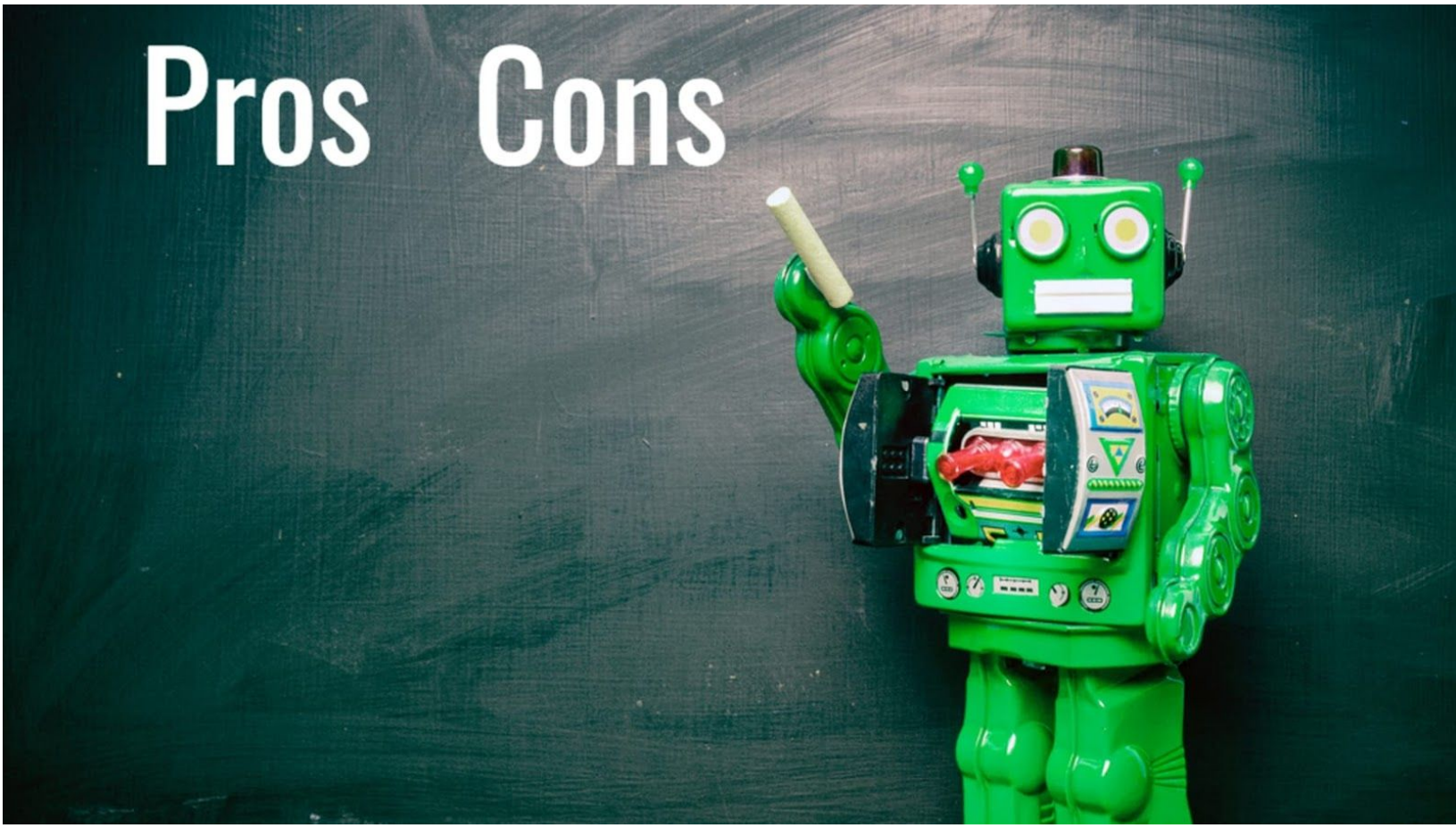
CSE373 Research Project

Winter 2021

Sean Sexton
Donavan Erickson
Melissa Truong

A Critique on AI within Content Moderation



Analyzing the Pros and Cons of Artificial Intelligence for Content Moderation

Why is AI an Issue in Content Moderation:

With the expansion of social media within our modern world, the use of Artificial Intelligence (AI) arguably seems like the perfect response to the growing struggles of content moderation. From Facebook to Reddit, we have seen an eruption of number of users and the immense scale of data within such platforms, combined with the relentlessness of violations and the need for human judgments without wanting humans to have to make them. With such issues, AI seems like the perfect solution, right? The answer: YES, but also NO. Yes if the AI is trained and used properly, but no, because AI are rarely 100% perfectly trained models and deployed properly. With this in mind, we must deploy a balance between our automated tools and human moderators, enabling us to promote a medium where everyone's voice is heard within the proper and appropriate constraints.

