AI4media
ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

www.ai4media.eu

@ai4mediaproject

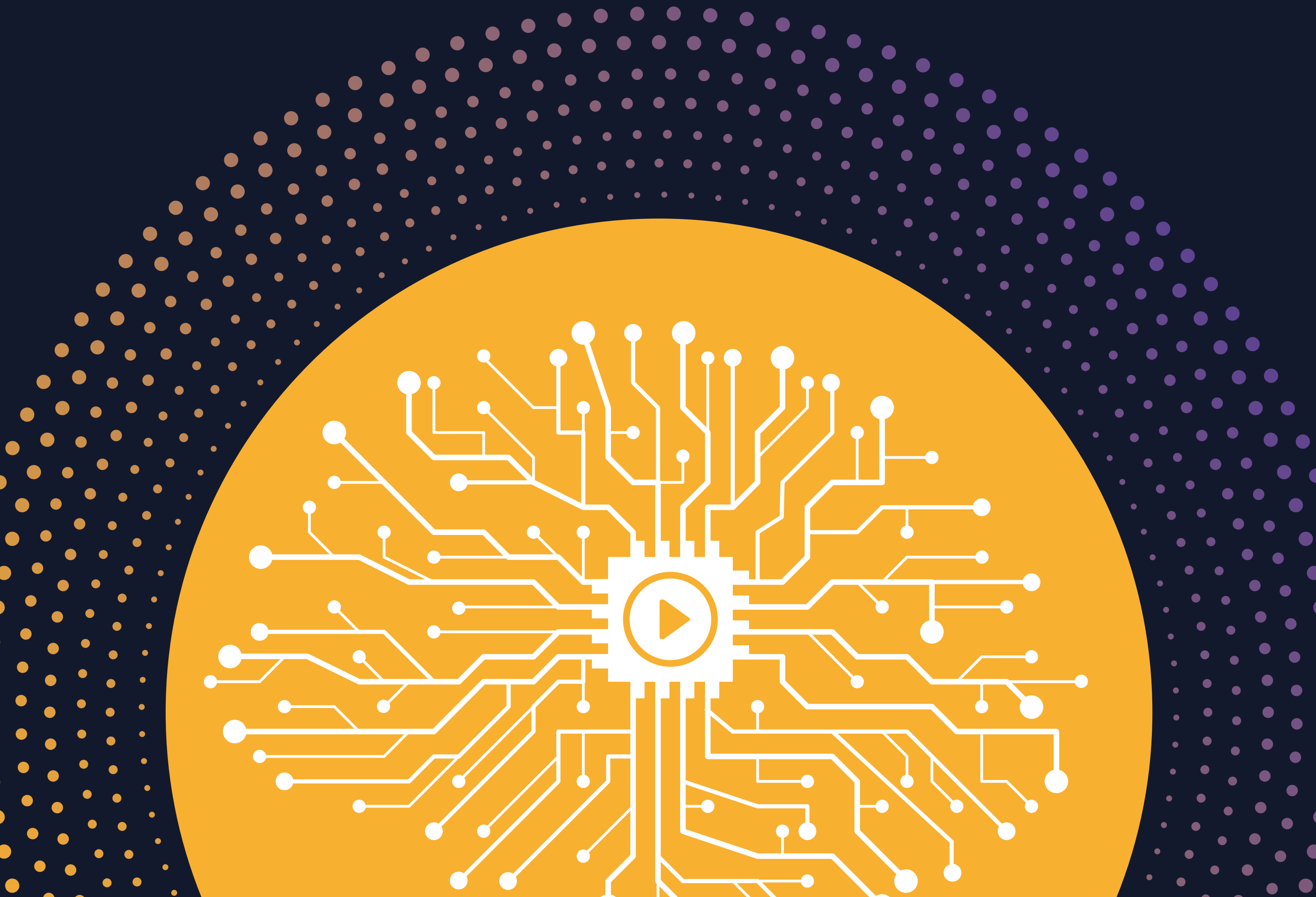
AI4Media Results in Brief: **AI and Content Moderation**

Identifying common challenges relating to the use of AI in content moderation.

Authors:

Authors: Anna Schjøtt Hansen (University of Amsterdam), Noémie Krack (KU Leuven), Lidia Dutkiewicz (KU Leuven), Aleksandra Kuczerawy (KU Leuven).

This report provides insights into the common challenges faced by industry actors using AI in content moderation. It is based on an online workshop organised in February 2023 with nine industry actors representing both small and large organisations based in Europe. The findings from the workshop have also previously informed the 'Report on Policy for Content Moderation'.



Key insights: Six core themes of AI-assisted moderation

During the workshop, six overarching themes emerged that described both what AI-assisted moderation can facilitate, but also pointed to issues of human oversight, transparency and inequalities in the quality of AI-assisted moderation.

Moderation at scale

AI systems remain key to ensuring moderation at scale and proactively classifying content or comments that violate the policies of either news outlets or social media. This can enable prioritisation of the worst content, which enables human moderators to quickly remove the most heinous content first but also protects the human moderators from seeing violent and degrading content, which can be filtered immediately. Thereby, producing better working conditions for moderators by minimizing their workload and protecting their mental health.

