Responsible AI in Media Organizations: Four Case Studies Implementing Ethical Tools in Practice

Maaike Harbers[1], Sophie Horsman[2], Pascal Wiggers[2], Huib Aldewereld[3], Marcio Fuckner[2], Coert van Gemeren[3], Oumaima Hajri[4], Floor Schukking[3], Nathalie Stembert[1]

[1]Rotterdam University of Applied Sciences

[2]Amsterdam University of Applied Sciences

[3]Utrecht University of Applied Sciences

[4]Autoriteit Persoonsgegevens

m.harbers@hr.nl, s.horsman@hva.nl, p.wiggers@hva.nl, huib.aldewereld@hu.nl, m.fuckner@hva.nl, coert.vangemeren@hu.nl, floor.schukking@hu.nl, n.stembert@hr.nl

Abstract

Ethical tools are frequently proposed as a means to promote the design and implementation of responsible Artificial Intelligence (AI). Yet many organizations designing and deploying AI make only limited use of ethical tools. This study explores the application of ethical tools for responsible AI in the media sector through four case studies conducted at three Dutch media organizations. Each case study involves the application of an ethical tool to improve the responsible design, development, or deployment of an AI application. The findings reveal that successfully implementing ethical tools is highly contextual, and requires more than their mere availability. Tools must be selected, adapted, or even partly developed to align with specific challenges. Additionally, successful adoption of ethical tools requires organizational awareness of AI ethics, knowledge of mitigation strategies, and governance supporting responsible AI. These insights thus highlight the importance of contextualization and organizational readiness for establishing a responsible AI practice.

1. Introduction

Artificial Intelligence (AI) is increasingly used in the media industry, for instance, for the automatic creation, personalization, distribution and archiving of media content (Trattner et al. 2022; Elahi et al. 2022). This rapid development, however, raises concerns in society and the media sector itself. There are worries, for instance, about the creation of deep fakes, the spread of disinformation and extreme content through algorithms, algorithms reinforcing and strengthening existing biases and stereotypes, and intellectual property infringements by large tech companies using media content to train generative AI models. To address these issues, many media organizations are creating or have created their own governance structures for the responsible design, development, and deployment of AI applications. Yet, due to the fast developments in AI technologies and upcoming regulations (such as the European AI Act), and because there few established practices for how to design, develop and deploy AI in a responsible way, media organizations are struggling with this process (Mioch et al. 2023).

The challenge to design, develop and deploy AI in a responsible way is not unique to the media domain. With the growing use of AI, in recent years, an increasing amount of attention has been paid to the ethical issues associated with AI. Many organizations and governmental bodies have formulated principles and guidelines for responsible AI. For example, Jobin, Ienca and Vayena (2019) identified 80 sets of AI guidelines that had been proposed up to 2019. Whilst a useful start, these principles and guidelines are not sufficient for realizing responsible AI in practice as they are general and high-level, and often lack practical instructions on how to operationalize AI ethics in practice (Morley et al., 2020). In the past few years, several ethical tools have been proposed to translate high-level responsible AI principles to practice (see e.g., Ayling and Chapman 2022). Ethical tools can take different forms, such as checklists, design methods, training, and prototypes. Yet, ethical tools alone still do not close the gap between principles and practice, as AI practitioners need the awareness, knowledge and skills to deploy the tools that are available (Morley et al. 2023), and as responsible AI practices have certain organizational requirements (Rakova et al. 2021). This paper therefore aims to contribute to (re-)design, the practical use and implementation of ethical tools for responsible AI in the media sector.

The work presented in this paper builds on previous work, in which we interviewed employees of Dutch media organizations about their beliefs and practices about responsible AI (Mioch et al. 2023). The main results of this study were that the media organizations included in the study all saw the importance of responsible AI, and were all working on creating a responsible AI practice, but none of them had a fully established responsible AI practice. Although there was some knowledge of ethical tools within the organizations, these tools were hardly used. An important issue that hindered the creation of a responsible AI practice according to the respondents in the study, was that existing ethical guidelines and tools for AI were not specifically geared to the media context. However, it is not clear to what extent more media-specific ethical tools would lead to better responsible AI practices, and what else is needed to establish successful responsible AI practices. Therefore, in this study we aim to get a better understanding of what it takes to apply ethical tools in practice, paying attention to the suitability of ethical tools, as well as social, cultural and organizational aspects affecting the adoption and use of ethical tools.

The use of ethical tools to achieve a responsible AI practice in actual media organizations is a complex challenge, in which it is impossible to control all factors in a research study. Therefore, we chose to follow a Research-through-Design (RtD) approach (Zimmerman, Forlizzi and Evenson 2007; Stappers and Giaccardi 2017). In this approach, research is conducted by creating so-called artifacts that transform the world from its current state to a more desired state. Through designing these artifacts and studying their effects on the world, new knowledge and insights are gained about the problem and possible solutions. In this study, we performed four case studies, which each involved the (re-)design and application of an 'ethical tool' in the context of a concrete AI project at a media organization in the Netherlands. The results of the four use cases are combined to gain insights and draw overall conclusions regarding responsible AI in media organizations.

The remainder of this paper is organized as follows. Section 2 discusses related. Section 3 describes the four case studies. Section 4 provides a discussion of the results of the case studies and the insights they provide into responsible AI practices in media organizations. Section 5 ends with a conclusion and suggestions for future work.

2. Related Work

This study aims to assist media organizations in navigating the transition towards a responsible AI practice. Therefore, in this section we primarily focus on empirical and/or applied work on operationalizing AI ethics in general (Section 2.1), and tailored to the media sector (Section 2.2).

2.1 Operationalizing AI Ethics in Practice

In the past couple of years, many tools and methods have been proposed to bridge the gap between high-level ethical principles and practice. Morley et al. (2020) proposed a typology of tools and methods enabling AI practitioners to translate principles into implementable design practices. They systematically reviewed 106 existing responsible AI tools and methods and mapped them against various components of the AI development lifecycle. Other researchers have also conducted systematic reviews and/or proposed frameworks to describe ethical AI tools and methods (Ortega-Bolaños et al. 2024; Schiff et al. 2020; Zhou and Chen 2023; Agbese et al, 2023; Prem 2023). Ayling and Chapman (2022) compared existing guidelines and ethical tools with validated methodologies from various disciplines, identifying several gaps in the current landscape of AI ethics tools.

