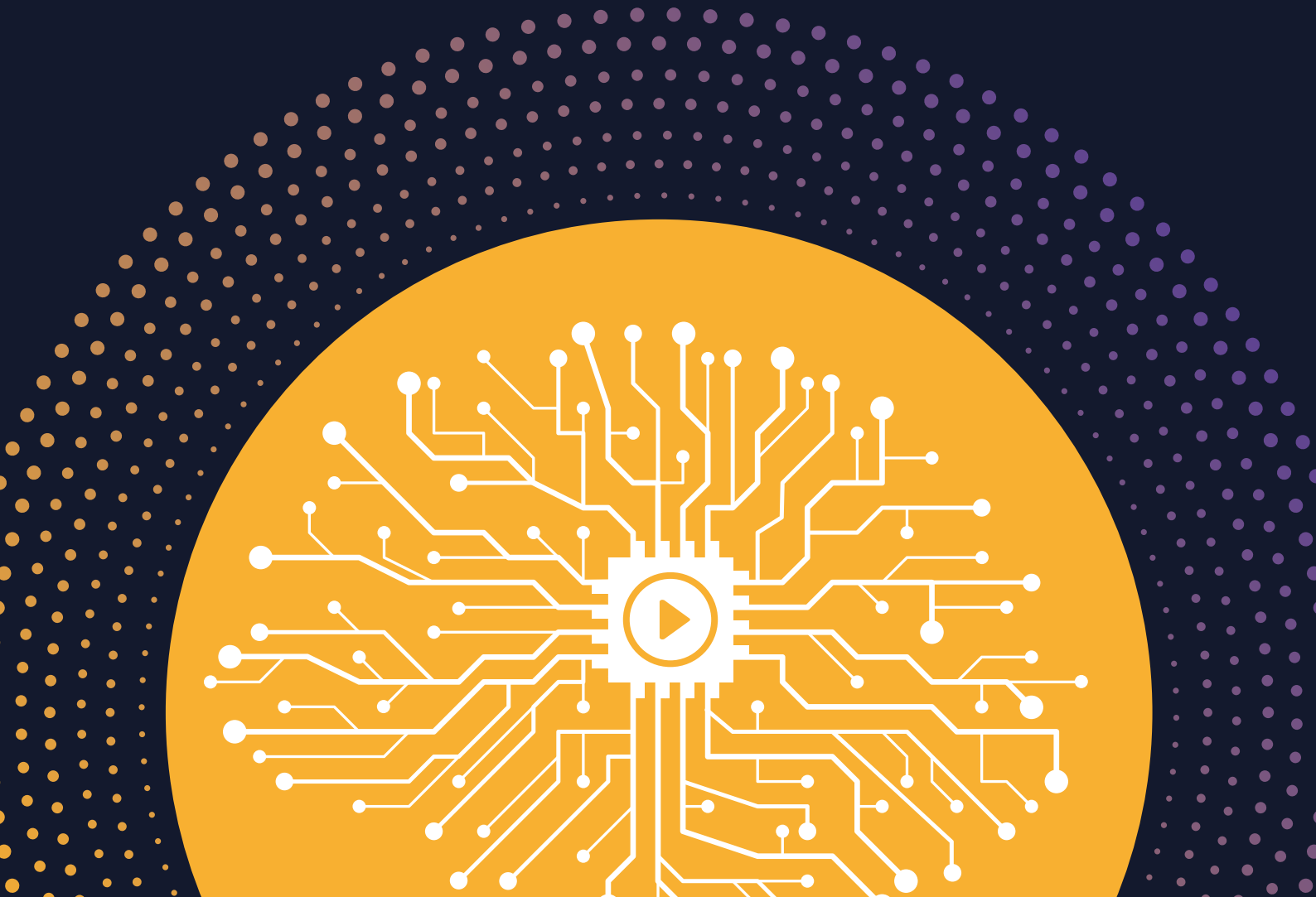
AI4media
ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

www.ai4media.eu
@ai4mediaproject

AI4Media Results in Brief: **Measuring the success of Recommender Systems for Media**

Authors:
Anna Schjøtt Hansen, University of Amsterdam, Natali Helberger, University of Amsterdam

This report provides insights into the current challenges, potentials and good practices for measuring the success of recommender systems for media. It is based on an online workshop organised in April 2022 with six commercial and public service media organisations from Europe.



Key insights: Three core questions

During the workshop, three core questions emerged that characterised the discussion on how to measure the success of recommender systems in media, namely (1) what to optimise toward, (2), how to balance costs and (3) how to include the voice of the audience. These questions were discussed via potential suggestions of how to address these questions, but equally via the challenges related to answering them.

Optimising toward what goal?