Hybrid solutions remain key

AI-facilitated moderation solutions cannot stand alone, they require human oversight to ensure the quality of the moderation and assess boundary cases where contextual factors or elements of humour and satire might not be detected by the AI system (i.e., false positives). Human oversight also remains key as an accountability mechanism in case moderation decisions are contested.

Supporting constructive debate

AI systems also offer different strategies of moderation by enabling the identification of 'good' content that provides constructive input to the debate. Beyond identifying and potentially removing or down-prioritising problematic content, AI can help to prioritize, for example, constructive comments as a way of positively reinforcing the practices in the communities and supporting constructive debate.

Putting numbers behind the abuse

AI systems can also contribute to putting numbers behind the abuse by illustrating, for example, the number of verbal attacks experienced by politicians or other public figures. Currently, limited evidence exists to illustrate the scale of these issues.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Key insights: Six core themes of AI-assisted moderation

Correcting the moderation gap

An important aim will also be to bridge the current moderation gap, where English-speaking contexts and spaces are currently much better moderated via AI systems, as opposed to non-English-speaking contexts. This gap was seen as also connected to the lack of willingness by large platforms to collaborate with third-party providers who develop local solutions. This gap leads to a dislocation of moderation where non-English speaking countries both experience over and under-moderation. The under-moderation emerges because of language issues, as current AI systems are predominately trained on English data, whereas the over-moderation occurs as a result of Western – often American – understandings and policies of moderation (e.g., of hate speech) imposed on these contexts.

Moderation transparency

There is a growing need to be more transparent about the ways in which AI is used in moderation practices, also induced by emerging laws in, for example, the EU. This could include publishing guidelines but also being more transparent about the ways in which the system classifies and the following process of human oversight, as well as better possibilities of contesting the decisions. However, this also requires clear transparency and explainability of the systems to the moderators for them to correct false decisions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Core challenges for using AI systems in content moderation

Several challenges were raised in relation to the use of AI in content moderation where the most predominant related to data, both in terms of access and diversity in data and noise in the data, but also data protection concerns. Equally, the lack of contextual understanding and clear definitions produced major challenges for the industry actors as well as unique issues related to evaluation and live moderation.

Data challenges

Data was considered the most important challenge, particularly because the field is now drawing from the same foundational AI model, which makes the data even more important to innovate the systems. These challenges included the following:

Access to datasets

The participants described how accessing quality datasets was an issue, as there was a lack of publicly available datasets to train the AI systems on and often the data was not from the exact context that the AI system was to be used in. For example, one participant described how they were building a moderation tool for Facebook, they could harvest data from Facebook's API, but that did not include the content that had been moderated already, which would serve as high-quality training data for their model.

Diversity in data

The participants also referred to issues of diversity in data relating to a lack of data on particularly small or minority languages and particular geographic regions. Particularly trustworthy and high-quality data was seen as highly difficult to attain when moving beyond large languages.

Dealing with noisy data

Another issue raised was problems with noisy data which resulted from inconsistent practices of moderation by human moderators. This could both be seen as a subjectivity bias but also related to the lack of clear and shared definitions of, for example, hate speech. Equally, many of the datasets risked perpetuating existing data by drawing on 'real' moderation data, which should also be mitigated when using such data.

Data protection and sharing data

Data was also connected with challenges of ensuring the data subject's data protection rights, for example, via anonymization. The GDPR was seen as a challenge for sharing datasets between multiple organisations.

Core challenges for using AI systems in content moderation

Definitional challenges

Another group of challenges related to the problems of defining what should be moderated and accounting for local contexts as well as finding a balance to uphold free speech. These challenges particularly included:

Defining content to be moderated

A main challenge related to the difficulties in defining the different forms of content to be moderated, such as hate speech, because the definition would often vary across different geographic contexts and also not be a stable definition, but rather evolving with societal and cultural changes, which is also highly situational.

Including the context

The emphasis on contextual understandings of what requires moderation challenges the use of AI, as it requires local retraining and ongoing adjustment, as societal events, such as war, change what can and cannot be posted.

Balancing moderation and freedom of expression

As many policies on moderation extend beyond what is legally defined as illegal speech and include problematic or hateful speech, a new critical question emerges on how to ensure safety, while also protecting free speech. This balancing is also related to the accuracy of the AI systems, as high numbers of false positives might threaten free speech, which might particularly occur in the moderation of small or minority languages with poorer training data.

Core challenges for using AI systems in content moderation

Challenges in evaluating AI systems for moderation

There are also specific challenges related to the lack of shared evaluation frameworks and difficulties in gaining insights into how these systems work. These challenges include:

Lack of shared evaluation frameworks

Some emerging evaluation frameworks exist for AI systems used for moderation, such as the ‘hate check’, which tests for what functions and attributes the system has. However, these frameworks are not applicable to all systems. Equally, there is a lack of shared benchmarks to help assess whether the systems work and how they work for different user groups (e.g., minority groups).