Morley et al. (2023) concluded that the tools and methods considered in their earlier work, did not actually close the gap between principles and practices, as long as AI practitioners themselves do not have the knowledge and/or methods and tools to operationalize AI ethics in practice. In their follow-up study, they followed a mixed-methods approach consisting of a survey and interviews with industry professionals to examine the level of awareness and motivation of AI practitioners regarding ethics, and the types of assistance they currently have and seek when trying to implement ethics in practice. One particularly interesting outcome was a disconnect between the availability and demand for ethical design resources, meaning that in all cases there were more practitioners that deemed ethical resources (from ethical principles, ethical frameworks to technical toolkits) useful, than there were people having access to them.

Rakova et al. (2021) conducted semi-structured interviews with industry professionals to identify common challenges, ethical tensions and organizational structures that currently support or hinder an effective implementation of responsible AI. They concluded that a lack of accountability, ill-informed performance trade-offs and an overreliance on external pressure currently hinder responsible AI in practice. However, they also noticed emerging practices that are beneficial to steer organizations towards more responsibility, such as the use of frameworks and metrics, and the proactive evaluation and mitigation of ethical issues.

Benjamins, Barbardo and Sierra (2019) described a case of a large organization that introduced a company-wide methodology to minimize the risk of undesired consequences of AI. The methodology includes AI principles, awareness and training, a set of questions to be asked in the development process, specific tools to help answering these questions, and a governance process defining responsibilities and accountability. They stress the importance of this methodology but also state that they consider their methodology as a starting point and that more research is needed to establish solid responsible AI practices.

2.2 Responsible AI Media Practices

In another line of work, researchers have directed their focus towards the media sector, recognizing the specific ethical considerations associated with the integration of AI into journalistic and media practices. Helberger et al. (2022), for instance, have contributed significantly by arguing for a normative perspective on journalistic AI. Their work emphasizes the necessity of translating normative ideals into tangible responsible AI technologies in the media landscape. They also stress the importance for media organizations to create space, capabilities and financial means to invest in responsible AI.

Elahi et al. (2022) focus on responsible recommendation of media content. Recommender systems are frequently used AI systems within media organizations, to recommend users tailored content on online platforms. There is a risk that these systems recommend too much personalized (and one-sided) content, which in the long term could lead to the amplification of filter bubbles and polarization. Therefore, various researchers are examining ways to design recommender systems with diversity in recommendations as a design criterion (Helberger, Karppinen and D'acunto 2018; Vrijenhoek et al. 2021).

In another study, researchers and engineers from Spotify describe what is necessary to move beyond literature on bias to implementing these insights in practice (Cramer et al. 2018). They describe that next to research and analysis, it is necessary to develop processes that can be integrated in existing product cycles. Furthermore, they state that it is also necessary to engage with external communities to exchange lessons learned with one another. In this study, a case study is described in which voice technology could be biased by misrecognizing alternative spellings which are often used in the music industry. They suggest that crowdsourced pronunciations could help to overcome this obstacle.

Helberger and Diakopoulos (2023) studied the implications of the European AI Act on media and journalism. According to them, most applications of AI in the media do not directly fall in the category of high-risk AI systems. However, depending on the use case, certain (mostly safety and transparency related) obligations hold. Therefore, it is crucial that media organizations prepare and get ready for the upcoming regulations. In this study (Helberger and Diakopoulos 2023) also propose that researchers should collaborate with media professionals, both "to unlock the potential of AI", but also to assist in the "transfer of critical skills" to evaluate new innovations.

2.3 Concluding Remarks

We conclude that even though many ethical guidelines and tools have been developed, these do not always land in practice. There are multiple studies that greatly contributed to the systemization of the (often fragmented) knowledge, methods and tools available regarding responsible AI. There are also some studies that have researched (obstacles hindering) the application of ethical tools, but research on this topic is still in its early stages. Research within the media sector also recognizes the importance of developing established responsible AI practices, and the current lack thereof. However, though many researchers stress the importance of developing responsible AI in practice, there is little work in which the operationalization of AI ethics has been researched through empirical work in which ethical tools are applied in a realistic setting for actual real-world problems.

3. Four Case Studies

To better understand what it takes to implement ethical tools in practice, we conducted four case studies at three prominent media organizations in the Netherlands. Following a Research-through-Design approach (Zimmerman, Forlizzi and Evenson 2007; Stappers and Giaccardi 2017), in each of these case studies, we applied an (existing, adapted or self-developed) ethical tool to support the responsible design, development or deployment of an AI application within a media organization, in collaboration with employees of the media organization. Table 1 provides an overview of the four case studies, showing which media organization was involved, which AI application it concerned, which ethical tool was used, and which process was supported by the ethical tool, respectively. The case studies all addressed actual ethical questions that the media companies had regarding the design, development or deployment of an AI application.

The three media organizations involved in this research are all prominent Dutch media organizations. MO1 is a public organization that manages one of the largest digitized media archives in the world. MO2 is an all-round media and entertainment company, with a focus on television. MO3 is a public organization that coordinates the programming of all public media networks, channels, and platforms. All three media organizations signed the Declaration of Intent ("Intentieverklaring") for responsible use of AI in the media sector (Media Campus 2021). The Declaration was created in 2021 by some of the most prominent Dutch media organizations and signed by fifteen major media organizations in the Netherlands. It contains media-specific ethical principles that are based on the ethics guidelines for trustworthy AI from the European Commission (HLEG 2019).

Case study	Media Organization	AI application	Ethical tool	Supported proces
1	MO1	Speaker labeling	AI impact assessment (Government of the Netherlands 2023)	Procurement and development
2	MO2	Music selection	AI impact assessment (ECP 2018)	Evaluation and deployment
3	MO2	Speech recognition	Interactive model cards (self-developed)	Measurement and documentation
4	MO3	Content recommendation	Prototyping (self-developed)	Design ideation

Table 1: Overview of the four cases studies.