What are the Ethical Pros and Cons?

Some can argue that with the use of AI, it will help us, humans, from making hard decisions like "Is this comment considered racist?", "Is this tweet invoking violence?", or "Does this comment suggest some racial bias?" Obviously, it takes a lot of brain power to sift through hundreds, if not thousands, of comments, tweets, posts, etc and gauge whether they violate any of the social media's rules of conduct. Therefore, many at first glance, will be like "HECK YES, fully automate AI's. Do your thing!".

Yet looking from an affordance analysis viewpoint, full autonomy of AI can lead to a pitfall where bias can be reinforced at a massive scale. Even with machine learning and its training and testing sets, AI is only good at detecting toxicity and inappropriateness when outlined by well-defined criterias that are seen in past training sets. Beyond the well-defined criteria, the grey area becomes murky and AI will produce bias. For example, an AI can mark a comment as inappropriate and "toxic" because it discriminates against Black English when it ethically should not have. It has a hard time accounting for context, subtlety, sarcasm, and subcultural meaning. Racial and gender bias, especially

non-transparent ones, are much harder to define algorithmically. From a Harvard Gazette post, Michael Sandel mentions, “AI not only replicates human biases, it confers on these biases a kind of scientific credibility.” Therefore, since AI/ machines learn from data sets they’re fed, chances are “pretty high” they may replicate bias such as systemic disparate treatment of African Americans and other marginalized groups.

What Needs to Be Done:

To “perfect” a model for content moderation, we need to fix both the AI and human portion. The word, “Perfect” in quotes because the definition of the term is very loose. In the case of AI, we need to use better training sets that are diverse and reflect the actual diversity in speech online. By doing so, we can avoid or minimize marginalization of certain groups on social media platforms. In the case of human moderators, we need to hire more diverse people, racially and gender-wise. By doing so, we can shift in assumptions about what kinds of language are considered “healthy” versus “toxic”.

If those two objectives are not performed and taken into account by companies producing the content moderation systems, perhaps, automated tools are best used to identify the bulk of the cases, such as those with well-defined criteria, and leave the less obvious or more “controversial” identifications to human moderators. For example, computer-automated tools could be designed and optimized for identifying the most egregious, scarring content (beheadings, violent pornography, and child abuse), in order to protect human moderators from having to view it at all and users should, therefore, accept less precision in these areas. Human moderators, instead, can focus on content that is not so damaging to encounter (Gillespie). Regardless of what decisions are made, automated tools should be designed to support human teams rather than supplant them. Like mentioned in the beginning paragraph, we should be creating a relationship between the two in the world of content moderation.

Last Thoughts:

With the thoughts and ideas outlined above, more companies need to create and embody the need for balance between technology and humans. By doing so, it will help control the spread of bias and marginalizing of certain group's voices. There needs to be a conversation on how we, as a company, can be more accountable for such balance. We should not "egg on" and encourage certain beliefs and prejudice about certain groups because our content moderation system failed and marked comments from those groups as "toxic". Companies, from its developers to its big stakeholders, need to hold more accountability on the products they are producing.

Lastly, while we acknowledge that the suggestions outlined above will not solve or hand out the perfect solution to how to combat the balance between AI and human interaction, we hope to invoke an important conversation to take into account these ethical problems regarding AI in, not only content moderation, but also in our everyday lives. As Sandel says in the Harvard Gazette, "Companies have to think seriously about the ethical dimensions of what they're doing and we, as democratic citizens, have to educate ourselves about tech and its social and ethical implications — not only to decide what the regulations should be, but also to decide what role we want big tech and social media to play in our lives," .

References:

Ethical concerns mount as AI takes bigger decision-making role in more industries by Christina Pazzanese
https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/
How Automated Tools Discriminate Against Black Language by Anna Woorim Chung
https://civic.mit.edu/2019/01/24/how-automated-tools-discriminate-against-black-language/
Content moderation, AI, and the question of scale by Tarleton Gillespie
https://journals.sagepub.com/doi/full/10.1177/2053951720943234