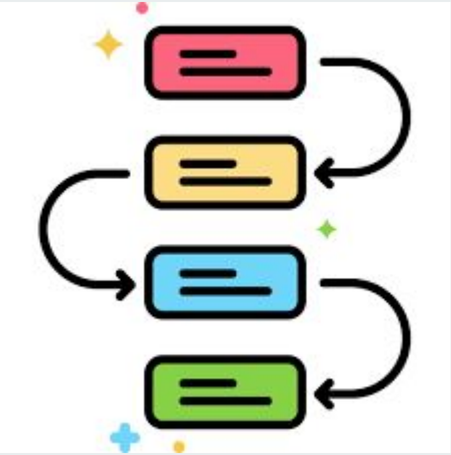
AI-Powered Social Media Moderation: Ethics and Human Cost

Pablo Mateo del Pino-Bonn Ethics Lab

Project Overview

1. The Problem- human moderators exposed to extreme content

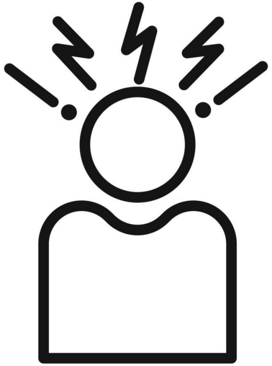
2. The Project- model based on trained AI
3. Key Observations/Conclusion-bias/political influence

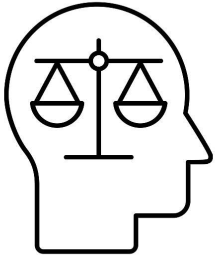


The Human Cost of Moderation

Human moderators see extreme content: violence, suicides, child abuse.

Psychological effects: PTSD, anxiety, depression, lifelong trauma.





Why Automate with AI?

Protect human moderators from trauma.

¿AI labelers?

Scale moderation to millions of posts daily.

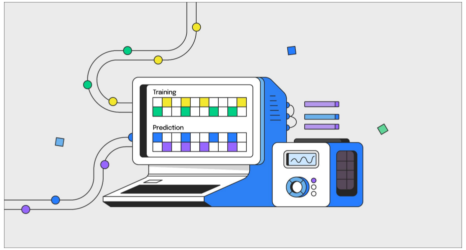
Reduce operational costs.

Current usage: Meta, YouTube, TikTok, Hive.

Hybrid model: AI filters, humans verify complex cases.

Economic Comparison

How AI is Trained – Ethical Challenges



Training datasets include real violent content, suicides, abuse.

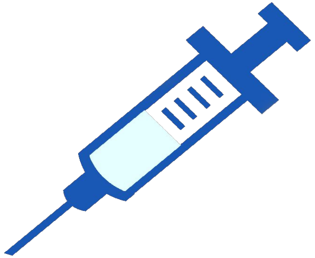
Child sexual abuse material cannot legally be stored

Human labelers are exposed to trauma during dataset creation.

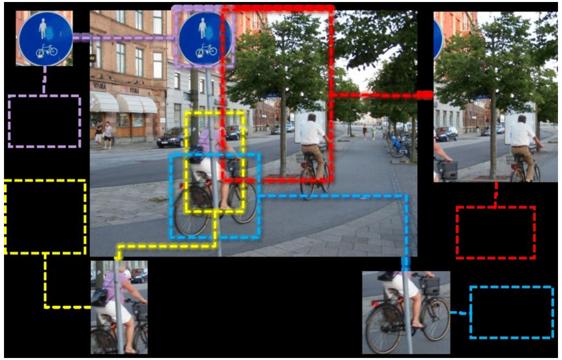
Bias, Politics, and Power

Biased data → biased moderation decisions.

Risk of “algorithmic truth”: AI may remove true content unintentionally.



Social Impact and Who Wins / Loses



Ethical Conclusion

AI can reduce trauma for future moderators.

Today, human suffering still enables the AI.

Technological progress is not free; it comes at a human cost.

Thanks for your time

“AI will likely solve the problem of traumatic moderation in the future, but someone had to see the horror to teach the machines today.”