The AI applications and ethical tools were selected in close collaboration with the respective media organizations. This selection process generally involved that we presented our aims and ambitions to our contact person(s) at the media organization, that they subsequently presented one or multiple AI applications they were currently working (sometimes after an internal inquiry), from which we, as researchers, and the media professionals collaboratively selected an AI application. Subsequently, as researchers, we suggested one or several possible ethical tools to apply to that case, from which one was selected, which again involved a joint decision between researchers and professionals at the media organization. This process often took about three to four meetings, which were needed to establish common ground, communicate goals (researchers), make internal inquiries (media professionals), sometimes finding the right people (media professionals), and research possible tools (researchers). We tried to stay as close to the wishes and needs of the media professionals as possible. Although this made it harder to control the direction of the research, it helped to establish trust and willingness to cooperate from the media professionals, and it also helped to work on case studies that were relevant, timely and realistic.

The case studies were performed by different researchers in parallel. There were regular meetings among the research team to update each other on progress and to exchange experiences. The outcomes of one case study may have influenced other case studies, but they were not explicitly used to inform the next case study. In the remainder of this section, we describe the four case studies in more detail.

3.1 Case Study 1: AI Impact Assessment for Speaker Labeling

Use Case

This case study concerns the (automatic) labeling of speakers at MO1. MO1 is a public service media organization that creates and maintains a large archive of audiovisual and other media-related content. Part of the archive is a database that includes information (meta-data), amongst other things, on who is speaking at each timestamp. These labels are automatically produced by an AI model and used internally and externally. This information can be used, for instance, to investigate the representation of various groups in television shows, e.g., the gender balance of guests in a specific talk show. Another use of the AI-produced labels could be to follow the career of a single actor, producer or comedian.

At the time of this case study, MO1 was outsourcing the process of building, maintaining, and deploying the AI model labelling speakers to a supplier. The contract with their supplier was about to expire, and the organization had to decide whether to renew the existing contract, switch to another supplier, or develop an in-house AI model for speaker labelling. According to MO1, strategic decisions such as this one were typically made “on the fly”. They were, however, aware that the decision involved ethical considerations, and they were interested in using an ethical tool to support the decision-making process.

Ethical Tool

The ethical tool selected in this case study concerns the Dutch national government's AI Impact Assessment (AIIA) (Government of the Netherlands 2023). An AIIA usually consists of a list of questions that can be used to evaluate the impact of a specific AI system. An AIIA seemed to be a useful tool for this situation, as MO1 wanted to consider multiple options for the AI application and understand its impact and related ethical issues in a broad sense. The AIIA from the national government was chosen specifically because it was one the most up to date AIIAs available at the time (considering the drafts of the AI Act at the time). Moreover, it was an adaptable AIIA, and provided suggestions for how to adapt it for specific purposes.

To prepare the deployment of the AIIA, the researchers and a media professional at MO1 determined how they wanted to use the AIIA. The list of questions in the AIIA is divided into three categories: (1) blue questions that are mandatory to facilitate a conversation about the desirability of the AI system; (2) green questions, which are more concrete and case-dependent follow-up questions; and (3) red questions that are specifically tailored towards high-risk AI systems (as indicated by the drafts of the AI Act). For each question, the researchers and the media professional determined its relevance for this case and whether to include it or leave it out. Furthermore, questions were ranked by priority. The subjects ‘fundamental rights and fairness’, ‘accountability’, ‘technical robustness’, ‘data governance’ and ‘risk management’ were picked as the main topics to discuss in the workshop.

The AIIA was used in an in-house ethical workshop at MO1 with twelve participants, consisting of employees from MO1 with varying expertise, researchers and graduate students. At the beginning of the workshop, two short presentations were given: (1) an employee of MO1 introduced the AI system, explaining how it worked and how they used it; and (2) a researcher explained the ethical tool, including the modifications that were made. Then, two groups were formed. The first group focused on the technical challenges, while the second group focused on the social, ethical and legal challenges. The workshop was hosted in a creative collaboration room, to enable participants to get out of their regular working flow and engage in meaningful ethical and strategical discussions to steer the project's future direction. The results of the workshop were captured in a detailed report by the researchers.

Outcomes

During the workshop, several important ethical issues were discussed, among which: fundamental rights and fairness; technical robustness, data governance and risk management; and accountability. The most important results were the following. First, the AI model showed gender bias in speaker recognition due to training primarily on male-dominated city council meetings, and while efforts were made to fix this, the lack of a systematic investigation of bias and limited representation of speakers from diverse genres remained problematic. Second, the system faces several technical and governance risks, including vulnerability to adversarial attacks, data poisoning risks and the risk of complete system failure if the supplier contract ends. Third, there are significant accountability issues due to the dependency on the supplier, with the employees of MO1 lacking knowledge about model functionality, training data, and monitoring procedures. Finally, transparency in the communication between the AI model and its users needs improvement.

In this case study we used an existing tool, making a series of adjustments to the AIIA to better suit the purpose of the AI application. For example, the AIIA follows a top-down approach. However, the AI solution in this case study was already in production, with an existing agreement between the organization and the supplier. We thus changed the dynamics and order of topics so that the main goal was to reveal hidden risks and potential ways to mitigate or remove them. The adjustments were helpful, as at the end of the workshop, it was clear that a series of changes in existing service level agreements must be included in the contract renewal with the supplier – to increase the model's fairness, technical robustness and add governance.

At the end of the workshop, the participants of the media organization mentioned there were some topics that would have been valuable to discuss, but that were not addressed by the tool. According to them, the modified tool would ideally have covered more strategic considerations as well, such as gaining insights in a cost-benefit analysis to make decisions about a potential new contract with a client.

We conclude that using the AIIA helped to get a general overview of the speaker labeling application's impact. Addressing the various ethical considerations was helpful and gave an overall picture of what is important for this use case. Yet, even though the tool was modified, not all topics that were relevant to discuss were covered by the tool.

