of the employees' responsibilities from other work responsibilities: Employees are limited in how long they may do this work per day and must go to a separate, dedicated office to do it; they can't do this work at home or on personal equipment.
4. Immediate breaks or reassignment: At their request, employees receive a break from doing this work or a day-off. If an employee no longer wishes to do this work, he or she will be assigned other responsibilities (Levy, 2017).

I think these guidelines could be a good basis for something larger scale. The issue is if other tech companies have similar guidelines, how well are they enforcing those?

Analysis:

Tech companies need to be held responsible for the psychological harm done to employees labeling and moderating AI data. It is unethical to force an individual to obscene and violent material.

However, it is important to acknowledge the other side of the argument. In terms of ethical principles, one runs into the Utilitarian principle. Much like the trolley problem. who is considered the greater good? The millions of moderators or billions of people on the internet?

Unfortunately in this case, I would have to say those on the internet. It prevents the distribution of obscene material and from viewing dangerous content.

There is a need for human AI data moderators, but greater measures need to be taken by tech companies to ensure not just our safety but the employees as well. Having strict guidelines companies need to follow keeps them ethically checked.

Sources:

Bernal, C. A. A. (2024, August 20). *The hidden health dangers of data labeling in AI development*. The Hidden Health Dangers of data labeling in AI development. https://4sonline.org/news_manager.php?page=36940

Chen, A. (2017, January 28). *The human toll of protecting the internet from the worst of humanity*. The New Yorker. https://www.newyorker.com/tech/annals-of-technology/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity

Criminal Division, U. S. D. of J. (2023, August 11). *Citizen's Guide to U.S. federal law on obscenity*. https://www.justice.gov/criminal/criminal-ceos/citizens-guide-us-federal-law-obscenity

Garrido, L. (2023, November 27). *The Psychology Behind AI Content Moderation: Understanding User Behavior*. Checkstep. https://www.checkstep.com/the-psychology-behind-ai-content-moderation-understanding-user-behavior/

Levin, S. (2017, January 12). *Moderators who had to view Child abuse content sue microsoft, claiming PTSD*. The Guardian.

https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd

Levy, N. (2017, January 12). *Lawsuit: Ex-Microsoft Employees Claim PTSD from jobs reviewing child porn and other "toxic content."* GeekWire. https://www.geekwire.com/2017/lawsuit-ex-microsoft-employees-claim-ptsd-jobs-reviewing-child-porn-toxic-content/

Wiessner, D. (2021, July 23). *Judge Oks $85 mln settlement of Facebook moderators' ptsd claims | reuters*. Reuters. https://www.reuters.com/legal/transactional/judge-oks-85-mln-settlement-facebook-moderators-ptsd-claims-2021-07-23/