Expert group for the EU Observatory on the online platform economy

Work stream 5: Algorithmic Discrimination in Platform Economy

Concept note

1. Policy context

This workstream would focus on patterns or processes that lead or may lead to illegal discrimination in content moderation.

The non-discrimination right is one of the fundamental values of the Union, as for art. 2 of the TEU. Additionally, Article 10 of the TFEU commits to combat discrimination based on sex, racial[1] or ethnic origin, religion or belief, disability, age or sexual orientation, when defining and implementing the European policies and activities. Article 21 of the EU Charter on fundamental rights defines the fundamental right to non-discrimination.

This research therefore aims to investigate the possible incidence of discriminatory practices in content moderation on online platforms. This is a particularly important, yet under-researched issue. The Digital Services Act proposal sets rules to address negative effects on the fundamental right to non-discrimination, including as regards the content moderation systems used by large online platforms.

For this research, we refer to the criteria of discrimination set out in Article 21 of the Charter. Discrimination can be based on a variety of grounds, such as race, ethnicity, nationality, sexual orientations or gender, reflecting a variety of algorithmic biases or design decisions in content moderation: decontextualized language, cultural diversity, asymmetric acculturation, stereotypes and prejudices, discriminatory patterns, etc. Such a moderation is then likely to generate discrimination, reproduced in biased algorithms and in the governance framing their use by online platforms.

It is important to analyse algorithmic systems, rather than just the software (design). The technical performance, but also the checks and balances, pressure valves and human intervention, reporting of algorithmic decisions do usually define the final output. Algorithmic systems are socio-technical systems; as such, the origin of discrimination is difficult to detect. Solutions might come both at the technical level (i.e. through algorithmic design and testing) but also through the governance set around the use of the algorithmic systems.

2. Research/Policy questions and methodology

This study will consider content moderation decisions, irrespective of the grounds for applying restrictions to the content (deletion of contents, deletion of accounts, downgraded contents,

[1] The document uses the term "race" to align with current legislation. An example is art. 21 of the EU Charter which articulates the provision on discrimination using the term. This notwithstanding, the Expert Group adopts the approach established with the Directive on equal treatment between persons irrespective of racial or ethnic origin and with which "(t)he European Union rejects theories which attempt to determine the existence of separate human races. The use of the term 'racial origin' (...) does not imply an acceptance of such theories". See Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin OJ L 180, 19.7.2000, p. 22–26. Recital no. 6.

slow or no more accessibility of content), which can have negative effects on freedom to distribute or receive information but can also be discriminatory in certain circumstances. Such risks occur where content moderation decisions target specific communities and groups. For example, optimizing for job recommendations based on certain criteria can also have a discriminatory effect.

First Goal: While the literature on discriminatory effects of content moderation on online platforms is documented in the United States, the aim is to document these risks, more specifically in the Union, to properly assess the situation. Therefore, the first objective is to gather facts and use cases as evidence of discriminatory content moderation.

This first goal will combine literature review and expert interviews, on what is actually known about content moderation and discrimination in the EU, and what is not known.

This first step will illustrate specific risks and open areas of further research. Eventually, the discussion aims at feeding into the analysis of systemic risks of discrimination in Europe in anticipation of the proposed rules under the Digital Services Act.

Second Goal: The proposed DSA sets out obligations on very large online platforms to assess (art. 26(1)b) and mitigate the risks of their algorithmic systems, including content moderation, recommender and ad-delivery systems have on certain fundamental rights, including non-discrimination. The risks of discrimination identified in the DSA are subject to mechanisms to mitigate against them (art. 26 and 27). Compliance with these obligations is scrutinised by independent auditors (Article 28), and further access to data for researchers to investigate the evolution of systemic risks such as discrimination is possible though a dedicated procedure (Article 31).

In this second step, the research will seek to identify parameters that can inform risk assessments. It will open further research questions and eventually contribute to the design of appropriate risk mitigation tools for non-discrimination on online platforms. The goal is to make a (non-exhaustive) inventory of the state-of-the-art tools and the possible methods and parameters. This toolkit review will provide a better knowledge of how to conduct in practice a risk assessment, as well as the strengths and weaknesses of these tools. Platforms will probably use the existing compliance tools and some of them may not be able to meet the requirements properly and are pitfalls to be avoided.

This analysis will contribute to study the evidence of the algorithmic discrimination risks and how platforms can mitigate them based on the DSA risks assessment requirements.