3.2 Case Study 2: AI Impact Assessment for Music Selection

Use Case

The AI application in this case study concerns automated music selection at MO2. MO2 developed an AI system for music selection called ProsAIc, which it developed in-house, in collaboration with partners. MO2 used it to automatically select appropriate background music for a given TV program. Professionals at MO2 were aware of ethical issues around AI-based music selection, for instance, the representation of different (groups of) artists in the AI model, e.g. based on ethnicity, gender and popularity. Media professionals at MO2 had a desire to develop and deploy the system in a responsible way and were interested in using an ethical tool to reflect on that. Besides wishing the ethical tool would assist them in making ethical choices, they saw the use of the tool as an opportunity to record their ethical decision-making process, thus promoting accountability.

Ethical Tool

The ethical tool selected in this case study concerned the ECP AI Impact Assessment (AIIA) (ECP 2018). Like the AIIA in the previous case study, this AIIA consists of a list of questions aiming to help people to get insight into the impact of a given AI system. It is, however, a different set of questions, organized in a different way. The AIIA of the EPC consists of three phases. The first phase (step 1) is the necessity phase, which consists of eight questions that are meant to determine whether there is a necessity for executing the AIIA assessment. If at least one of these first eight questions is answered with “yes”, ECP advises to perform the full assessment. The second phase (steps 2-5) is the description phase, and consists of describing the application, describing the gains, analyzing ethical and legal responsibilities, and analyzing the reliability, safety and transparency of the application. Each step contains several questions, some further detailed in sub-questions. The last phase (step 6-8) is the decision and reporting phase. This phase contains questions that help to make decisions, document them and periodically evaluate them.

MO2 had indicated their interest in using the AIIA, but that they often had to deal with time constraints. They indicated that they were most likely to use an ethical tool when it would only cost a limited amount of time. We therefore decided to focus only on the first two phases (steps 1-5) of the AIIA, as these matched closely with the need of MO2.

A session was organized at MO2 to assess the impact of the ProsAIc AI system for music selection with the ECP AIIA. A group of professionals from MO2 consisting of a senior data science manager, a senior developer and a data analyst participated in the session, which was guided by a researcher. During the session, first the AIIA was shortly introduced by the researcher, then the participants answered the questions in the AIIA, and finally, the participants were asked to answer several questions to evaluate the AIIA as an ethical tool. In the evaluation, participants were asked, for instance, which questions in the AIIA they deemed relevant and useful (or not), what insights they gained during the AIIA, and whether and why (not) they would consider doing an AIIA for new projects. Based on these inputs, a revised (shortened) version of the AIIA was made by a team of three researchers. The shortened version of the AIIA was eventually evaluated with MO2 to determine appropriateness.

Outcomes

In the final session, in which the use of the shortened AIIA was evaluated, the participants expressed that they were positive about the adapted assessment. They indicated that the process of adapting and applying the AIIA had contributed to creating awareness in the media organization about the importance of reflecting on ethical aspects in the development and deployment of AI. There were two main reasons why the professionals at MO2 had not made use of ethical tools such as impact assessments until that point.

A first reason why ethical tools were not (regularly) used by professionals at MO2 was because of unfamiliarity and ignorance about the value and application of such assessments. Within the media organization, there was a shared belief that ethical tools such as impact assessments should be geared for the media practice specifically. However, after using the adapted AIIA (which was not media-specific), MO2 participants expressed their value in using the tool, and that they no longer thought that an ethical tool should necessarily be designed for a media context for it to be useful.

A second reason why ethical tools were not (regularly) used at MO2 was because MO2's professionals were reluctant to invest (much) time to ethical aspects of AI development and deployment. They stated that as a commercial media organization, they gave priority to commercial considerations, i.e., time investment versus pertained value of using an ethical tool. During the first session we held, it took 2.5 hours to fill in the AIIA, which MO2 participants deemed too much for regular use of such a tool. During the evaluation part of the session, they indicated that large differences in the extent to which different parts or steps of the AIIA they considered valuable. For instance, step 4 ("Are the goal and the way the goal is reached ethically and legally justifiable?") was seen as most interesting, and MO2 participants suggested including the company's values in this step. In contrast, steps 1-3 ("Determine the need to perform an AIIA", "Describe the AI application", and "Describe the benefits of the AI application") were deemed of no or little use. During the session it took 1 hour and 20 minutes to perform step 1-3, whereas the MO2 participants stated that the information discussed in these steps was documented internally for each project, and therefore these steps felt obsolete.

After the experience with using the adapted AIIA, the MO2 participants stated that they were interested in implementing it into their work processes. However, they thought that a reduction in time required to perform the assessment was key in lowering the threshold to use it. They came up with several suggestions for further adaptations to the AIIA to lower the time investment of using it. For instance, as projects at this media company were often continuations of previous projects, they suggested to make use of earlier assessments if there was overlap in risk profile and ethical implications. Participants also suggested to create a company-wide template for the documentation of the AIIA, including all questions and existing documentation.

3.3 Case Study 3: Interactive Model Cards for Speech Recognition

Use Case

This case study concerns an Automatic Speech Recognition (ASR) system at MO2. Many media organizations use ASR to generate subtitles for television programs and other content. Generating automatic subtitles involves extracting audio, identifying the spoken language, recognizing speech, and, in some cases, translating content. ASR performance has improved significantly in recent years, driven by the introduction of end-to-end transformer-based models and unsupervised training techniques (Radford et al. 2023). MO2 has a platform with a popular video-on-demand streaming service that uses AI for various tasks, including automatic subtitle generation. Currently, for some shows a fully automated procedure is used to generate subtitles, while for other shows a semi-automated procedure is used.

Although the quality of AI-generated subtitles has drastically improved over the past few years, it is well-known that ASR systems do not work equally well for all speakers. Speech of young children and seniors, as well as that of people with a (strong) accent, is often not properly recognized by an ASR system (Feng et al. 2023). This can result in subtitles containing more errors for certain speaker groups. MO2 therefore finds it important to evaluate the outcomes of ASR systems, and to correct their biases. This evaluation and correction process is ongoing work for MO2, because of the rapid advancements in the field. Changes in the deployed AI models are becoming increasingly frequent, and pressure media organizations to constantly replace models with those with better accuracy, with risks of introducing bias embedded in new versions.