Lack of researcher access

Another challenge relates to the minimal access to information on how these systems work, particularly on large platforms where limited researcher access is given. As these systems moderate public speech, it will be important to be able to understand how norms of speech are evolving as well as have an accountability function that assesses the function of these systems. This would also better enable public deliberation over what should and should not be considered problematic speech, which currently is left to the organizations building and deploying these tools.

Challenges for live moderation

Specific challenges also emerge or are exacerbated when AI systems are used for live moderation, such as challenging the infrastructure and computational power needed and also the human oversight.

AI infrastructure

When AI systems are used to moderate speech live in, for example, comments sections, it requires a more complex infrastructure and more computational power, which might not be accessible to all organizations, inducing a gap in the access to real-time moderation in certain contexts.

Lack of human oversight

Using AI systems in live content moderation also challenges the aim to always have human oversight and the potential ramifications of mistakes made by AI systems.

Potential ways forward

The workshop also provided insights into potential ways forward that could enable the responsible use of AI systems for content moderation and alleviate some of the identified challenges. Particularly these recommendations were made relating to producing more local and open-source solutions, using transfer learning to alleviate the current moderation gap, and producing better conditions for sharing datasets.

Local and open-source solutions

One way forward focused on providing better conditions for making local and open-source AI solutions that could complement the AI-assisted moderation systems on, for example, large platforms to minimise the moderation gap that currently exists between English and non-English contexts. These conditions could be provided via either targeted funding or policy but also would require new forms of collaboration between third-party providers and large platforms, who are both important sources of data and where such solutions could have a high impact. Such local solutions could also better consider the specificities of the cultural context in which the moderation would take place, thereby, also alleviating the dislocation of moderation.

Transfer learning across regional languages

To enable a way forward with such solutions, the participants looked towards transfer learning and the ability to provide regional models rather than purely local models, as this would require less training data to retrain models in similar languages. Such experimentation had already been carried out by two participants with high accuracy rates, where only 10-20% labelled data was needed to retrain the models, which would place much less stress on small or minority languages to be fully responsible for building their own system and training them on complete datasets.

Publicly available datasets

The last way forward, which also would be incremental for any of the solutions to be attainable, was to create publicly available datasets, as this remained the core barrier for most of the projects described during the workshop. While a few examples, such datasets relating to sexism have been produced, they remain English-focused, and the quality of available data varies significantly across different geographic contexts. High-quality datasets would better enable the foundational training of regional models, which could then be retrained on smaller amounts of locally labelled and contextual data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

BACKGROUND INFORMATION

This mini report is based on an online workshop organized by KU Leuven and the University of Amsterdam as part of the Horizon2020 project AI4Media on February 6th 2023. The participants included industry participants from nine industry actors representing both small and large organizations based in Europe and in some cases representing partner organizations in AI4Media.

The workshop was conducted under the Chatham House Rules, but a participant list is provided below that provides some contextual information regarding the participants.

The purpose of the workshop was to identify the common challenges faced by industry actors who engage with AI in the context of content (comment) moderation and learn from their respective experiences on the use of AI systems assisting their content moderation efforts.

The workshop included:

- **Introductory talk:** The workshop starts with a short Introductory talk by, Distinguished University Professor Law & Digital Technology, with a special focus on AI, Natali Helberger from the University of Amsterdam (UvA).
- **Round table:** Each participant shared what they consider their main challenge when working with AI-enabled content moderation (e.g., technical, or ethical challenges)
- **Discussion of good practices:** Based on the round table the last part of the workshop focused on identifying potential good practices. Two discussants were also present to help and guide the discussion: Bernhard Rieder, Associate professor in New Media and Digital Culture at the UvA and Aleksandra Kuczerawy, postdoctoral researcher at KU Leuven focusing on online Content Moderation and the Rule of Law.

Participants

The participants who were recruited via the network of the organizers and the wider consortium included:

- A European company producing image recognition solutions for developers and businesses.
- A European consultancy doing content moderation analysis.
- An AI4Media-funded project focusing on robust and adaptable comment filtering.
- A prominent newspaper from Austria.
- A UK company developing socially Responsible AI for Online Safety. They develop AI-powered tools to find and stop toxic content.
- A German local broadcast media production and distribution company doing responsible journalism and professional entertainment.
- An American technology company that owns a very large online platform(s).
- A European company developing trustworthy, transparent and explainable human-centred AI solutions that read and understand large amounts of text.
- A researcher from a well-known university in the Netherlands and consultant for the United Nations Department of Political and Peacebuilding Affairs (DPPA) Innovation Cell.

CONTACT AND MORE INFORMATION

This report was produced by Anna Schjøtt Hansen, Noémie Krack, Lidia Dutkiewicz and Aleksandra Kuczerawy for more information or questions feel free to contact them.

To cite to this document: A. Schjøtt, N. Krack, L., Dutkiewicz & A. Kuczerawy, 'AI and Content Moderation: Identifying common challenges relating to the use of AI in content moderation'(AI4Media Results in Brief), December 2023, AI4Media.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911