Participants raised the question of **what should drive the optimization** of the recommender systems deployed by media, which raised a related question of **what metrics to use in evaluating the systems.**

Suggestions for drivers of optimisation:

- Measuring the success of a recommending system not only by what it shows to us but by **what it shouldn't show** to us to avoid harmful effects of recommenders (e.g., inducing rabbit holes or harmful feedback loops).
- Measuring the success not only on an individual level but also how the system manages to **capture community interest** (e.g., addressing minority groups).
- Measuring the success by the **diversity in content discovery**, so that users are exposed to content they might not normally have come across.
- Measuring the success according to **editorial values**, so that the system recommends in a way that aligns with the local editorial mission of the media organisation.

Challenges for evaluation:

- **Moving beyond click-based accuracy metrics**, as this remains the common standard to test and evaluate the performance of the systems against, particularly when deploying systems by third-party providers.
- **Operationalizing alternative metrics** proves a challenge as editorial values can be hard to translate into metrics and require a close editorial-tech collaboration, which is often even more difficult when deploying systems by third-party providers. Being able to do so also requires investments on the side of the organisation, in skills, expertise as well as time and room for experimentation.
- **Lack of benchmarks**, to test how the system performs in comparison to other systems that optimise towards media values, as existing benchmarks also remain focused on click-through rates or have simplistic understandings of, for example, diversity.

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

Key insights: Three core questions

How to balance short and long-term implications?

During the workshop the participants also raised the question of how to balance the **economic costs of building and scaling recommender systems**, but also how to balance **non-economic costs such as losses of audience privac**y. These 'costs', which are both economic and non-technical represent the core question of how to balance the short and long-term implications related to implementing recommender systems.

Suggestions for ways of balancing implications:

- Making **infrastructures as efficient as possible** to reduce costs of scaling recommender systems and being conscious of what is **minimally required to technically** deliver on the proposed mission.
- Approaching **data management as a service** in which users have a choice of how their data is used and managed (e.g., what type of data or opt-out), to also **maintain trust in media**, which can be threatened by misuse or untransparent use of, for example, personal data.
- Being **conscious of data decisions**, and considering using the minimal amount of data needed and avoiding using personal data, when possible, as opposed to engaging with data practices through a lens of more data is better.
- Being **transparent to the audience regarding what data is collected and how it is used** within the systems.

Challenges in balancing implications:

- **Deploying ML at scale** requires high costs for data storage and training, which can challenge even large well-resourced media organisations.
- **Cloud computing as a cost reduction** is often viewed as one solution as it provides options for scaling up and down but also induces new dependencies on commercial infrastructures and risks adding greater costs to, for example, audience privacy.
- **Balancing the cost of transparency** for the user experience, as large technical descriptions might produce barriers to audience engagement without delivering the goal of transparency.

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

Key insights: Three core questions

Where is the voice of the audience?

The last core question that was raised by the participants, was how to better include **the voices of the audience in the development of recommender systems**, but it was also questioned **to what extent this should be the case.**

Suggestions for ways of balancing implications:

- **Ensuring strong editorial-tech collaboration via cross-disciplinary teams** where the editors speak on behalf of the audiences' interests.
- **Including the audience in defining the optimization goals for the recommender system** via, for example, focus groups with an emphasis on getting insights from **marginalised groups** whose perspectives are often threatened to be left outside of these discussions of how to assess recommenders.
- **User control**, not only over data collection and usage, but also over the recommendations themselves via, for example, different ways to tweak the optimising goals of the recommenders towards more topic or genre diversity, and being able to provide feedback.

Challenges in balancing the role of the audience's voice:

- **Balancing the audience's voice against editorial concerns** with the aim to ensure that the mission of the publication remains present, as the aim of media is also to push perspectives and democratically inform about multiple perspectives (at a minimum by exposing the audience to such diverse content).
- **How to meaningfully include audience perspectives**, provides a challenge as often the groups that are most at risk of being marginalised are harder to reach and the discussion can easily become too abstract.

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

Good practices and policy recommendations

During the workshop, several best practices were discussed together with some potential recommendations for how policy could support these best practices. The best practices included insights into establishing evaluation frameworks for assessing and testing recommender systems, which are presented via a case study based on a presentation during the workshop and in a timeline that raises evaluative questions not only at the end but throughout the pipeline of making a recommender system and addresses the compromises that must be made here.

CASE STUDY

A VALUE-DRIVEN FRAMEWORK FOR THE EVALUATION AND DESIGN OF AI SYSTEMS FOR NEWS

The case study is based on a presentation by a large Danish commercial media organisation who are developing an in-house recommender system with the aim to (1) deliver a more engaging and informing news experience, (2) deliver AI systems that are aligned with the outlet's editorial mission and minimise the dependence on the tech giants by building in-house, and (3) push for healthy norm setting for the use of AI in news, which is currently still in development.

As part of the project, the media organisation has developed a value-driven framework that outlines four different value domains, which need to be considered and balanced during the design and evaluation of AI systems. These include:

- Moral responsibility via ethical AI values (e.g., transparency/explainability, fairness, justice, privacy, and avoiding harm).
- Public Service via journalistic values (e.g., truthfulness, objectivity, credibility, pluralism, relevance, identification, and sensation).
- Economic value creation via business values (cost-effectiveness, superior value to attractive customer segments and strategic independence).
- Technical Excellence via technical values (best performance against optimization goal, efficiency in resource use, such as data and computing).