In the initial phase of this case study, a bias evaluation pipeline was built to quantify bias for ASR systems, and was executed for two state-of-the-art ASRs, namely Wav2Vec2 and Whisper (Fuckner et al. 2023). The evaluation focused on the Dutch language and investigated the performance of different groups in age, gender, speaker region and language proficiency. The result was a comprehensive report of error rates for different groups, fine-grained data containing segmented audio, ground truth, generated transcripts, and metrics at word, character, and phoneme levels. MO2 was interested in using an (ethical) tool to disclose these results to stakeholders accountable for the ASR and those interested in the ASR system's performance.

Ethical Tool

The ethical tool selected in this case study were model cards. Model cards, proposed by (Mitchell et al. 2019), are structured documents that provide essential information about an AI model, such as the model's characteristics, performance metrics, potential biases and harms, and limitations. They thus serve as a communication tool. The content of a model card is often provided by the developer of the AI system, and often, important details useful for experts and non-experts are omitted. Examples are the model performance for different groups and intersections, such as the ones generated by the evaluation pipeline in this case study.

To increase the transparency of the bias evaluation pipeline and broaden the target audience to non-experts, we proposed a variation of model cards called interactive model cards (IMC). IMCs introduce different modalities for interrogating the model's performance without additional work from the developer (Crisan et al. 2022). The IMC allows experts and non-experts to perform quantitative assessments of model performance. This interface lets users compute average word error rates for different groups such as age, gender, region, language proficiency and different combinations of groups.

We gave a presentation about the IMCs to data scientists and other interested colleagues (about 20 people) in MO2. The team of data scientist responsible for the subtitle generation process used the model card for decision-making regarding MO2' ASR system for their streaming platform. MO2 was considering replacing the ASR system used for subtitle generation, more specifically, to replace Wav2Vec2 by Whisper. The team was aware that Whisper outperformed Wav2Vec2 in general, but they needed to determine how well the ASR performed for different groups and intersections of different groups to mitigate bias. The IMC was used to obtain more fine-grained statistics (e.g. per groups) of the performance of both models, which showed that Whisper outperformed Wav2Vec2 because it showed smaller performance disparities between different groups. The team reported their decision to replace Wav2Vec2 by Whisper to the upper management, including the IMC metrics to justify their decisions.

Outcomes

The development and use of an interactive model card to inform decision making about ASR tools had clear benefits. The use of the tool supported the decision-making process in MO2 in choosing between multiple ASR systems, by providing insight into the biases of both systems in a clear way, also suitable for non-developers. The availability of performance indicators across different groups and intersections in the model card also boosted 'productive skepticism' within the team. This term, coined by (Crisan et al. 2022), refers to a neutral orientation towards models, not dismissive or over-trusting AI models. For example, checking how an ASR system performs based on different combinations of groups or performing fine-grained checks on how the ASR system performs at the sentence level was useful for the team in understanding in which situations the system performs poorly.

The close collaboration between researchers and specialists from MO2 helped refine the process of selecting ASR systems, by creating more explicit model selection and exclusion criteria based on the metrics collected from the model card. As part of this refinement, the team chose to reduce the emphasis on global performance scores and instead consider a broader set of indicators related to age, gender, language proficiency, accent and their intersections.

3.4 Case Study 4: Prototypes for Content Recommendation

Use Case

This case study describes the development of a recommendation system in MO3. MO3 has a website providing recommendations for radio and television programs to the public. Most of the recommendations on the website are handpicked by curators and displayed to all users of the platform, and a small number of the recommendations are selected automatically by a PSM recommender to match a user's preferences. During this case study, MO3 has been deploying this relatively limited version of their PSM recommender for a number of years, and they were working on developing a more extended version of the recommendation system. They had started an internal project to develop a prototype of a more advanced recommendation system, serving as a proof of concept. If the prototype would be considered successful, MO3 would implement the solution on their regular website.

As a public media organization, MO3's recommender had to meet multiple requirements. To be of value, the recommender needed to match their users' interests and preferences. Additionally, the recommender needed to be in line with MO3's mission, which included informing the public and exposing them to a balanced mix of different views and perspectives. In other words, MO3 had to balance the requirement of personalization with the requirement to provide diversity (also called pluriformity) in their recommendations. More generally, MO3 faced the challenge to decide what criteria their recommendation system should optimize for, how these criteria could be operationalized, how they could weighed against each other, and how the recommendations should be presented to the user. These decisions needed to be made before the system could be developed, since design choices have ethical implications regarding the way the Dutch audience is informed.

Ethical Tool

The ethical tool used in this case study were prototypes that embodied different ways to balance the requirements of personalization and diversity. We selected this tool because prototypes can help to facilitate a conversation between people with different backgrounds. This was needed, since the decisions to be made in the design phase of the recommendation system, required input of people with different kinds of expertise, such as content curation, UX design and technical AI expertise. Having a meaningful conversation between people with different backgrounds can be challenging, and artifacts such as prototypes can serve as a boundary object to facilitate such conversations (Star 1989; Van der Horst, Overdiek and Harbers 2023). We believed that providing concrete examples of various expressions of diversity in recommenders through prototypes would help the team to better express their preferences and make decisions.

To foster a conversation about strategic choices about the recommendation system, we developed four prototypes based on a conceptual framework on diversity in recommendation systems from (Helberger 2019), further elaborated on in (Vrijenhoek et al. 2021). The conceptual framework contained four models, which we translated into prototypes of recommendation systems, taking the available metadata at MO3 into account (Harbers, Hajri and Stembert 2024). The four models are the following.

