# Moderação de Conteúdo por IA

**1. Viés e Justiça**
- Tipos de viés: **Viés de dados**: Treinamento com base em padrões de linguagem de certos grupos culturais/linguísticos (limita a capacidade do algoritmo de lidar com dados fora desses padrões) e **Viés algorítmico**: Regras que interpretam publicações de forma simplista, faltando uma complexidade de análise maior.
- Grupos afetados: Minorias linguísticas, grupos culturais com gírias/dialetos próprios (graças a falta de um treinamento específico a IA pode simplesmente considerar essas gírias e dialetos como ofensiva), jornalistas (já que precisam tratar de temas sensíveis em notícias e podem ter seu alcance limitado por isso) e ativistas (os quais passam por algo similar a jornalistas, onde tópicos mais sensíveis de protestos são censurados pelo filtro.)
- Justiça: a distribuição de riscos e benefícios não é justa, pois alguns grupos claramente são mais afetados que outros.

**2. Transparência e Explicabilidade**
- Sistemas de moderação geralmente não são transparentes: o usuário recebe no máximo uma justificativa “genérica” do porque algum conteúdo seu foi marcado como ofensivo.
- A lógica de decisão é pouco explicável para o usuário final, funcionando assim como uma “black box”
- A dificuldade de contestar decisões é grande, na maioria das vezes você não consegue restaurar seu conteúdo, e quando consegue é com limitações.

**3. Impacto Social e Direitos**
- Impacto social: risco de censura, de desinformações permanecerem online ou até mesmo a remoção de críticas legítimas.
- Mercado de trabalho: moderadores humanos são substituídos em maior parte, mas ainda são necessários em casos complexos onde a IA não consegue ter autonomia.
- Direitos fundamentais: Impacto direto na liberdade de expressão e, indiretamente, na privacidade (dados de interação sendo processados pelo algoritmo de cada plataforma). Possível conflito com a LGPD.

**4. Responsabilidade e Governança**
- Equipe de desenvolvimento: poderia ter testado o sistema com bases de dados mais diversas, implementado auditorias independentes e revisões

humanas em contextos mais específicos, além de uma análise mais humana em contestações.
- Princípios aplicáveis: “Ethical AI by Design”:
Transparência: O usuário deve poder saber claramente o porquê sua publicação foi removida, não com um comentário genérico, mas sim com uma explicação automática mais detalhada, mostrando quais palavras e frases, assim como o contexto que levou à remoção.
Justiça: A IA deve tratar os usuários igualmente, sem punir indevidamente. Isso pode ser alcançado através de um treinamento com dados mais diversificados, aprofundando em diversos grupos linguísticos diferentes a fim de evitar confusões com gírias em diferentes contextos.
“Accountability”: Deve ser claro identificar o responsável pelas decisões da IA e se ter meios de contestação com análise humana para casos onde a IA possa ter falhado.
- Leis/regulações: Além da LGPD, discussões do Marco Civil da Internet e legislações emergentes sobre moderação e IA na União Europeia (AI Act) se aplicam a esse caso.

**Posicionamento:**
O sistema atual deveria passar por um aprimoramento com treinamentos mais aprofundados em dados de grupos e contextos específicos, juntamente de uma transparência maior nas penalidades aplicadas e uma possibilidade maior de contestação com análise humana para casos onde a IA possa vir a falhar.