These four value domains can help to inform the compromises needed to be made throughout the process of developing a recommender system and to balance the known challenges of AI, such as the lack of ethical and journalistic considerations in commercial recommender systems.

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

TOWARDS ADAPTABLE AND PROCESS-ORIENTED ASSESSMENT FRAMEWORKS

One insight that stood out during the workshop was the need for more adaptable and process-oriented assessment frameworks of recommender systems. It was highlighted how each media organisation will have their own values and goals that need to be foregrounded in the assessment framework, but also how it would have to be more of an assessment pipeline or procedure because assessment must be made throughout the developing process, not only in the final stages. Below is a figure that outlines evaluative questions across the pipeline and what compromises are raised.

1. **Mission definition:** Initially it should be discussed what values are essential to the media and discussed how these can be optimised toward in the future recommender system. This process should be inclusive in the sense of including representatives from the different parts of the organisation (management, legal, marketing, technical, journalists, etc.). This could be represented in a framework, such as the one presented above.

    - **Compromise(s):** The framework can help to assess what values can be negotiated and which are non-negotiable, but it will also be important to ensure how questions of audience voice will be addressed.

2. **Buy from a third-party provider or build in-house:** This decision should be based on the above goals and values, in-house expertise and resources, business case and whether the system will be mission critical (see also Council of Europe Procurement Guidelines).

    - **Compromise(s):** Such decisions might be impacted by cost and resource efficiency concerns, and it will be important to assess when it will not be acceptable to use a third-party provider.

3. **Data collection and cleaning:** Consider the dataset to ensure the data is not biased in ways that counteract the goals set out, and that training data sets or the pool of information to draw on is sufficiently diverse (e.g., due to the composition of the dataset) and assess how the data choices will impact the performance of the model.

    - **Compromise(s):** Data collection and cleaning might be impacted by the quality of the data available, which might not be of high enough quality and could lead to compromises in expected performance. In the consideration of what data is the 'best; to use there might also emerge trade-off, where 'good' data will affect the values of data privacy of the audience.

4. **Model selection, training, and testing:** Consider different recommender models (e.g., collaborative or content filtering) and how they can contribute differently to the goals. Testing multiple models allows for an assessment of how they produce different effects.

    - **Compromise(s):** resource constraints might impede testing a large variety, so it will be important to first understand what is at a minimum required to gain the needed insights. Equally, training the models is expensive and there might be compromises in how much training is possible at a loss of performance of the number of tested recommender systems.

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

5 **Testing across locations:** Testing the different models across locations on the news sites might reveal how they might perform better or worse in some locations and where they might induce unwanted harm, allowing for a discussion of how to compose the recommender infrastructure to best reach the goals set out.

- **Compromise(s):** There might be conflicting goals between increasing, for example, engagement and diversity, so evaluations at this stage will need to compromise on what values matter at what parts of the news site.

6 **Testing beyond accuracy:** Testing the models for different metrics also enables nuance regarding how some models might enable other values beyond engagement, such as diversity in exposure and consumption.

- **Compromise(s):** It might be difficult to operationalize measures beyond accuracy and some compromises might arise in how and what can be measured.

7 **Scaling:** When scaling it will be necessary to assess what computational power will be needed and what efficiency measures are attainable.

- **Compromise(s):** Here it will be necessary to decide, where technical compromises can be made in terms of running the systems live or online a few times a day or assess whether a lower performance might be acceptable if it significantly reduces costs.

POLICY RECOMMENDATIONS

During the workshop, concrete suggestions for how to create the conditions for better measuring the success of recommender systems were also discussed and here three main conditions were highlighted to be better supported by policy.

Improving knowledge and negotiating power when working with third-party providers: This could include the ability to ask for optimizations beyond click-rate through accuracy to alleviate the current market gap in which only certain media organisations have the agency to adjust their recommenders.

Better benchmarking practices: This could include the ability to benchmark against other media systems and with metrics beyond accuracy, which could help better illuminate the public value of recommender systems.

Better industry-academia collaborations: This could include collaborations on technical solutions but also to help develop value frameworks, which could help reduce the costs of the projects and support more responsible development.

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:

- A short case study presentation by the project leader of a recommender project at a Danish media organisation who presented insights from their recent tests and evaluations of their recommender systems.
- A roundtable by all participants where they each reflected on how they had previously engaged with the question of what quality criteria are important when evaluating recommender systems in the media context.
- A shared discussion focusing on sharing knowledge and developing some initial core principles or questions that should be considered by the media and policymakers.

CONTACT AND MORE INFORMATION

This report was produced by Anna Schjøtt Hansen and Natali Helberger for more information or questions feel free to contact them.

To cite to this document: A. Schjøtt & N. Helberger, 'Measuring the success of Recommender Systems for Media' (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