- **Liberal model**. This model promotes autonomy, self-development and dispersion of power by facilitating the specialization of a user in an area of his/her choosing and by tailoring to the user's preferences.
- **Participatory model**. This model promotes inclusiveness, participation and active citizenship by making sure that different users do not necessarily see the same content, but they do see the same topics, the recommended content's complexity is tailored to a user's preference and capability, and it reflects the prevalent voices in society.
- **Deliberative model**. This model promotes deliberation, tolerance, open-mindedness and public sphere by focusing on topics that are currently at the center of public debate, and within those topics, presenting a plurality of voices and opinions.
- **Critical model.** This model promotes including marginalized voices and defying prejudices by emphasizing on voices from marginalized groups.

The prototypes showed for the different models which recommendations the system would provide and what it would look like for a user, making use of MO3's own content.

We held a workshop in which two researchers presented the four prototypes to the team that was responsible for developing the new recommendation system at MO3, consisting of a project manager, two content curators, two AI developers and a UX designer. After the short presentation we moderated a discussion in which the workshop participants discussed the advantages and disadvantages of the different prototypes, including their ethical implications.

Outcomes

The four prototypes evoked a lot of discussion, e.g., on diversity versus personalization, accuracy, privacy and transparency. In the presentation of the prototypes and the following discussions, team members expressed their (individual) preferences and opinions in ways they had not done before. The workshop did not lead to concrete decisions during the workshop itself. Some participants were positive about the workshop and deemed it highly valuable (participants with a non-technical background), other participants were positive about the workshop and the discussion, but expressed disappointment about the fact that no decision were made during the workshop (the developers).

The development of the prototypes involved that the researchers were regularly present at project meetings to understand the complexity of the case, a short presentation of the researchers to the recommendation system team, and a session to surface which metadata MO3 had available as input for the recommendation system. At the start of the project it was clearly agreed that we as researchers were not responsible for the (normative) decisions regarding diversity of the recommendation system. However, throughout the project, other team members were not working on it actively, possibly because they knew we were involved. Though we tried to engage the team in working on the questions regarding the diversity of the recommendation system, and they actively engaged in the discussions about it, we did not manage to transfer ownership of the problem to the team.

4. Findings

Based on the four case studies described in the previous section (summarized in Table 1), we can draw several general insights into the process of implementing ethical tools in media organizations to design and establish a responsible AI practice**.** In this section we discuss findings about the use cases, ethical tools, their application, and the organizational context in which they were applied.

The responsible AI **use cases** in the four case studies showed a great variety. First, there is a high diversity in the type of AI application at stake (e.g., speaker labelling in case study 1 and a recommender system in case study 4), and whether the focus is on developing a new application, evaluating an existing AI system, or an AI system supplied by a third party. There are also differences in the phase of the process that is being supported by applying the ethical tool (e.g., strategic decision-making, design, development, or deployment). Another important difference between the case studies are the organizational structures and cultures in the different media organizations. For example, MO2, being a commercial organization, seemed to focus more on the time and resources of applying an ethical tool than MO1 and MO3, both public organizations. Case study 4 showed that being funded by public resources can come with stricter responsibilities regarding ethical implications, as MO3's mission to provide diversity in information is determined by law. Finally, we found that the awareness and knowledge of responsible AI varied a lot for different people that were involved in the case studies.

The specific characteristics of a use case, such as described above, pose different requirements on an ethical tool. In literature on ethical tools for AI, however, relatively little attention is paid to the characteristics of a use case to which an ethical tool is applied. In the typology proposed by Morley and colleagues (2020), one of the most cited overviews of AI ethics tools, tools are organized according to 1) ethical principles (e.g. beneficence, autonomy, and explicability) and their requirements, and 2) seven stages of machine learning algorithm development (e.g., design phase, testing, and monitoring). Ayling and Chapman (2022) classified ethical tools according to which sector the authors/users were from, which stakeholder would use the tool or result, type of tool, internal or external use, stage in AI production in which it was used, whether it addressed model, data or both. In a review of ethical tools, Prem (2023) classified approaches according to category (type of approach), ethical issue addressed, and phase in AI application development process in which it was used. While all these researchers consider phase in the development process, they do not consider organizational structure and culture in which a tool is or should be applied, or the (required) awareness and knowledge of people using the tools. This seems to indicate a gap in current research: whereas our case studies showed that contextual factors highly matter in applying ethical tools, attention for the context in which a tool is or should be used is often missing in current overviews on ethical tools.

Our results show that merely selecting a suitable **ethical tool** does often not adequately address what the situation required. Instead, our findings suggest that it is necessary to tailor ethical tools to the situation at hand. In the case studies we used modified lists of assessment questions for evaluating AI-systems of which the purpose was relatively clear (cases 1 and 2), self-developed prototypes if the consequences of design choices were less clear (case 4), and self-developed interactive model cards when it concerned an AI-system that is frequently updated and consequences need to be documented (case 3). The process of selecting and possibly modifying tools, or developing specialized tools sometimes included adding media-specific aspects (case 4), but tailoring the tools also largely depended on other factors. The specific AI technology or application, the stakeholder group who would use the tool, and the phase in development/deployment process all provide requirements to an ethical tool. In addition to that, important factors to consider when selecting and/or modifying a tool include the amount of time people are willing or able to spend (case 2), the work that has already been done (case 1), or that is lacking (case 3 and case 4).

The tailoring of ethical tools to specific situations has received little attention in current literature on AI ethics. Articles introducing an ethical tool often include one or more examples of its application (e.g., Franzke, Muis and Schäfer 2021). In those cases, however, the application context was selected for the tool. There is very little work in which a real-world situation forms the starting point for which a tool has to be selected. In such situations, it is more likely that there is no perfectly fitting existing tool, and that it is therefore needed to develop or adapt an ethical tool to fit the needs of the specific context. As mentioned in Section 2 of this paper, in general, there is still limited work in which actual ethical tools are applied in a realistic setting for actual real-world problems.

Following the previous finding, the case studies show that **applying ethical tools** requires in-depth knowledge and skills. As mentioned in Section 2, Morley and colleagues (2023) concluded AI practitioners require knowledge and/or methods and tools to operationalize AI ethics in practice. Our results show that the required knowledge and skills do not only include selecting the proper tool and applying it, but possible also tailoring the tool to the specific situation at hand. This requires a good understanding of ethics, a sufficiently large toolbox and the experience to go beyond predefined tools. One could argue that translating ethical guidelines into practice requires not necessarily the availability of ethical tools, but rather professionals capable of designing and deploying context-specific ethical tools.

In addition to selecting, adapting and applying ethical tools, in all four case studies we did additional 'invisible' work as researchers, such as explaining our research goals, creating awareness about the importance of responsible AI, and finding the right person in the organization. This work may seem trivial, but it took quite some time and involved overcoming obstacles. More than once, we encountered a certain level of reluctance among employees of the Mos to engage in ethical work. A reoccurring observation in all organizations was the desire for fast, simple and clear solutions, whereas ethics involves complex processes and making actual tradeoffs and does typically not lead to fast and clear solutions. Our work thus indicates that establishing a responsible AI practice may involve the affective labor of convincing others to put time and effort in responsible AI. This is in line with previous work, in which it has been argued that part of the work on AI ethics involves ‘being brave’ (Gambelin 2021) and that AI ethicists ‘require protection’ (Cocchiaro et al. 2025).

The openness and willingness of employees to put time and effort in AI ethics is influenced by the **organizational context**. None of the three media organizations in this study had discussions about the ethical aspects of an AI application or used ethical tools as part of their existing workflows, since that was not part of their organization’s governance. This is in line with research arguing that AI (ethics) governance is a crucial part of establishing a responsible AI practice (Mäntymäki et al. 2022). The lack of a proper AI ethics governance in the MOs may explain why it took effort to have employees engage in the case studies in the first place. It might also explain why applying the ethical tools did not always lead to practical solutions. In case study 1 and 3, the application of ethical tools led to specific practical insights and action points, but in case study 2 and 4 it mostly led to an increased (ethical) awareness, even though this research was motivated by practical efforts to help media organizations on small scale projects by extrapolating ethical tools in specific contexts. This shows that the application of ethical tools does not necessarily fully address the underlining multi-faceted challenges that society and media organization are facing.

5. Limitations

This study has a number of limitations. There are many ethical tools and approaches available for each use case and corresponding (ethical) challenges. The tools and approaches we selected were based on our knowledge and experience as researchers in the field of responsible AI. We do not suggest that we made optimal choices. Our knowledge and experience is (by definition) limited, and we may have biases and blind spots that influenced our choices.

Another limitation of this study is that the four case studies are different in nature and sometimes difficult to compare. The reason for this variety in nature is that we were interested in realistic situations, and for that we were dependent on the use cases of the media organizations. In selecting the use cases we tried to stay as close as possible to the wishes of the media organizations, and as such, we had less control over the research process. The benefit of our approach is that all four case studies are realistic, actual problems that media organizations were dealing with.

6. Conclusion

In this study we investigated what it takes to apply ethical tools in the real-world practice of designing, developing and deploying AI in the media sector. We conducted four case studies at three media organizations in the Netherlands, in each of which we applied an (adapted or self-developed) ethical tools to improve the responsible application of a given AI application. The case studies all involved realistic examples of challenges and practices of responsible AI.

The four case studies convincingly showed that the application of an ethical tool is highly contextual, and the mere availability of ethical tools is not enough for establishing a responsible AI practice. Besides ‘just’ applying an ethical tool, a suitable ethical tool needs to be selected, and the selected ethical tool needs to be to be tailored and contextualized to specific needs, related to e.g., the AI technology or application that is being considered, the stakeholder group involved, the phase in the design, development or deployment process, the time available, and the information available. Moreover, organizational AI governance highly influences how and where attention is paid to AI ethics, and thus the establishment of a responsible AI practice.

Selecting and adapting an ethical tool requires *awareness* of the importance of paying attention to the ethical implications of AI, *knowledge* about how to address ethical implications of AI, and *willingness* to invest time and resources in mitigating ethical risks associated to AI. Therefore, these factors are imperative for the implementation of ethical tools in an organization to establish a responsible AI practice.

Acknowledgments

This research was supported by NWO-SIA under grant RAAK.PUB08.040 GRANT. The authors thank the involved MOs for their collaboration and all participants in the activities for their valuable contributions.

References

Agbese, M.; Mohanani, R.; Khan, A.; and Abrahamsson, P. 2023. Implementing AI Ethics: Making Sense of the Ethical Requirements. In Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering. Oulu Finland. https://doi.org/10.1145/3593434.3593453.

Ayling, J.; and Chapman, A. 2022. Putting AI ethics to work: are the tools fit for purpose?. *AI and Ethics*, 2(3): 405-429. https://doi.org/10.1007/S43681-021-00084-X.

Benjamins, R.; Barbado, A; and Sierra, D. 2019. Responsible AI by design in practice. arXiv:1909.12838.

Cocchiaro, M. Z.; Morley, J.; Novelli, C.; Panai, E.; Tartaro, A.; and Floridi, L. 2025. Who is an AI Ethicist? An empirical study of expertise, skills, and profiles to build a competency framework. *AI and Ethics*, 1-13. https://doi.org/10.1007/s43681-024-00643-y.

Cramer, H.; Garcia-Gathright, J.; Springer, A.; and Reddy, S. 2018. Assessing and addressing algorithmic bias in practice. *Interactions*, 25(6): 58-63. https://doi.org/10.1145/3278156.

Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive model cards: A human-centered approach to model documentation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul, Republic of Korea. https://doi.org/10.1145/3531146.3533108.

ECP. 2018. Artificial intelligence impact assessment. https://ecp.nl/artificial-intelligence-impact-assessment/. Accessed: 2024-03-21.

Elahi, M.; Jannach, D.; Skjærven, L.; Knudsen, E.; Sjøvaag, H.; Tolonen, K.; Holmstad, Ø.; Pipkin, I.; Throndsen, E.; Stenbom, A.; and Fiskerud, E. 2022. Towards responsible media recommendation. *AI and Ethics* (2)1: 103–114. https://doi.org/10.1007/s43681-021-00107-7.

Feng, S.; Halpern, B.M.; Kudina, O; and Scharenborg, O. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84. https://doi.org/10.1016/j.csl.2023.101567.

Franzke, A.S.; Muis, I; and Schäfer, M.T. 2021. Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands. *Ethics and Information Technology*, 23(3): 551-567. https://doi.org/10.1007/s10676-020-09577-5.

Fuckner, M.; Horsman, S.; Wiggers, P.; & Janssen, I. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In Proceedings of the 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 146-151, IEEE. https://doi.org/10.1109/SpeD59241.2023.10314895.

Gambelin, O. 2021. Brave: what it means to be an AI Ethicist. *AI and Ethics*, 1(1), 87-91. https://doi.org/10.1007/s43681-020-00020-5.

Government of the Netherlands. 2023. AI Impact Assessment. https://www.government.nl/documents/publications/2023/03/02/ai-impact-assessment. Accessed: 2024-03-20.

Harbers, M.; Hajri, O.; & Stembert, N. (2024). Responsible AI in Practice: A Case Study on Designing a PSM Recommender. In Proceedings of HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), Malmö, Sweden.

Helberger, N. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8: 993–1012. https://doi.org/10.1080/21670811.2019.1623700.

Helberger, N.; and Diakopoulos, N. 2023. The European AI act and how it matters for research into AI in media and journalism. *Digital Journalism*, 11(9): 1751–1760. https://doi.org/10.1080/21670811.2022.2082505.

Helberger, N.; Karppinen, K.; and D'acunto, L. 2018. Exposure diversity as a design principle for recommender systems. *Information, communication & society*, 21(2): 191-207. https://doi.org/10.1080/1369118X.2016.1271900.

Helberger, N.; Van Drunen, M.; Moeller, J.; Vrijenhoek, S.; and Eskens, S. 2022. Towards a Normative Perspective on Journalistic AI: Embracing the Messy Reality of Normative Ideals. *Digital Journalism* 10, 10: 1605–1626. https://doi.org/10.1080/21670811.2022.2152195.

HLEG. 2019. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 2025-01-19.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9): 389-399. https://doi.org/10.1038/s42256-019-0088-2.

Mäntymäki, M.; Minkkinen, M.; Birkstedt, T; and Viljanen, M. 2022. Defining organizational AI governance. *AI and Ethics, 2*(4), 603-609. https://doi.org/10.1007/s43681-022-00143-x.

Media Campus. 2021. Brede steun voor ethische richtlijnen voor gebruik AI in media. https://mediacampus.nl/kennisbank/brede-steun-voor-ethische-richtlijnen-voor-gebruik-ai-in-media/. Accessed: 2025-19-05.

Mioch, T.; Stembert, N.; Timmers, C.; Hajri, O.; Wiggers, P.; Harbers, M. 2023. Exploring Responsible AI Practices in Dutch Media Organizations. In: Abdelnour Nocera, J., Kristín Lárusdóttir, M., Petrie, H., Piccinno, A., Winckler, M. (eds) Human-Computer Interaction – INTERACT 2023. INTERACT 2023. Lecture Notes in Computer Science, vol 14145. Springer, Cham. https://doi.org/10.1007/978-3-031-42293-5_58.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; and Gebru, T. 2019. Model cards for model reporting. In Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency Atlanta GA USA. https://doi.org/10.1145/3287560.3287596.

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4): 2141-2168. https://doi.org/10.1007/s11948-019-00165-5.

Morley, J.; Kinsey, L.; Elhalal, A.; Garcia, F.; Ziosi, M.; and Floridi, L. 2023. Operationalising AI ethics: barriers, enablers and next steps. AI & SOCIETY, Volume 38: 1-13. https://doi.org/10.1007/s00146-021-01308-8.

Ortega-Bolaños, R.; Bernal-Salcedo, J.; Germán Ortiz, M.; Galeano Sarmiento, J.; Ruz, G.A.; and Tabares-Soto, R. 2024. Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*, 57(5): 110. https://doi.org/10.1007/s10462-024-10740-3.

Prem, E., 2023. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, 3(3): 699-716. https://doi.org/10.1007/s43681-023-00258-9.

Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (ICML'23). Honolulu, Hawaii, USA. PMLR 202:28492-28518.

Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. In Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1): 1-23. https://dl.acm.org/doi/10.1145/3449081.

Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A; and Lennon, M., 2020. Principles to practices for responsible AI: closing the gap. arXiv:2006.04707.

Stappers, P.J.; and Giaccardi, E. 2017. Research through design. In *The Encyclopedia of Human-Computer Interaction*, edited by Kurian, G. T.; and I. N. Chief, 1-81. Aarhus, Denmark: Interaction Design Foundation.

Star, S.L. 1989. The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In *Distributed artificial intelligence*, 37-54. Morgan Kaufmann.

Trattner, C.; Jannach, D.; Motta, E.; Costera Meijer, I.; Diakopoulos, N.; Elahi, M.,; Opdahl, A.L.; Tessem, B.; Borch, N.; Fjeld, M.; and Øvrelid, L. 2022. Responsible media technology and AI: challenges and research directions. *AI and Ethics*, 2(4): 585-594. https://doi.org/10.1007/S43681-021-00126-4.

Van der Horst, T.; Overdiek, A; and Harbers, M. 2023. Designing a Physical Boundary Object to Invite Dialogue about Power Relations behind AI Systems. In Proceedings of Related Systems Thinking and Design Symposium. October 6-20, 2023, Georgetown, USA.

Vrijenhoek, S.; Kaya, M.; Metoui, N.; Möller, J.; Odijk, D.; and Helberger, N. 2021. Recommenders with a mission: assessing diversity in news recommendations. In Proceedings of the 2021 conference on human information interaction and retrieval. https://doi.org/10.1145/3406522.3446019.

Zhou, J.; and Chen, F. 2023. AI ethics: From principles to practice. AI & SOCIETY, Volume 38, 38(6): 2693-2703. https://doi.org/10.1007/s00146-022-01602-z.

Zimmerman, J.; Forlizzi, J.; and Evenson, S. 2007. Research through design as a method for interaction design research in HCI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, San Jose California USA, 493–502. https://doi.org/10.1145/1240624.1240704.